VACCINE

Visual Analytics for Command, Control and Interoperability Environments
A U.S. Department of Homeland Security
Science and Technology Center of Excellence

VACCINE ANNUAL REPORT – YEAR 5

Addendum B - Publications

July 1, 2013 – June 30, 2014

Cooperative Agreement No. 2009-ST-061-CI0001

PURDUE UNIVERSITY

HOMELAND SECURITY UNIVERSITY PROGRAMS
TODAY'S RESEARCH & EDUCATION, TOMORROW'S SECURITY

Addendum B – Publications

Table of Contents

PENNSYLVANIA STATE UNIVERSITY

Purdue – Delp

Purdue – Ebert

Purdue – Elmqvist

University of North Carolina at Charlotte

University of Oxford

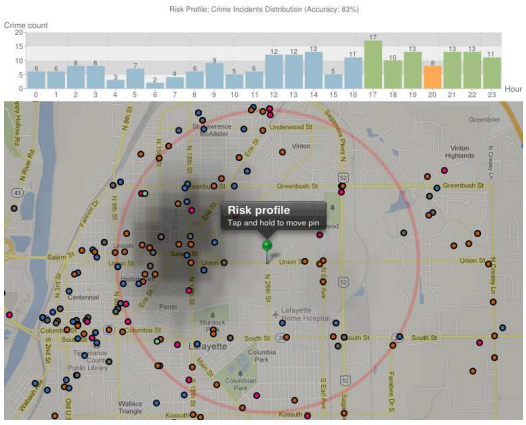
A Mobile Visual Analytics Approach for Law Enforcement Situation Awareness

Ahmad M. M. Razip, Abish Malik, Shehzad Afzal, Matthew Potrawski*
Purdue University

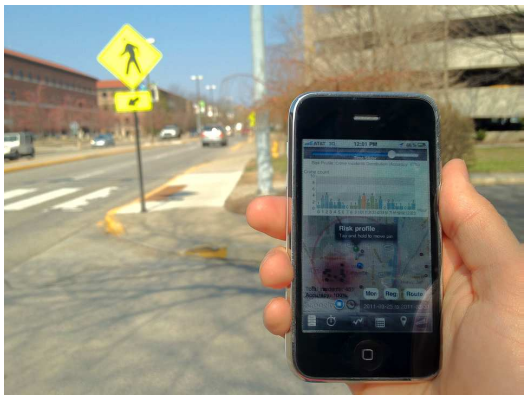
Ross Maciejewski†
Arizona State University

Yun Jang‡
Sejong University

Niklas Elmqvist, David S. Ebert§
Purdue University



(a) Risk profile tools give time- and location-aware assessment of public safety using law enforcement data.



(b) The mobile system can be used anywhere with cellular network coverage to visualize and analyze law enforcement data. Here, the system is being used as the user walks down a street.

Figure 1: Mobile crime analytics system gives ubiquitous and context-aware analysis of law enforcement data to users.

Abstract

The advent of modern smartphones and handheld devices has given analysts, decision-makers, and even the general public the ability to rapidly ingest data and translate it into actionable information on-the-go. In this paper, we explore the design and use of a mobile visual analytics toolkit for public safety data that equips law enforcement agencies with effective situation awareness and risk assessment tools. Our system provides users with a suite of interactive tools that allow them to perform analysis and detect trends, patterns and anomalies among criminal, traffic and civil (CTC) incidents. The system also provides interactive risk assessment tools that allow users to identify regions of potential high risk and determine the risk at any user-specified location and time. Our system has been designed for the iPhone/iPad environment and is currently being used and evaluated by a consortium of law enforcement agencies. We report their use of the system and some initial feedback.

Keywords: Mobile visual analytics, situation awareness, public safety

Index Terms: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques; I.3.8 [Computer Graphics]: Applications—Mobile Visual Analytics

*e-mail: {mohammea|amalik|safzal|mpotraws}@purdue.edu
†e-mail: rmacieje@asu.edu
‡e-mail: jangy@sejong.edu
§e-mail: {elm|ebertd}@purdue.edu

1 Introduction

In 2010, the number of handheld devices reached a staggering volume of 4 billion devices globally [9]. With a large and diverse user base, it is the only truly universal computational platform today. With this global explosion in the usage of modern smartphones and handheld devices, users now have more connectivity to the digital world and the ability to ubiquitously ingest data and transform it into knowledge that enables them to comprehend a situation better and make more effective decisions. However, challenges associated with the data processing, exploration and analysis on a mobile platform are becoming prominent due to the increasing scale and complexity of modern datasets and limited screen space of mobile devices. These challenges are being addressed by the emerging field of visual analytics [40]. Visual analytics in the mobile domain utilizes state-of-the-art mobile devices and provides users with the ability to effectively and interactively analyze large and complex datasets on-the-go; thereby, providing analysts, in-field workers, responders, decision makers and other users insights into any emerging or emergent situation in real-time (Figure 1).

In this paper, we present a mobile visual analytics approach for solving one such problem in the public safety domain. Our work leverages the ubiquity of the mobile platform and focuses on creating an effective, interactive, client-server-based situational analysis and analytics system for the geotemporal exploration of criminal, civil and traffic (CTC) incidents. Our mobile system (Figure 1b) provides on-the-go situational awareness tools to law enforcement officers and, in the future, to citizens. Our system has been designed in collaboration with a consortium of law enforcement agencies and first responder groups and has been developed using a user-centered approach. Designing a mobile visual analytics system in this domain provides a unique set of challenges that range from identifying the in-field needs of a diverse group of end-users to understanding the applicability of a mobile solution to improve the day-to-day operations of law enforcement officers. We discuss these challenges

and present our mobile solution that has the following main contributions:

- *Identifying the needs of first responders and law enforcement agencies in the mobile domain.* Our collaboration with the law enforcement agencies enables us to explore and discuss the use of our system in their day-to-day tasks at different organizational levels in the agencies.

- *Discussion of the design of a public safety mobile visual analytics solution for the first responder community.* We adapt several methods and techniques to work effectively in a mobile environment that has unique interaction methods, use cases, and objectives (e.g., risk profile, hotspot alerts, plume visualization).

2 Related Work

The use of mobile devices in visual analytics has proliferated greatly over the past few years. Mobile devices should not be thought of merely as auxiliary devices for use while on the road but they are the new personal computers [9]. With the explosion of the number of users of smartphone, tablets and other mobile devices, the interest to develop visual analytic systems on the mobile platform has greatly increased. For example, work by Kim et al. [23] presents a mobile visual analytics system for emergency response cases and highlights the use of such systems in time-critical applications. However, these systems face a unique set of challenges compared to the commonly used desktop systems. Some of the main constraints of mobile systems include limited performance, small displays, and different usage environments [30, 37]. Novel methods that deal with these unique constraints are thus a must. Pattath et al. [31], for example, focus their work mainly on addressing the display and interaction area size constraint by utilizing the focus+context technique. Additionally, the development of mobile visual analytics systems provides novel use cases that traditional desktop systems cannot provide.

For a dataset of geospatiotemporal nature (data with information to geographical location and time) such as public safety data, geographic information systems (GIS) play an important role in the exploration and analysis processes in decision support environments [3]. There exist many GIS systems that provide tools to support the Exploratory Data Analysis (EDA) [41] process. Many of these systems provide a geospatial interface along with statistical tools for analysis and are usually designed to enable users to explore trends, patterns and relations among such datasets [6, 16]. For example, Andrienko et al. developed the Descartes system [5] that automates the presentation of data on interactive maps. The GeoDa system [8] also provides an interactive environment for performing statistical analysis with graphics. Many systems are also tailored to focus on specific applications in a given domain. Examples of these include GeoTime [21] and Flow Map Layout [32] that focus on tracking movements of objects in time. Other domain specific applications include traffic evacuation management [19] and urban risk assessment [24].

Similarly, GIS systems provide important support in the public safety domain [12]. In analyzing and modeling the spatiotemporal behavior of criminal incidents, Chen et al. [13], developed a crime analysis system with spatiotemporal and criminal relationship visualization tools. However, in contrast to our focus on risk assessment and situational awareness in the mobile domain, their system focuses on finding relations between datasets and is developed for a desktop environment. Moreover, there exist many web-based crime GIS tools (e.g., [14, 15, 28, 29, 33, 36, 39]). However, most of these tools offer only basic crime mapping and filtering functionalities and provide basic analytical tools to allow users to perform EDA. Also, many of these systems target casual users only. Similarly in the GIS domain, some systems also target non-GIS specialists (e.g., [4, 8]). Our system is designed to enable domain experts and, in the future, ordinary citizens explore and analyze spatiotemporal CTC incidents. Our work uses the notion of casual visualization [34], but further provides statistical tools that help users identify trends, patterns and relations within the data in the EDA process.

The ability to have situational awareness in law enforcement is essential in maintaining public safety. Situational awareness by definition [18] enables one to perceive, comprehend, and project into the future to make more effective decisions. Law enforcement officers make decisions in resource allocation and patrol planning to reduce crime, while citizens may make decisions to be at a place at any particular time. Critical to gaining awareness of a situation are the fundamental questions of where, what, and when an event/incident occurs [7]. Our mobile system, similar to many GIS tools, presents the law enforcement information in a place-time-object organization to allow for information exploration and sense-making using the different tailored views that we have developed based on the user-driven requirements of different situations.

3 Requirements and Challenges

Our system's design was driven by the task, environment, and device factors gathered from our end users through a user-centered design approach.

3.1 Domain Analysis of Requirements

Our collaboration with local law agencies started with our work on an earlier desktop-based system [27]. Having seen the benefits of a desktop based visual analytics system in their operations, the local law agencies approached us with the idea to develop a system for the mobile platform that addresses their mobile needs. As such, we had several meetings and informal discussions about the use of such a mobile system in their day-to-day tasks to derive the system requirements.

Below, we identify the requirements of the law enforcement agencies and explore the use of such a mobile system based on our formative engagements with officers ranging from shift supervisors, patrol officers, detectives, and crime analysts.

R1: Easy operation — Mobile systems are often used ubiquitously, sometimes in less than ideal conditions. Additionally, using a mobile system in such conditions often requires users to divide attention between the system, the task they are performing, and their surroundings. Thus, the visualizations need to be easily comprehensible in a short glance and interaction should be simple and easy.

R2: On-the-go risk assessment and emergency management — Our end users emphasized the need for a mobile analytics system that would provide on-the-go risk assessment to in-field officers. Crime trends are affected by the hour-of-the-day and day-of-the-week effects, and, as such, our risk profile system factors in the current location and time of the officers to provide them with a situational awareness of their surroundings. Additionally, because our end user group included first responders, there was a need for emergency management tools in case of accidents that potentially affected a large population. As such, we provide tools that allow them to visualize the impacts of chemical spills for use in emergency and evacuation situations.

R3: Near real-time data — The shift supervisor is primarily responsible for resource planning and allocation. He needs to have access to the most recent data and look at all the CTC incidents from the prior and current day to prepare for the shift change briefing and roll call. Thus, we design the mobile system on a server-client architecture that centralizes the entire data on the server, which is always kept up to date. Based on recommendations made by our end-users, the CTC data is currently acquired four times a day and put in to our database server which allows the data to be up to date before a patrol shift change and during patrol.

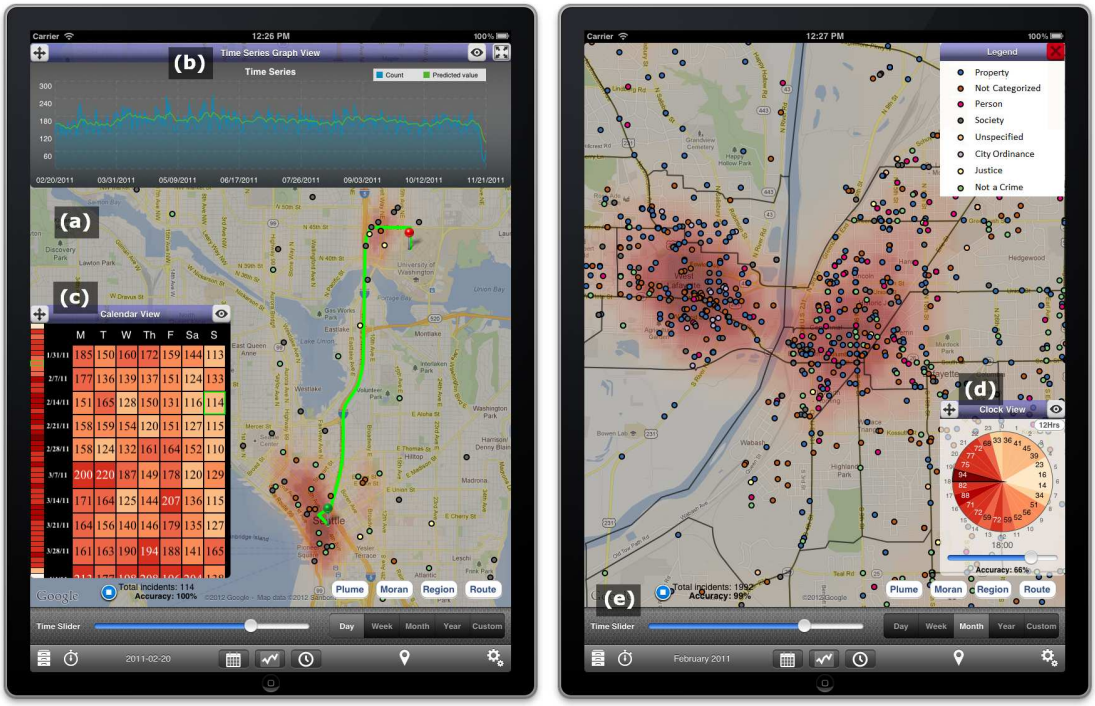


Figure 2: A screenshot of our mobile visual analytics law enforcement system. (left) Visualizing all CTC incidents for the city of Seattle, WA, USA on February 20, 2011. The map view (a) plots the incidents as color-coded points on a map (the legend for the points has been shown in the top-right window). The interactive time series view (b) plots the incident count over time with the estimated weighted moving average (EWMA) control chart overlaid. The bottom-left image (c) shows an overview+detail calendar view of the CTC incidents. (right) Visualizing all CTC incidents for Tippecanoe County, IN, for the month of February 2011. The county's census tracts have been overlaid on the map. The bottom right image (d) shows the interactive clock view to provide an hourly view of CTC incidents for the month selected. The interactive time slider (e) allows users to scroll through time and offers various temporal aggregation levels.

R4: Trend analysis and visualization — During shift briefings, shift supervisors discuss the current crime trends with patrol officers and review, for example, what has been happening over last week, day, this day last week, and so far today. They use this information for tasks including planning their patrols and deciding on what they should be on the lookout for. This is also important for detectives and crime analysts to be able to see the patterns of specific crimes over space and time. Interactive visualization of CTC trends over time is thus an important task required for the mobile system.

R5: Mobile information — The shift supervisors typically use paper print outs of the incidents that happened the previous day on a map during shift briefings to point out certain incidents of interest and show their geospatial trends. We identify that the need to offload information from a workstation to something a person can carry around (e.g., paper) is essential and that the person must be able to explore the information (they may carry multiple print outs of the map showing different information).

3.2 Mobile Challenges

Today, tasks that were once only done commonly using desktop computers can now be done using devices that can fit into a pocket [9]. But developing such a system brings about unique foci, tasks, goals, and constraints. In the following paragraphs, we discuss how some of these constraints affected the final design of our mobile system.

C1: Different task and user intent — Due to the form factor of mobile devices, certain tasks are better done on mobile devices than on desktop workstations. For example, long-term risk assessment and resource allocation are usually done using desktop workstations due to better hardware and screen space for advanced analytics. On the other hand, mobile devices are often used when officers are away from their workstations, or when they want to get a quick access to the data and focus on rapid situation assessments.

C2: Different usage environments — As briefly discussed above, the usage conditions of these systems also differs in that desktop systems are often used in a controlled office environment; whereas, mobile systems are more geared for use on-the-go. For law enforcement officers working in the field, our system is highly beneficial in providing them with tools to increase their situational awareness within their areas of responsibility [17]. Additionally, using such a mobile system out in the field introduces new challenges in that it requires the officers to have divided attention between their surroundings and the system, thereby stressing the need of having views that are intuitive and easy-to-read. As such, our system makes use of casual visualization [34] concepts in addressing these issues.

C3: Limited computing resources — The limited physical size imposed on mobile devices, which restricts their hardware capability such as computing power, memory, display area, and input capabilities due to miniaturization also limits the device performance. Speed is a key feature that users look for when performing EDA, especially because mobile systems are often used in time-critical situations. For example, officers may use this system as they respond to dispatch or emergency calls; and citizens may want to get a quick check of the safety around their area as they walk to their office. With our mobile system, we choose to determine a default data size limit (the number of incidents to load) depending on the device used and the network to guarantee an interactive performance. In this case, the most recent incidents are chosen and shown using the system.

C4: Limited/varying display and interaction area — Device fragmentation in terms of display area was another key factor in the system design process. The screen of a smartphone can only display a fraction of what is displayable on a tablet's screen (e.g., compare the iPhone's 3.5-inch screen to the iPad Mini's 7.9-inch and iPad's 9.7-inch screens). Additionally, interaction also becomes an issue where users would get a richer experience using the iPad's bigger interaction area on its touchscreen as opposed to the small touchscreen on the iPhone [20]. We thus designed the system to be context-aware and behave differently based on the device used.

C5: Security — One of the main benefits of having a risk assessment system on the mobile platform is its ubiquity. The system can be used by law enforcement agencies in situ while responding to dispatch or emergency calls. This increases the risk of having the device misplaced or stolen and thus the risk of having sensitive data falling into the wrong hands. Being used in the public, the system is also more susceptible to being a target of data sniffing. To ensure data confidentiality, we use a secure protocol to transmit encrypted data to the end-devices. We also utilize secure authentication services to authenticate the user logging on to the database in order to ensure the data is protected in case the device is misplaced.

4 Mobile Visual Analytics Environment for First Responders and Law Enforcement

Our mobile visual analytics system provides users with an overview of public safety data in the form of criminal, traffic and civil (CTC) incidents. It comprises a suite of tools that enables a thorough analysis and detection of trends and patterns within these incident reports. Our system has been developed for visualizing multivariate spatiotemporal datasets, displaying geo-referenced data on a map, and providing tools that allow users to explore these datasets over space and time (*R2*). We further provide filtering tools that allow users to dynamically filter their datasets. Our system also incorporates linked spatiotemporal views that enhance user interaction with their datasets.

Figure 2 shows two snapshots of our system. Figure 2 (left) shows all CTC incidents for the city of Seattle, WA, USA occurring on February 20, 2011, and Figure 2 (right) shows the CTC incidents for Tippecanoe County, IN, USA occurring in the month of February 2011. The main view of the system is the map view (Figure 2(a)) that provides users with the ability to plot the CTC incidents as color-coded points on the map (Section 4.3).

With temporal data, the system allows for visualization using several views, namely the time series graph view (Figure 2(b)), the calendar view (Figure 2(c)), and the clock view (Figure 2(d)). The time series graph view allows users to visualize the temporal aspect of the incident data using line graphs and model the data for abnormal event detection. The calendar view [43] lays the temporal data in the format of a calendar, allowing users to visualize the weekly and seasonal trends among the CTC incidents. The clock view, on the other hand, allows users to visualize the hourly distribution of the CTC incidents. A time series slider (Figure 2(e)) is provided to allow users to scroll in time, updating all linked views dynamically. Furthermore, our system also provides users with risk profile tools that allow them to dynamically assess the risks associated with their neighborhoods and surroundings.

4.1 Public Safety Data

Our system was developed using CTC data collected by a consortium of law enforcement agencies in our local county and from publicly available data [38, 39]. Each report entered into the database consists of, among other fields, the date and time of when the incident was reported, the time range between which the incident was thought to have occurred (e.g., in case of burglaries), the geolocation and the charges associated with the incident. Additionally, we provide multiple aggregation levels to group the different crime incidents. Our system provides support for the Uniform Crime Code (UCR) categorization of CTC offenses utilized by the Federal Bureau of Investigation [42] that helps increase familiarity with the system (*R4*).

4.2 System Design

Our system consists of two main components, a server back-end for processing and computation, and a client front-end composed of our interactive mobile visual analytics system. The server back-end consists of a database that enables querying and provides data to the client. The data going into the server undergoes a pre-processing and data cleaning stage so it becomes ingestible to our client system. The front-end consists of the mobile device that provides a user interface for the visualization, exploration, and analysis of the spatiotemporal public safety dataset. The exploration and analysis of data is done per user on his/her device and our end users have indicated that it will be advantageous to have the ability to share this visualization or data exploration state with another user. However, since this is not in our initial list of requirements (Section 3.1), we leave this for future work.

4.3 Geospatial Displays

Our visual analytics system provides multiple views to visualize the spatial component of the datasets. We allow users to plot the incidents as points on the map that are color coded [10] to represent the different parameters of the datasets (e.g., agencies responding to the incident, offense type). The map has been dimmed so as to distinguish these color-coded incident points from the map colors. The radius scaling of the points is dependent on the zoom level. Moreover, in order to tackle the issue of over-plotting the incidents on the map, we provide interaction methods where the users can zoom in and tap on incidents and drill down to his or her level of interest. A better approach in dealing with the over-plotting issue is to show the aggregate sum of the overlapping incidents on the map at different zoom levels, and provide interaction techniques to show details on demand. We leave this as future work.

Furthermore, we utilize a kernel density estimation technique [25] to allow a quick exploration of the incidents on the map and to identify hotspots. The system also allows users to overlay different layers on the map (e.g., law beats, census tracts, bus routes) [44] and allows them to place custom placemarks. The users can also overlay driving and walking routes on the map, enabling them to visualize the CTC spatial distributions along their intended routes. An example of this has been shown in Figure 2(a).

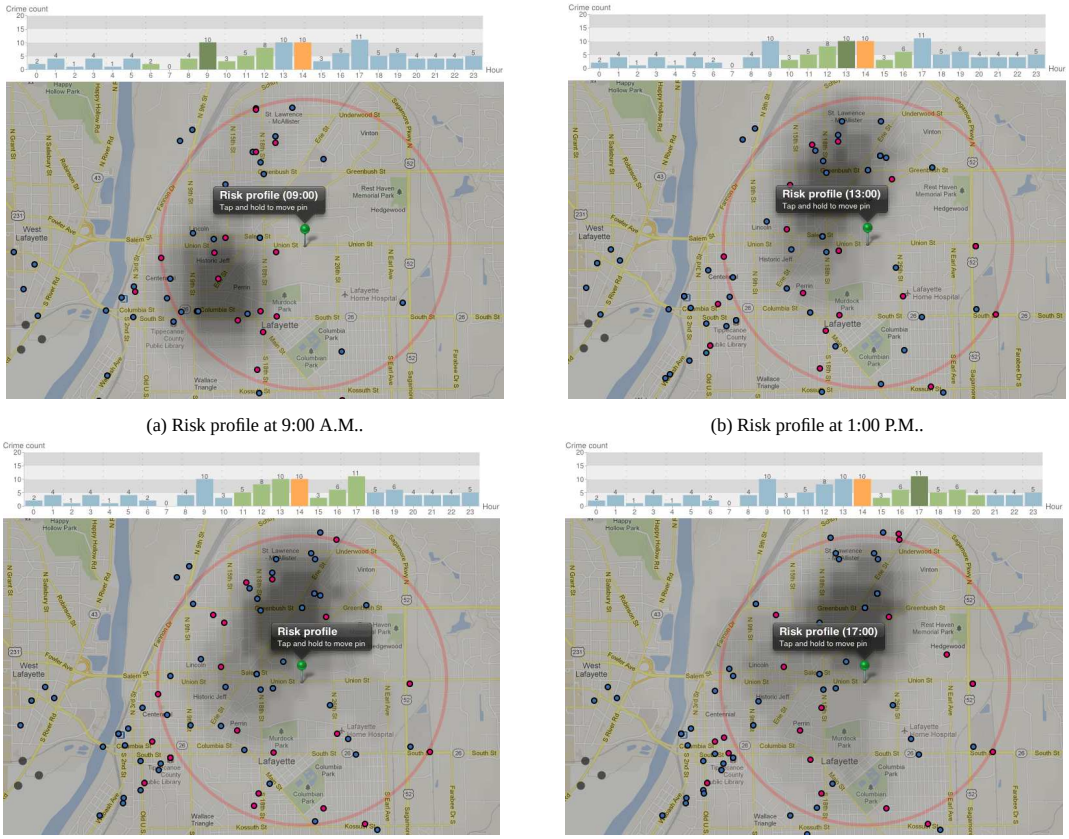
Furthermore, as is the case with most multivariate datasets, the CTC dataset is often incomplete. For example, many of the incidents do not contain valid geolocation data, causing uncertainty in the analysis process. In order to account for the uncertainty caused by this incompleteness in the dataset, we provide the ratio of correct incidents as a percentage value to show the accuracy of the visualization. This becomes important for users to accurately extract information from their dataset [22].

4.4 Temporal Displays and Analysis

Our system provides users with three temporal displays that allow them to visualize the temporal distribution of their datasets. We provide a time series display that presents the distribution of the incidents over time as a line graph, a calendar view visualization that lays the temporal data in the format of a calendar and a clock view that visualizes the hourly distribution of incidents.

4.4.1 Time Series View

The system allows users to simultaneously select multiple offenses and displays them as time series line graphs highlighting the trends between multiple datasets (Figure 2(b)). Furthermore, we provide



(a) Risk profile at 9:00 A.M..

(b) Risk profile at 1:00 P.M..

(c) Risk profile at the user's current time (2:00 P.M.).

(d) Risk profile at 5:00 P.M..

Figure 3: Risk profile visualizations at different times of day. The selected hour is shown by the dark green colored histogram bin, and the current time is shown as the orange colored bin. The ± 3 window from the current selected hour are shown as the green colored histogram bins, and the rest of the histogram bins is colored as blue.

users with several temporal aggregation options to aggregate their datasets. For example, users may choose to aggregate the incidents by day, week, month or year, and visualize the results dynamically. We further note that the time series display is interactive, allowing users to touch the screen to get the incident count at any particular time. A time series tape measure tool [26] in the graph view allows users to determine the temporal distance between any two points on the graph. We also provide users with tools that allow them to accurately model the data using an Exponentially Weighted Moving Average (EWMA) control chart for event prediction [26].

4.4.2 Calendar View

We adopt the calendar view visualization developed by van Wijk and Selow [43] to provide a temporal overview of the data in the format of a calendar, allowing users to visualize the data over time. Each date entry is colored on a sequential color scale that is based on the overall yearly trend to show the relative count of incidents for each day with respect to the maximum daily count over that calendar year.

To account for the smaller screen space of mobile devices (*C4*), we provide an overview+detail view. This is shown in Figure 2(c), where the left portion of the calendar view shows the weekly overview of the entire calendar. The overview display draws the individual rows based on the selected aggregation level (e.g., week, month) and are colored on a sequential color scale to reflect the weekly count of incidents. Users may tap on any portion of the calendar overview, updating the calendar view visualization dynamically to the position touched by the user. The overview+detail calendar view allows users to quickly identify weeks of high activity, and provides an easy way to scroll to them.

4.4.3 Clock View

In order to visualize the hourly distribution of the CTC incidents, we implement a clock view (Figure 2(d)) that organizes the data in the format of a clock. The clock view is a radial layout divided into 24 slices that are colored on a sequential color scale [10] to reflect the number of incidents that occur during each hour of a day.

As is the case with geospatial data (Section 4.3), many of the incidents do not contain a valid time field, causing uncertainty in the analysis process. Thus, we use the same method of displaying uncertainty for geospatial data in this domain, by showing the accuracy of the clock view visualization.

4.5 Risk Profile

The risk profile visualization shows the spatial and temporal distribution of incidents with respect to the current location and time of the user.

Our system utilizes the GPS feature of the mobile device and factors in the geospatial location of the user, the current time, and the historic CTC incidents occurring within their neighborhoods to provide estimates of CTC activity in their surroundings. These provide users with an overview of all the incidents and allows them to increase their level of situational awareness of the safety risks involved in their surroundings. Now, in order to show the spatial distribution of historic incidents, we utilize the map view and plot the incidents as points on the map. When users enable the risk profile feature, the system shows the current location of the user as a green colored pin on the map, and draws a circle (of a user-controlled radius) around their current location. The system then performs a query to the server and acquires all incidents that occur within this circle within a ± 3 hour offset (adjustable by the user)

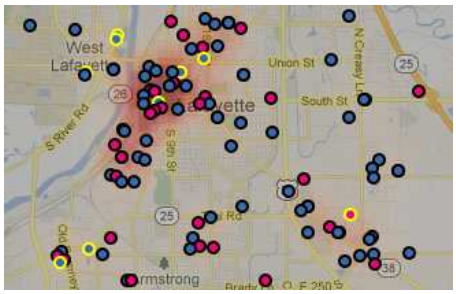


Figure 4: Incidents with nearby localized high-frequency activity are highlighted in yellow to alert users of areas with suspicious criminal activity.

with respect to the current time, for a date range specified by the user. The resulting incidents are then displayed as points on the map. Also, the temporal distribution of all incidents falling within the circle is shown in an interactive bar graph (Figure 3a (top)). We highlight the hours on this graph to reflect the hours on which the incidents are being displayed. Note that in addition to visualizing the risk profile for the current location and time, users may also choose to generate risk profiles for any desired spatial locations (by dragging the pin provided on the map) and for any hour of the day (by selecting a time by tapping on the risk profile time series graph (Figure 3a (top))). The users can then get a geotemporal overview of risk at any location and time.

In order to show which incidents have happened nearer in *time* (with respect to the current time of day), we modify the kernel density estimated heatmap (as described in Section 4.3) to encode the temporal distance from the current time. So in this case, the heatmap gives more weight to those incidents that fall closer to the current time within the ± 3 hour window, than to those that are farther away from the current time. The hotspots that so emerge provide an estimate of those incidents that happen closest in time with respect to the current time. An example of this approach is shown in Figure 3, where the user is visualizing all offenses against person and property for the month of February 2013. We can see hotspots emerging in different locations that show the distribution of incidents that happened closest to the current selected time.

4.6 Hotspot Alert

Our system also provides a feature to help users identify unusual localized high-frequency patterns of crimes and identify crime hotspot locations. Each data entry in the database is checked for other crimes with similar properties (defined by the user). This is done within a 200 meter block radius of the incident location, and for a 14 day period that extends from the day the incident occurred backwards in time. The system then highlights the incidents with the most number of similar incidents within the space-time window (Figure 4) which is updated dynamically as new data is entered into the database by the user. This alerts law enforcement officials of higher probability regions with nearby localized suspicious criminal activity and allows for more effective resource allocation and patrol planning.

4.7 Chemical Plume Modeling

Our system also provides a dynamic chemical plume modeling tool that provides law enforcement officers and first responders with better situational awareness in emergency situations resulting from chemical releases. Our system uses the ALOHA (Areal Locations of Hazardous Atmospheres) [1] software for chemical dispersion modeling and generating threat zones to assess the potential hazards caused by chemical releases. ALOHA is part of the CAMEO software suite [11] and can be used as a standalone desktop program.

Figure 5: Plume visualization. (left) The plume visualization displaying the different levels of danger zones. (right) The plume visualization showing census tracts colored on a sequential color scale to encode the number of people affected in each census tract, with respect to the maximum people affected in any tract.

In the case of a hazardous chemical release, our system can display the threat zones (plume) as a geospatial visualization. Each threat zone defines an area where the hazard level exceeds some 'Levels of Concern' (LOC) [2]. The concentration of chemicals is measured in ppm (parts per million) and each LOC specifies an area with a certain range of the released chemical concentrations in ppm. The innermost zone shows the area of most hazard and the outer zone shows the area of least hazard. The number of zones shown on the map can be changed by changing the number of LOCs, but is set to three LOCs by default [2].

The threat zones are plotted on top of a geographic map to help first responders quickly identify the regions under the most threat and also the safest roads for travel and evacuation. The threat zones are dynamic in nature due to changing weather conditions, and their visualization on the mobile system can be updated regularly by the user. Our system allows two types of plume threat zone visualizations to be shown. Figure 5 (left) shows the first visualization where the plume is visualized as three threat zones, colored red, orange and yellow to encode the levels of danger from most danger to least danger. The visualization also takes into account the changing weather conditions and shows the plume model confidence interval, represented by the outermost threat zone boundary (dark outline). By default, only the confidence line of the outermost threat zone is shown, but the user can also choose to show confidence lines for each individual threat zone.

In the second visualization (Figure 5 (right)), we display color-coded census tracts to show the distribution of people affected by the chemical plume in addition to the plume LOC visualization. For the color encoding, we first calculate the ratio of the census tract area covered by the plume. By assuming uniform population distribution, we calculate the expected number of people affected by the chemical release in each census tract by multiplying this ratio with the census population data. We take the number of people affected in each census tract and normalize this number by the maximum number of people affected in any tract. This obtained value is used to pick a color from a sequential color scale to provide information about the census tracts that have larger percentages of people in danger due to the chemical release. Detailed statistics of the affected people can also be shown.

4.8 Implementation Notes

Our mobile system front-end has been developed for the iPad/iPhone environment on the iOS platform. The system requires iOS version 5.0 onwards and is developed in the Objective-C environment using XCode 4. The PROJ.4 Cartographic Projections Library [35] is used to switch between map projection systems and the Google Maps library is used for routing.

On the back-end of our server-client architecture, the database is managed by MySQL that enables data querying and the web service that handles data requests by the client uses PHP. Data is transmitted securely (*C5*) using HTTP over SSL/TLS in XML or plain text format to keep the data format simple and generic for other uses.

5 User Evaluation

In this section, we provide user feedback of our mobile visual analytics law enforcement system from domain experts from our local law enforcement agencies. For the purpose of evaluating our system, we formed a focus group of 15 domain experts consisting of several police departments in our local county as well as the county legal offices. We deployed our mobile system to this group in December 2012 and have made continuous modifications to the system based on their responses and feedback obtained from our meetings which include several demonstrations, structured interview questionnaires, and informal in-person discussions. Here, we report some of the group's responses over the past 9 months.

We have received generally very positive feedback from this group. At a high level, officers, especially the chief of police in our local community, saw great value and potential of the system for certain daily tasks. He cited one use of this system to increase public awareness of their safety from the criminal activities that occur in their surroundings during safety campaigns. He also mentioned its use to visualize the impacts of active crime prevention actions (e.g., resource allocation for patrol, public safety campaigns).

At another level, the shift supervisor responsible for overseeing the patrol officer assignments during their shift periods and communicating information of the previous shifts to the incoming officers viewed the application as a major improvement to their existing tools used for briefings and roll call (e.g., print outs of incident reports from the previous 24 hour period). The features within our system that allow the display of the heat map of incidents displayed provided an added level of focus for incident location. The feedback provided by the supervisor indicated that they utilized the temporal reporting features in conjunction with geospatial features to assist in resource allocation planning. They also found the convenience of having a mobile version useful for their roll call briefings that allows them to look at the past incidents and have discussions outside of the office space for instance while having coffee in the lounge.

Patrol officers saw the value of using the system while on the field to get CTC incidents data in a visual form to have an increased situation awareness. Furthermore, the officers particularly liked the risk profile feature that provides the regions with historically higher incident levels based on the patrol officers current position and time. They indicated that this feature was especially useful as it factored in the current time-of-day information while they were patrolling their area of responsibility. The Mobile Computer Terminal (MCT) systems in their patrol cars are currently more geared towards reactive policing for reacting to an emergent situation, and as such, our application provides them with tools that enable proactive policing. The officers, however, also indicated that our system should be better integrated with the functions provided by their MCTs (e.g., dispatch calls, ability to enter detailed reports to be stored directly in their database) for a more effective tool. They also indicated a need for voice controlled commands for performing common functions.

During discussions with law enforcement personnel, additional information would be useful to officers responding to incidents (e.g., protective orders, non-contact orders). However, this information does not currently reside within their systems and, as such, the officers do not get access to such data while responding to incidents. We plan on incorporating these datasets in our system to assist the law enforcement personnel with operational decisions when responding to service calls.

At the role level of detectives and crime analysts, the usage patterns begin to shift more towards the predictive end of the spectrum as well as an increased usage of the temporal reporting features. Detectives are responsible for, among other things, solving crimes by, for example, investigating crime patterns; whereas, crime analysts provide insights and analysis into patterns and trends in crime to their police departments. Existing analysis tools tend toward a list generating, selection/filtering tool. Geospatial mapping also seems to appear to be prominently used at the present time as an analysis tool for performing hotspot analysis. The overall graphical user interface of our system looks promising for this role because of the ease of use and the widely known mobile application selection, zoom, and positioning conventions. One feature that appears to be the most promising is the ability to dynamically create regions or utilize shape files for geographic boundary definition. These features allow the crime analyst and/or detective to zero in on specific geographic areas of interest, like neighborhoods or law beats, without having to manually sort through data. These features also assist in generating historical reports for use in detailed crime analysis or investigation. The detective/crime analyst also found the selection capability within the calendar tool useful for seasonal crime analysis.

From our engagements, an intelligence analyst responsible for providing intelligence and planning for strategic, operational or tactical situations also saw value in the system for his day-to-day tasks. He particularly liked the ability of the system to apply any desired filters on-the-fly in a visual interface in order to narrow down his investigative analysis and to observe any spatial or temporal patterns. He was also interested in the crime monitoring process with the system and the potential to quickly identify incidents that were viewable by security cameras and pull up live and historical feeds by those cameras. Additionally, he suggested adding a layer of surveillance camera locations along with their range and angle of view information in the system for use in identifying incidents that may have been captured by any of the cameras for investigative purposes.

6 Conclusions and Future Work

In this paper, we have presented our mobile visual analytics system that has been designed to equip law enforcement personnel and, in the future, citizens with effective situation awareness and risk assessment tools. Our current work demonstrates the benefits of visual analytics in the mobile domain, and shows the effectiveness of providing users with on-the-go tools that allow them to make effective decisions. We have learned that the use of a mobile risk assessment system differs slightly in providing real-time situation awareness from a full, thorough analysis of CTC incidents with a full-fledged system on a desktop. From our interaction with our end users, we know that they do not have the time to tinker with a technology: it has to just work. Also with law enforcement agencies, they are trained to handle risky situations and would not rely on the system to gauge risk and determine if backup is needed. The use of the system is also limited in tactical situations where attention in certain actions takes precedence (e.g., focusing on the road while driving). In this case, a technology like speech recognition would significantly help but this remains a research direction we have not yet explored. Finally, as has been shown in this paper, we argue that there is great potential for the use of mobile visual analytics solutions to serve multiple role levels in the law enforcement and other related domains.

Future work includes adding advanced analytical and predictive capabilities into the system. Furthermore, we plan on incorporating image-based glyphs to represent the incidents on the map for easier identification of the charges associated with the incidents. We also plan on enhancing our routing methodology to factor in the risk and other parameters in order to allow users to plan safer routes. In addition, we think that incorporating multiple datasets (e.g., census data, street light locations, court records, weather) in the system would be an interesting research direction to explore the correlations between CTC incidents and other datasets. We also plan on modifying the kernel density estimation technique to further incorporate road network information to factor in for incident types that are road bound (e.g., traffic accidents). Finally, we plan on investi-

gating porting the system to other mobile platforms (e.g., Android, Windows RT).

ACKNOWLEDGEMENTS

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003. Jang's work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170). We would like to thank the reviewers for their valuable suggestions and comments, which helped to improve the presentation of this work. We also would like to thank Shantanu Joshi for his contribution in the project.

REFERENCES

[1] ALOHA. Internet: `http://response.restoration.noaa.gov/aloha`, [Mar. 23, 2012].

[2] Ask Dr. ALOHA: Choosing toxic levels of concern. Internet: `http://archive.orr.noaa.gov/book_shelf/1475_ToxicLOCs.pdf`, [Mar. 23, 2012].

[3] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M. Kraak, A. MacEachren, and S. Wrobel. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science*, 21(8):839–857, 2007.

[4] G. Andrienko, N. Andrienko, and H. Voss. GIS for everyone: the CommonGIS project and beyond. In M.Peterson, editor, *Maps and the Internet*, pages 131–146. Elsevier Science, 2003.

[5] G. L. Andrienko and N. V. Andrienko. Interactive maps for visual data exploration. *International Journal of Geographic Information Science*, 13(4):355–374, 1999.

[6] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer - Verlag, 2006.

[7] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

[8] L. Anselin, I. Syabri, and Y. Kho. GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22, Jan. 2006.

[9] P. Baudisch and C. Holz. My new PC is a mobile phone. *ACM XRDS*, 16(4):36–41, June 2010.

[10] C. A. Brewer. *Designing Better Maps: A Guide for GIS users*. ESRI Press, 2005.

[11] CAMEO software suite. Internet: `http://response.restoration.noaa.gov/oil-and-chemical-spills/chemical-spills/response-tools/cameo-software-suite.html`, [Mar. 23, 2012].

[12] S. Chainey and J. Ratcliffe. *GIS and Crime Mapping*. Mastering GIS: Technology, Applications & Management. John Wiley & Sons, 2006.

[13] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey, and Z. Huang. Coplink: visualization for crime analysis. In *Proceedings of the Conference on Digital Government Research*, pages 1–6, 2003.

[14] CrimeMapping. Internet: `http://www.crimemapping.com`, [Mar. 23, 2012].

[15] CrimeReports. Internet: `http://www.crimereports.com`, [Mar. 23, 2012].

[16] J. Dykes, A. MacEachren, and M.-J. Kraak. *Exploring Geovisualization*. Elsevier, 2005.

[17] M. Endsley and D. Garland. *Situation awareness: analysis and measurement*. Lawrence Erlbaum Associates, 2000.

[18] M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[19] G. Hamza-Lup, K. Hua, M. Le, and R. Peng. Dynamic plan generation and real-time management techniques for traffic evacuation. *IEEE Transactions on Intelligent Transportation Systems*, 9(4):615–624, Dec. 2008.

[20] C. Harrison. Appropriated interaction surfaces. *Computer*, 43(6):86–89, 2010.

[21] T. Kapler and W. Wright. GeoTime information visualization. *Information Visualization*, 4(2):136–146, Jun. 2005.

[22] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the information age: Solving problems with Visual Analytics*. EuroGraphics, 2010.

[23] S. Kim, R. Maciejewski, K. Ostmo, E. J. Delp, T. F. Collins, and D. S. Ebert. Mobile analytics for emergency response and training. *Information Visualization*, 7(1):77–88, 2008.

[24] O. Linda and M. Manic. Online spatio-temporal risk assessment for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):194–200, Mar. 2011.

[25] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland, and D. S. Ebert. Forecasting hotspots – a predictive analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):440–453, April 2011.

[26] A. Malik, S. Afzal, E. Hodgess, D. Ebert, and R. Maciejewski. VACCINATED - visual analytics for characterizing a pandemic spread VAST 2010 mini challenge 2 award: Support for future detection. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 281–282, Oct. 2010.

[27] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert. Visual analytics law enforcement toolkit. In *Proceedings of IEEE Conference on Technologies for Homeland Security*, pages 222–228, 2010.

[28] Metropolis police service crime mapping. Internet: `http://maps.met.police.uk`, [Mar. 23, 2012].

[29] Oakland crimespotting. Internet: `http://oakland.crimespotting.org`, [Mar. 23, 2012].

[30] A. Pattath, D. S. Ebert, R. A. May, T. F. Collins, and W. Pike. Real-time scalable visual analysis on mobile devices. *Multimedia on Mobile Devices*, 6821(1):682102, 2008.

[31] A. Pattath, D. S. Ebert, W. Pike, and R. A. May. Contextual interaction for geospatial visual analytics on mobile devices. *Multimedia on Mobile Devices*, 7256:72560H, 2009.

[32] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *Proceedings of IEEE Symposium on Information Visualization*, pages 219–224, 2005.

[33] Police.uk. Internet: `http://www.police.uk`, [Mar. 23, 2012].

[34] Z. Pousman, J. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, Nov.-Dec. 2007.

[35] PROJ.4 cartographic projections library. Internet: `http://trac.osgeo.org/proj/`, [Mar. 23, 2012].

[36] R. Roth, K. Ross, B. Finch, W. Luo, and A. MacEachren. A user-centered approach for designing and developing spatiotemporal crime analysis tools. In *Proceedings of GIScience*, Zurich, Switzerland, 2010.

[37] A. Sanfilippo, R. May, G. Danielson, B. Baddeley, R. Riensche, S. Kirby, S. Collins, S. Thornton, K. Washington, M. Schrager, J. Van Randwyk, B. Borchers, and D. Gatchell. InfoStar: An adaptive visual analytics platform for mobile devices. In *Proceedings of the ACM/IEEE conference on Supercomputing*, pages 74–83. IEEE Computer Society, Nov. 2005.

[38] Seattle data. Internet: `https://data.seattle.gov/`, [Nov. 5, 2013].

[39] Spotcrime. Internet: `http://www.spotcrime.com`, [Nov. 4, 2013].

[40] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[41] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[42] Uniform Crime Reports, Federal Bureau of Investigation. Internet: `http://www.fbi.gov/about-us/cjis/ucr/ucr`, [November 28, 2011].

[43] J. J. van Wijk and E. R. van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of IEEE Symposium on Information Visualization*, pages 4–9, 1999.

[44] J. Wood, J. Dykes, A. Slingsby, and K. Clarke. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183, 2007.

A Visual Analytics Process for Maritime Resource Allocation and Risk Assessment

Abish Malik * Ross Maciejewski*, *Member, IEEE* Ben Maule† David S. Ebert*, *Fellow, IEEE*

*Purdue University Visualization and Analytics Center (PURVAC) †United States Coast Guard

ABSTRACT

In this paper, we present our collaborative work with the U.S. Coast Guard's Ninth District and Atlantic Area Commands where we developed a visual analytics system to analyze historic response operations and assess the potential risks in the maritime environment associated with the hypothetical allocation of Coast Guard resources. The system includes linked views and interactive displays that enable the analysis of trends, patterns and anomalies among the U.S. Coast Guard search and rescue (SAR) operations and their associated sorties. Our system allows users to determine the potential change in risks associated with closing certain stations in terms of response time, potential lives and property lost and provides optimal direction as to the nearest available station. We provide maritime risk assessment tools that allow analysts to explore Coast Guard coverage for SAR operations and identify regions of high risk. The system also enables a thorough assessment of all SAR operations conducted by each Coast Guard station in the Great Lakes region. Our system demonstrates the effectiveness of visual analytics in analyzing risk within the maritime domain and is currently being used by analysts at the Coast Guard Atlantic Area.

Keywords: Visual analytics, risk assessment, Coast Guard

1 INTRODUCTION

As modern datasets increase in size and complexity, it becomes increasingly difficult for analysts and decision makers to extract actionable information for making effective decisions. In order to better facilitate the exploration of such datasets, tool sets are required that allow users to interact with their data and assist them in their analysis. Furthermore, such datasets can be utilized to explore the consequences and risks associated with making decisions, thereby providing insights to analysts and aiding them in making informed decisions.

Besides the sheer volume and complexity of such datasets, analysts must also deal with data quality issues, including uncertain, incomplete and contradictory data. Moreover, analysts are often faced with different decisions and are required to weigh all possible consequences of these decisions using such datasets in order to arrive at a solution that minimizes the associated risks within a given time constraint. Using traditional methods of sifting through sheets of data to explore potential risks can be highly inefficient and difficult due to the nature and size of these datasets. Therefore, advanced tools are required that enable a more timely exploration and analysis. Our work focuses on the use of visual analytics [17, 31] in the realm of risk assessment and analysis and demonstrates the effectiveness of visual analytics in this domain. The work described in this paper is based on the application of visual analytics to analyze historic response operations and assess the potential risks in the maritime environment based on notational station closures. Our work was done in collaboration with the U.S. Coast Guard's Ninth District and Atlantic Area Commands that are responsible for all Coast Guard operations in the five U.S. Great Lakes. In particular, we focused on the Auxiliary stations that are staffed by Coast Guard volunteers and civilians. These Auxiliary stations assist their parent stations in their operations and usually operate on a seasonal basis using a small fleet of boats for conducting their operations. However, the number of Auxiliary personnel that volunteer their time at these stations has decreased over recent years. This has required Coast Guard analysts to develop possible courses of action and analyze the risks and benefits with each option. Several options include seasonal or weekend only staffing of these units, or at worst, closure. Closure, however, may involve increased risks to the boating public and a complete analysis of the risks associated with closing an Auxiliary station needs to be evaluated. The results of this type of analysis would assist the decision makers in determining the optimal course of action.

In particular, the analysts are interested in determining the spatial and temporal distribution of response cases and their associated sorties (a boat or an aircraft deployed to respond to an incident) for all SAR operations conducted in the Great Lakes and how closing certain Auxiliary stations affects the workload of the stations that absorb these cases. Coast Guard policy mandates the launch of a sortie within 30 minutes and have an asset (boat or aircraft) on scene within two hours of receiving a distress call [32]. Closing these stations implies a potential for longer response times that could potentially translate into the loss of lives and property.

To address these challenges, we developed a visual analytics system that supports decision making and risk assessment and allows an interactive analysis of trends, patterns and anomalies among the U.S. Coast Guard's Ninth District operations and their associated sorties. Our system, shown in Figure 1, allows enhanced exploration of multivariate spatiotemporal datasets. We have incorporated enhanced tools that enable maritime risk assessment and analysis. Our system includes linked spatiotemporal views for multivariate data exploration and analysis and allows users to determine the potential increase or decrease in risks associated with closing one or more Coast Guard stations. The system enables a thorough assessment of all operations conducted by each station. In addition, the system provides analysts with the tools to determine which Coast Guard stations are more optimally suited to assume control of the operations of the closed station(s) by comparing the distances from available stations to all SAR cases previously handled by the closed station(s). Our system features include the following:

- Risk profile visualizations and interactive risk assessment tools for exploring the impact of closing Coast Guard stations
- Optimization algorithms that assist with the interactive exploration of case load distribution in resource allocation
- Linked filters combined with spatial and temporal views for interactive risk analysis/exploration

Our work focuses on providing analysts with interactive visual analytics tools that equip them to deal with risk assessment scenar-

*e-mail: {amalik|rmacieje|ebertd}@purdue.edu
†email: Ben.J.Maule@uscg.mil

IEEE Symposium on Visual Analytics Science and Technology
October 23 - 28, Providence, RI, USA
978-1-4673-0014-8/11/$26.00 ©2011 IEEE

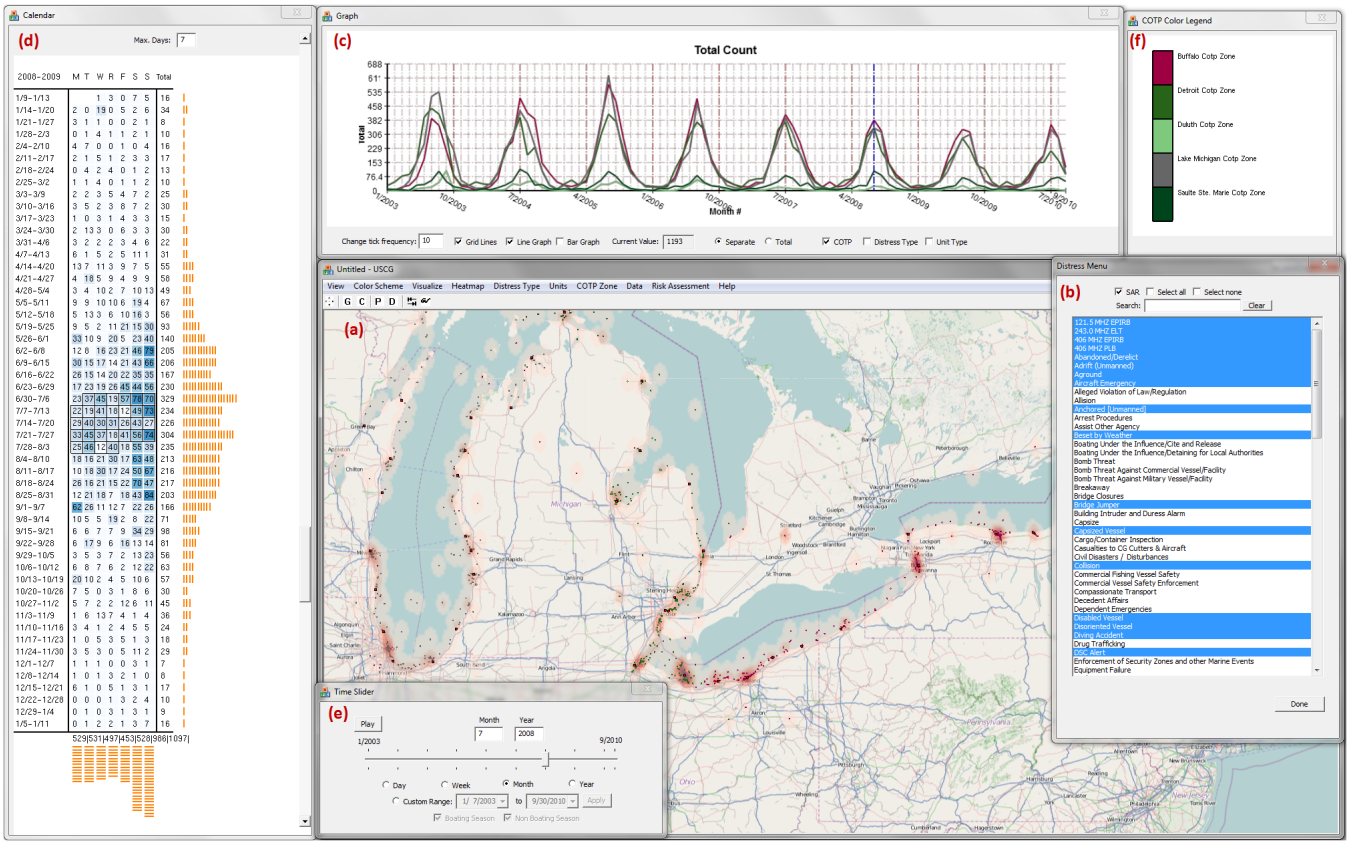


Figure 1: A screenshot of our risk assessment visual analytics system. Here, the user is visualizing all search and rescue (SAR) operations conducted by the U.S. Coast Guard in the Great Lakes region in July 2008. The main viewing area (a) shows the map view with the circles showing the locations of SAR incidents in the Great Lakes. The right-most window (b) shows an interactive menu showing all distress types with SAR cases selected in blue. The top window (c) shows the time-series and the left window (d) shows the calendar views of the SAR incident report data. The bottom-left window (e) shows the time slider with radio buttons that allow different temporal aggregation levels. A legend for all District Nine maritime zones is shown in the upper right (f).

ios associated with closing Coast Guard stations. We emphasize that although our risk assessment toolkit and the examples given in this paper have been based in the maritime domain, these techniques apply equally as well to other domains (e.g., criminal offense analysis, syndromic surveillance).

2 Related Work

In recent years, there has been a rapid growth in the development of new visual analytics tools and techniques for advanced data analysis and exploration (e.g., [30, 34]). From traditional scatterplots [8] and parallel coordinate plots [16] to tools like Theme River [15] and spiral graphs [7], these systems incorporate different forms of visualizations to provide enhanced analytical tools to users. Although these tools allow users to explore their data and assist them in their decision making process, researchers have only recently started to employ visual analytics techniques for risk-assessment and decision-making domain that allow users to perform a thorough analysis of risks associated with different decisions.

Migut and Worring [21] propose an interactive approach to risk assessment where they demonstrate a risk assessment framework that integrates interactive visual exploration with machine learning techniques to support the risk assessment and decision making process. They use a series of 2D visualizations including scatterplots and mosaic plots to visualize numerical and ordinal attributes of the datasets. While the authors demonstrate the effectiveness of using visual analytics in the field of risk assessment, their work is mainly focused on building classification models that users may interactively use to classify their data entities and visualize the effects on their classification. Gandhi and Lee [13] also apply visual analytics techniques to the realm of requirements-driven risk assessment. Specifically, they use cohesive bar and arc graphs to illustrate the risks due to the cascading effects of non-compliance with Certification and Accreditation requirements for the U.S. Department of Defense. Sanusi and Mustafa [25] introduce a framework to develop a visualization tool that may be used for risk assessment in the software development domain. Their proposed framework allow users to identify the components of the software system that are likely to have a high fault rate. Direct visualizations of risk use tools like bar graphs and confidence interval charts to visualize measures of risk and are usually constructed using spreadsheet programs like Microsoft Excel [12, 13]. Although widely used, these techniques fail to work for our purposes primarily due to the nature of the risk analysis that is required. The Coast Guard SAR dataset is spatiotemporal in nature and the exploration of risk requires domain knowledge that is difficult to incorporate algorithmically.

With respect to the temporal nature of risk assessment, researchers have also developed different visualization systems that allow users to explore risks associated with financial decisions related with investments and mutual funds, among other financial planning scenarios. Rudolph et al. [24] propose a personal finance

decision making visual analytics tool that allows users to analyze both short-term and long-term risks associated with making investment decisions. Savikhin et al. [26] also demonstrate the benefits of applying visual analytics techniques to aid users in their economic decision making and, by extension, to general decision making tasks. Both of the previous examples only explore temporal datasets. In this work, we apply visual analytics techniques to explore risks using multivariate spatio-temporal datasets that guide analysts in making complex decisions.

As is the case with most multivariate datasets, data tends to be inherently unreliable, incomplete and contradictory. In order to reach to correct conclusions, analysts must take these into account in their analysis. In this regard, Correa et. al. [9] describe a framework that supports uncertainty and reliability issues in the different stages of the visual analytics process. They argue that with an explicit representation of uncertainty, analysts can make informed decisions based on the levels of confidence of the data. Our system factors data reliability issues in the risk assessment process and provides confidence levels at all stages of risk assessment that, in turn, enable analysts to better understand the underlying nature of the data and guides them in making effective decisions.

There also exist many geospatial and temporal analytical systems that provide users with the ability to explore their spatiotemporal datasets in order to find patterns and provide an overview of the data in a visual analytics platform (e.g., [1, 2, 3, 14]). As the needs of our end users are unique, this warrants developing a stand alone system to address the challenges faced by the Coast Guard analysts. We plan on further examining these robust geo-temporal analysis tools and the degree to which they can be extended to meet the Coast Guard requirements that have been identified in this paper.

There has also been much work done in visualizing large datasets using interactive cross-filtered and linked views that allow users to explore their datasets. Stasko et. al. [30] use multiple coordinated views of documents to reveal connections between entities across different documents. Eick and Johnson [10] utilize multiple linked views to visualize abstract, non-geometric datasets in order to reduce visual clutter and provide users with insights into their datasets. Eick and Wills [11] also demonstrate the effectiveness of linking and interaction techniques in the visualization of large networks. Our system utilizes these practices and allows users to interactively explore their multi-dimensional and multi-attribute datasets using a series of multi-coordinated linked views.

Researchers have also explored different methods to address the challenges posed to maritime security and safety. Willems et. al. [34] introduce a novel geographic visualization that supports coastal surveillance systems and decision making analysts in gaining insights into vessel movements. They utilize density estimated heatmaps to reveal finer details and anomalies in vessel movements. Scheepens et. al. [27] also present methods to explore multivariate trajectories with density maps and allow the exploration of anomalously behaving vessels. Lane et. al. [18] present techniques that allow analysts to discover potential risks and threats to maritime safety by analyzing the behavior of vessel movements and determining the probability that they are anomalous. Some other models for anomaly detection in sea traffic can be found in [19, 22]. Researchers have also proposed several approaches to maritime domain awareness. For example, Roy and Davenport [23] present a knowledge based categorization of maritime anomalies built on a taxonomy of maritime situational facts involved in maritime anomaly detection. We observe that these methods and models may help in risk analysis and understanding the impact of weather and varying speeds of Coast Guard vessels in the Great Lakes to identify high risk regions.

There has also been much work done to assess and mitigate risks to critical infrastructure and transportation in the maritime domain. Adler and Fuller [5] provide dynamic scenario- and simulation-based risk management models to assess risks to critical maritime infrastructure and strategies implemented for mitigating these risks. Mansouri et. al. [20] also propose a risk management-based decision analysis framework that enables decision makers to identify, analyze, and prioritize risks involved in maritime infrastructure and transportation systems. Their framework is based on risk analysis and management methodologies that allows understanding uncertainty and enables analysts to devise strategies to identify the vulnerabilities of the system. Furthermore, work has been done to quantify risks in the maritime transportation domain, a summary of which can be found in [29]. While these methods facilitate maritime infrastructure risk analysis, our work is focused on assessing maritime risks from multivariate spatiotemporal SAR data sets. In this paper, we present a visual analytics approach to maritime risk assessment and provide examples that demonstrate the advantages of applying visual analytics in this domain.

3 Visual Analytics Risk Assessment Environment

Our visual analytics system provides enhanced risk assessment and analytical tools to analysts and has been built to operate for SAR incident report data. Our system has been implemented in a custom Windows-based geographical information system that allows drawing on an OpenStreetMap map [4], using Visual C++, MySQL and OpenGL. The system displays geo-referenced data on a map and allows users to temporally scroll through their data. We provide linked windows that facilitate user interaction between the spatial and temporal domains of the data. We also provide advanced filtering techniques that allow users to interactively explore through data. In addition, we have adapted the calendar view presented by vanWijk and Selow [33] and extended it to explore seasonal and cyclical trends of SAR operations and also as means to filter data to support advanced analysis.

Figure 1 presents a screenshot of our system. The main viewing window (Figure 1 (a)) shows the map view where the user can explore the spatial distribution of all cases handled by the Coast Guard. We utilize density estimated heatmaps (Section 3.2) to quickly identify hotspots. Users may draw a bounding box over incident points on the map that generates a summary of all incidents enclosed by the box. We also provide tape measure tools that allow users to measure the distance between two points on a map. The top-most window (Figure 1 (c)) shows the time-series view of the data where multiple lines graphs can be overlaid for comparison and analysis. Users may visualize time-series plots by department, distress type and Coast Guard Captain of the Port (COTP) zone to explore summer cyclical patterns. The left-most window (Figure 1 (d)) shows the calendar view of the selected Coast Guard cases. The total number of columns on the calendar may be changed as desired to reveal seasonal trends and patterns. The bottom window (Figure 1 (e)) shows the time-slider widget that is used to temporally scroll through the data while dynamically updating all other linked windows. The radio buttons beneath the time slider provide several temporal aggregation methods for the data. The right-most window (Figure 1 (b)) shows the distress type menu where all SAR cases (highlighted in blue) have been selected for visualization. Users may select multiple distress types using this menu, dynamically updating all linked views. We use similar menus to filter cases by other data fields. Users may also interactively search the menu using the search box provided on top of the menu. Finally, the top-right window (Figure 1 (f)) shows an interactive legend of the different Coast Guard District Nine maritime zones. This legend allows users to click on any of the zones that highlights all cases falling in the zone by filling the circles on the map with a solid color and dimming out the other cases being displayed on the map.

A key feature of our system is the interactive distress, station and COTP zone filtering component. Users interactively generate combinations of filters that are applied to the data being visualized

through the use of menus (like the one shown in Figure 1 (b)) and edit controls. The choices of filters applied affects both the geospatial viewing region and all temporal plots.

3.1 Coast Guard SAR data

The SAR data is collected by all U.S. Coast Guard stations and stored in a central repository. When the Coast Guard is called into action, a response case is generated, usually by the maritime zone that has authority in that region that receives the distress call (referred to by the Coast Guard as the Search Mission Coordinator or SMC). Upon receiving the call, this authority will determine if resources will be applied, including which unit will provide the resource, the resource type and number. Therefore, a response case may generate zero, one, or many sorties to respond to an incident. While analyzing risks associated with the various mitigation options, including station closure, analysts are interested in analyzing the spatiotemporal distribution of both the response cases and their associated sorties.

The SAR data consists of two main components: (1) response cases and (2) response sorties. Each entry in the response case and sortie dataset contains information that provides details of the incidents (e.g., number of lives saved, lost, assisted) and contains the geographic location of the distress.

Uncertainty in decision making

As is the case with most large datasets, anomalies and missing data introduce errors and uncertainty. The SAR data is no exception. We find that many SAR cases do not have an associated geographic location, or have a wrong geographic location associated with them. These inherent errors in data affect the spatial probability estimates and introduce a certain amount of uncertainty in the decisions that must be considered for an effective risk analysis and assessment. As noted in [17], visual analytics methods help people make informed decisions only if they are made aware of data quality problems. In this regard, we incorporate uncertainty and confidence levels associated with the SAR dataset in our visualizations by displaying the accuracy of the results at each step of the risk assessment process. This is shown as a percentage that shows the total cases with reliable data that can be used in the decision making process (Figure 2). This percentage is calculated by using the following formula:

$$Accuracy = \frac{N-G}{N} \times 100 \tag{1}$$

Here, N is the total number of cases and G is the number of cases with unreliable values (e.g., unknown geographic coordinates, swapped negative signs). When such errors are not obvious, the data is assumed to be correct and is displayed to the analyst on the map. The analyst can further report errors in the data and contribute to the data cleaning process.

3.2 Geospatial displays

Our system provides analysts with the ability to plot incidents as points on the map and as density estimated heatmaps (Figure 1 (a)). In addition, we provide users with the option of coloring each incident circle with a color on a sequential color scale [6] that represents its data value. For example, users may choose to visualize the average response time to respond to an incident for all SAR cases on the map and identify cases with higher response times. Furthermore, to explore the spatial distribution of the SAR cases and quickly identify hotspots, we employ a modified variable kernel density estimation technique (Equation 2) that scales the parameter of estimation by allowing the kernel scale to vary based upon the distance from the point X_i to the kth nearest neighbor x in the set comprising of N [28].

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{min_{(h,d_{i,k})}} K\left(\frac{x-X_i}{min_{(h,d_{i,k})}}\right) \tag{2}$$

Here, N is the total number of samples, $d_{i,k}$ is the distance from the i-th sample to the k-th nearest neighbor and h is the maximum allowed kernel width. We choose the maximum kernel width based on asset speed and travel time. Furthermore, we use the Epanechnikov kernel [28] (equation 3) to reduce calculation time:

$$K(\mathbf{u}) = \frac{3}{4}(1-\mathbf{u}^2)1_{(||u||\leq 1)} \tag{3}$$

where the function $1_{(||u||\leq 1)}$ evaluates to 1 if the inequality is true and to zero otherwise.

3.3 Time series displays

Along with the graphical interface, our system provides a variety of visualization features for both spatial and temporal views. For temporal views, we provide line and stacked bar graphs and calendar views to visualize time series SAR incident report data.

The line graph visualization allow users to overlay multiple graphs for easy comparison and to visualize trends. Both line graph and stacked bar graph visualizations are supported and can be interchanged using the radio buttons provided. Users may choose to visualize SAR cases handled by individual stations or maritime zones, or visualize them by distress types. The data is plotted based on a temporal aggregation level that the user selects on the time-slider widget (Figure 1 (e)). In Figure 1 (c), we show the line graph display of all SAR cases aggregated by month. We can easily observe peaks in the number of SAR cases in the summer months for all maritime zones in the Great Lakes region.

The calendar view visualization was first developed by van Wijk and Selow [33]. This visualization provides a means to allow the visualization of data over time, laid in the format of a calendar. In our implementation (Figure 1 (d)), we shade each date entry based on the overall yearly trend. Users may interactively change the total number of columns of the calendar thereby changing the cycle length of the calendar view, enabling users to explore both seasonal and cyclical trends of their datasets. The system also draws histograms for each row and column. This allows analysts to visualize weekday and weekly trends of SAR incidents and further assists them in determining an effective resource allocation scheme. Furthermore, we have modified our calendar view to support an interactive database querying method for easily acquiring summary statistics from the SAR database.

4 Risk Assessment process

In this section, we describe the different methods and techniques that we apply in the Coast Guard risk assessment process.

4.1 Problem description

To bound the problem, the Coast Guard analysts provided a series of questions for use in their analysis. These questions are briefly summarized below.

1. Assuming a maximum transit speed of 15 nautical miles per hour, how many cases occur per year in which a parent station could not have a surface asset on scene within two hours?
2. For each Auxiliary station, what are the types (by percentage) of SAR response cases occurring per year?
3. For each Auxiliary station, what is the temporal (by hour, month and day of week) distribution of the response case load?

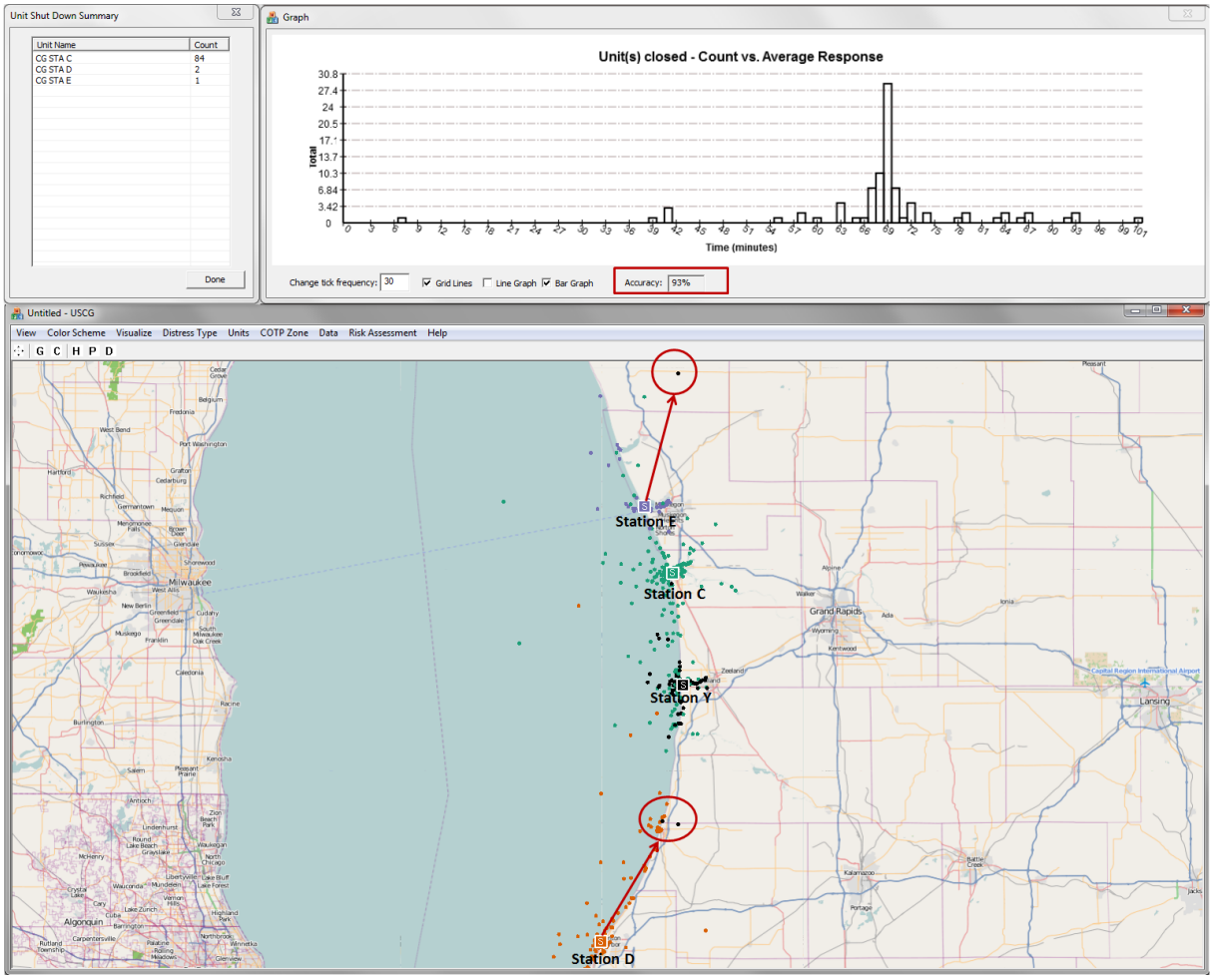


Figure 2: Average response time risk assessment when Auxiliary station Y is closed. The system automatically chooses the stations (shown in the upper-left window) that are optimally suited to respond to cases previously handled by station Y, along with a count of cases that each station absorbs. The station Y cases (black circles) to be handled by stations D and E are circled in red. The top graph shows the average response time distribution of these stations to respond to station Y's cases. We find that only 93% of the cases responded to by station Y have an associated geographic location.

4. What is the average annual case load that would be absorbed by each parent station in the absence of the Auxiliary station and what percentage increase would this represent to the parent station's annual case load?
5. Based on the historical data for all cases (SAR and others), what is the expected annual response case demand broken down by response type (i.e., Person in Water, Vessel Flooding, etc.)?
6. Assess the potential risks associated with closing certain Auxiliary stations in terms of additional case load absorbed, lives potentially lost, and other available factors.

Our visual analytics system was developed to assist the Coast Guard analysts in answering these questions and to model the potential risks of closing one or more Auxiliary stations. Furthermore, we allow analysts to explore the effects of closing multiple stations and provide a summary of stations that are most optimal to absorb the work load of the closed stations. Analysts may restrict the stations that absorb the work load of the closed stations to determine the stations that prove most effective, thereby informing optimal operational execution for the station that is nearest to respond to the distress case.

We perform our analysis under the assumption that the path between a station and a distress location is a straight line. While this assumption presents a best-case scenario to the analyst, discussions with our Coast Guard partners indicated this was an acceptable approximation as using channel and waterway information would result in a large computational overhead. With this assumption in place, if a station absorbing an Auxiliary station's cases increases the maritime risks in the region (e.g., if the average response time exceeds the two hour time limit for most SAR incidents), then closing the Auxiliary station could prove to be dangerous for the maritime and public safety of the region. This straight line approximation provides details on the best case scenario.

4.2 Average response time for SAR incidents

As stated before, a Coast Guard policy mandates the rescue resource to be on scene within two hours of a distress (e.g., disabled vessel, person in water). Given the cold water temperatures in the Great Lakes, even in the summer, increase in response time can potentially impact the success of a case. Therefore, given the option of closing a station, the analysts desire to know the nearest available resource to respond and calculate the time to respond to the scene. A typical Coast Guard vessel travels at a speed of 15 nautical miles

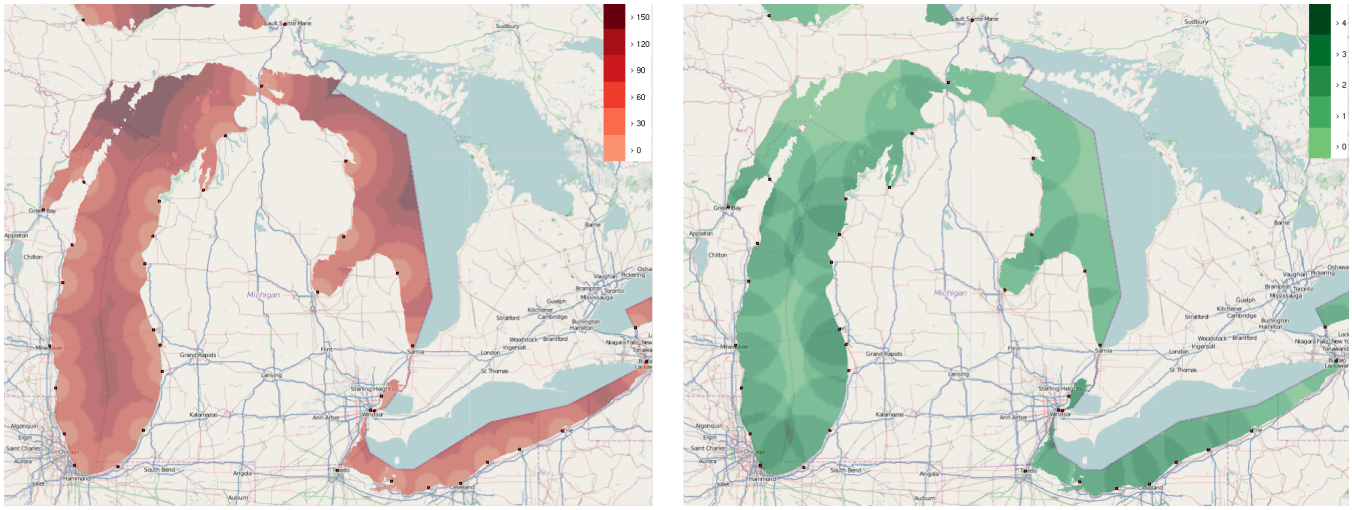


Figure 3: Risk profile. (Left) A heatmap showing the time taken (in minutes) by the Coast Guard stations to deploy an asset to the Great Lakes to respond to a SAR incident, assuming a speed of 15 nautical miles per hour. (Right) A heatmap showing the Coast Guard SAR coverage (i.e., the number of stations that respond to a particular region) in the Great Lakes. The squares along the coast show the locations of the stations.

per hour. After an Auxiliary station is closed, the parent station should still be able to reach most of the cases handled by the Auxiliary station within the two hour limit. In this section, we describe how our system can be used to determine the average response time for cases if a parent station (or any combination of stations) absorbs an Auxiliary station's cases.

In order to generate the average response time for the station(s) that absorb the work load of the closed station, we sift through all the incidents that the closed station handled and find the closest station (excluding the closed station) for each incident by comparing the distance between all stations and the incident. This distance between the closest station and incidents may also be visualized separately to reveal more details. Once the closest station is found, we obtain the time for an asset to reach the incident location using the distance formula $Time = Distance/Speed$. Users may also change the speed of the asset, changing the results dynamically.

We provide users with several filtering options while performing average response-time analysis. Users may choose to analyze the average response-time temporal distribution of incidents by applying any possible filters on distress type, department or maritime zone. Users may also analyze the distribution of only the non-SAR cases. Moreover, users may choose to close several stations all at once and model the resulting effects. They may also specify which stations absorb the cases of the closed stations and thus determine the stations best suited for closing and the optimal methods for reallocating available resources. We also note that our system can be easily modified to incorporate other risk metrics including, for example, normalizing SAR cases by the underlying population density, correlating SAR incidents with other parameters, etc.

Figure 2 shows the output generated when the analyst opts to close Auxiliary station Y. In this example, we examine all cases responded to by station Y between January 2004 and September 2010. The system automatically suggests the stations that should absorb Auxiliary station Y's cases along with the total number of cases that each station absorbs. We find that stations C (the parent station of Y), D and E absorb Auxiliary station Y's cases, with each absorbing 84, 2 and 1 cases, respectively. The analyst may instead select a specific station to absorb station Y's cases and analyze the results generated. In Figure 2, the map view shows all cases that each of the four stations responds to during this time period (shown as circles, with each case color coded by its station). We have also highlighted the two cases that station D and the one case that station E responds to in Figure 2. It may be noted that the one case absorbed by station E appears to be out of place (possibly due to a human error in entering the geographic coordinates for that particular case). The top-right bar graph shows the count of all SAR cases handled by station Y during this time period versus the average response time (in minutes) taken by the resulting stations to reach these cases, assuming a transit speed of 15 nautical miles per hour. From this time-series plot, we observe that all cases responded to by the Auxiliary station would fall well within the tolerance level of 120 minutes when the suggested stations take over. The system also determines the accuracy of the results dynamically by determining the number of cases that have no associated geographic coordinates. We find that 93% of the cases responded to by station Y in the time range January 2004 and September 2010 have an associated geographic coordinate (as seen from the accuracy percentage in Figure 2-top-right). Data integrity is a necessary parameter to report to the analysts and decision makers. Thus, the user is made aware of these uncertainties at every step of the risk assessment process.

4.3 Temporal distribution of response case load

One important aspect of risk assessment is analyzing the work load and distribution of response cases of the stations being analyzed over different temporal ranges. This becomes necessary to determine the feasibility of a station to be closed and to determine how the available resources may be reallocated (e.g., what times of day and what months would the stations need to have more personnel deployed). Analysts also use their domain experience and expertise to determine whether a particular station can absorb a closing station's cases. In particular, the Coast Guard officials were interested in understanding the hourly, daily and monthly trends of SAR cases occurring in the Great Lakes.

Using traditional methods of sifting through SAR datasets turns out to be highly inefficient for determining the temporal distribution of the SAR cases and, as such, advanced database querying tools are necessary to facilitate this process. To this end, we adapt the calendar view for querying the SAR database. We provide three different interaction methods within the calendar view widget (Figure 1 (d)) to obtain a detailed summary of response cases occurring over the selected date-range. Users can select date ranges by simply clicking on the start and end dates that selects all the dates between the two clicked dates. Users may also select one or more columns

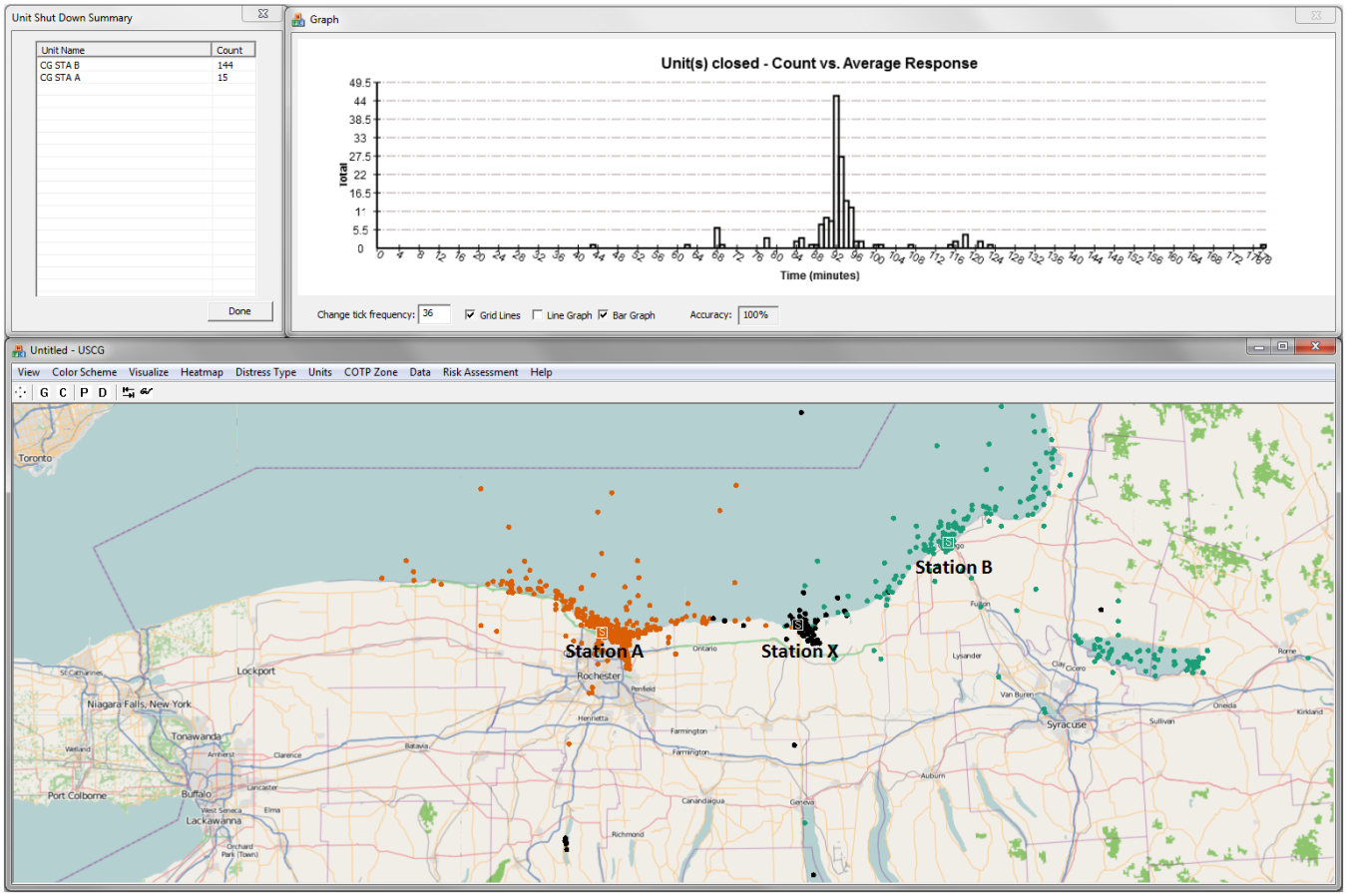


Figure 4: Risk assessment using linked views. Here, the analyst has chosen to close Auxiliary station X and is analyzing the risks associated with this decision. Each circle on the map represents a SAR case and has been color coded by its owner station. The system shows that stations A and B are best suited to absorb station X's cases, with station B absorbing 144 cases and station A absorbing 15 cases.

of the calendar to generate the summary statistics. This allows them to query the database and acquire the summary of events occurring, for example, only on a particular week day. Finally, users may select any combination of individual dates and obtain the summary of all selected response cases on those dates. These querying methods allow analysts to easily determine the temporal patterns of response cases over any date range. The system provides summary statistics of SAR incidents for all stations and includes the total number of lives saved, assisted, affected, total property damaged and saved and the count of all cases occurring over the selected date range. Users may select any date, row, column, or combinations thereof in the calendar view using the mouse to access the summary statistics. Furthermore, the system also allows users to visualize the hourly and monthly distribution of cases for any time period after all filters are applied.

4.4 Risk profile

Our system also provides users with the ability to interactively generate risk profiles that can be used to identify regions with little SAR coverage by the Coast Guard stations in the Great Lakes. Figure 3 illustrates the risk profile heatmaps that present an overview of the Coast Guard SAR coverage in the Great Lakes. Selected filter settings affect the visual output, and in this case, we are looking exclusively at small boat station coverage. When areas of low coverage exist, resources with additional capability (e.g., aircraft) are often provided to ensure coverage of all areas. Figure 3 (Left) shows the time (in minutes) that the Coast Guard stations would take to respond to a SAR incident in the Great Lakes, assuming a transit speed of 15 nautical miles per hour. This profile is generated assuming that the station closest to a location responds to an incident in the Great Lakes. The regions in the Great Lakes that take the longest time for the Coast Guard to respond to a SAR case can be clearly seen in this figure. Users may interactively close stations, filter on a different resource type (e.g., boat, aircraft), or change the asset speed, updating the risk profile interactively. This further enables the analysts to visualize the increase or decrease in risk when a station is closed. Moreover, analysts can set the lower threshold of the color scale to 120 minutes (or any arbitrary time), thereby allowing them to easily identify regions that may take more than 120 minutes to respond. We plan on incorporating contour lines into our system to demarcate the regions that may take more than the set threshold response time.

Figure 3 (Right) provides another risk profile visualization that allows officials to identify regions with low Coast Guard coverage for SAR operations in the Great Lakes. Regions with high SAR coverage by the Coast Guard stations are shown by darker colors. This further allows analysts to identify stations where resources may be reallocated without increasing maritime risk.

5 Exploring Risk Using Spatiotemporal Linked Views

While examining which Auxiliary stations are most suitable to close, analysts need to weigh all options and analyze the potential increase or decrease in associated risks. They must also consider

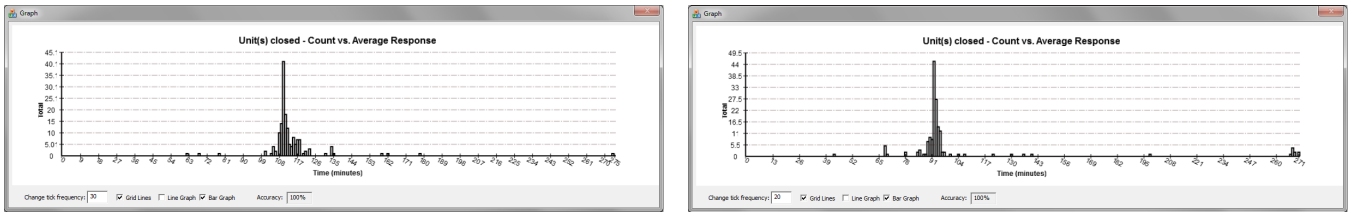


Figure 5: The average response time distribution for the additional cases when parent station A absorbs station X (Left) and when station B absorbs station X (Right). We see a left shift in the median time indicating that station B may be a better candidate to absorb station X's cases.

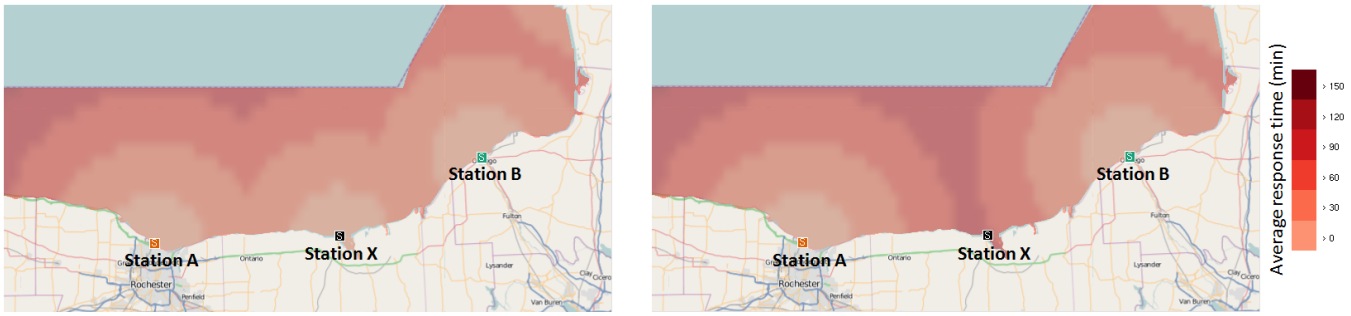


Figure 6: The change in average response time risk profile when Auxiliary station X is closed. (Left) Risk profile when station X is operational. (Right) Risk profile when station X is closed. We see an increase in risk in station X's area of operation when the station is closed.

the increase in workload of the stations that absorb the closed station's cases to effectively determine the optimal response of available resources. In this section, we describe a typical scenario where an analyst is trying to determine the risks associated with closing an Auxiliary station in one of the sectors in the Great Lakes and the stations that absorb the work load of this closed station.

Suppose the analyst chooses to close Auxiliary station X whose parent station is A. Since the parent station of X is station A, the analysts' first inclination may be to assign all cases to station A after station X is closed. The analyst first uses the system's calendar view and finds that the maximum number of cases that station X responds to in one day is 7 in the peak boating season. He also visualizes the hourly distribution of cases for station X and determines that the incidents are spread out during the day. Next, the analyst uses our risk assessment system to perform an average response time risk analysis over a time range of January 2006 to September 2010 and selects Auxiliary station X to be closed. Once station X is closed, the system automatically generates the result seen in Figure 4. As seen in the figure, the system determines stations A and B to be the optimal stations to respond to the cases handled by Auxiliary station X. In this figure, we see the spatial distribution of all SAR cases that the three stations responded to during this time range (seen as circles that are color coded by station). We observe that station B absorbs 144 additional SAR cases as opposed to parent station A which only absorbs 15 cases. Moreover, with this case load distribution, we find that 154 out of the total 159 cases are responded to within the 120 minute limit (as seen from the time series plot in Figure 4-top-right). These results suggest that station B is better suited to absorb most of the cases of Auxiliary station X.

In order to get a better picture, the analyst now restricts the stations that respond to the cases handled by station X to first, its parent station A, and then to station B and analyzes the average response time distribution of cases for each of the two stations separately. The results of this step are shown in Figure 5, with the left graph showing the average response time taken by station A, and the right graph corresponding to station B. As the analyst compares the two graphs, he realizes that if only station B is allowed to absorb station X, there are 14 cases that take more than 2 hours for the Coast Guard to arrive on scene. On the other hand, if station A is allowed to absorb station X, 16 cases take longer than 2 hours. However, as we can clearly see from Figure 5, station A takes a longer time to respond to most cases than station B, with the median time of station A being 110 minutes, and that of station B being 92 minutes. The analyst also notes that station B takes between 264-271 minutes to respond to about nine cases, whereas station A takes 275 minutes to respond to one case. To get a better understanding of why this may be happening, the analyst explores the spatial distribution of station X's cases on the map and discovers that some cases of station X get mapped to an inland lake (which requires trailering the boat to the scene). In order to confirm that these cases do not occur as a result of errors in the database, he draws a bounding box over these incidents on the map and obtains a summary of these incidents. The summary confirms that these incidents do indeed occur in that particular lake. As station A is closer to these cases, the analyst concludes that station A would be a better candidate to absorb these cases as opposed to station B. But for the rest of the cases, the results clearly suggest that station B would be a better candidate to absorb station X's cases, and that station A would increase the maritime risks if allowed to absorb station X's cases alone. Thus, this analysis confirms that a combination of these two stations yields the best results. With these results at hand, the analyst may also recommend using an aircraft to respond to cases that take more than 120 minutes. Or, as our preceding analysis showed, stations A and B may absorb the work load of station X together, with station B receiving a higher share of resources than station A. We also note that in the future, our system could be modified to perform a real time analysis of SAR cases and could then be used to assign each case to the correct station in real time.

The analyst now uses the risk profile tools to observe the increase in risks when station X is closed. This is shown in Figure 6, with the left figure showing the average response time risk profile when Auxiliary station X is functional whereas the right figure shows the risk profile when station X is closed. The analyst explores the new average response time on the map when station X is closed and determines the potential increase in maritime risks in the region. The

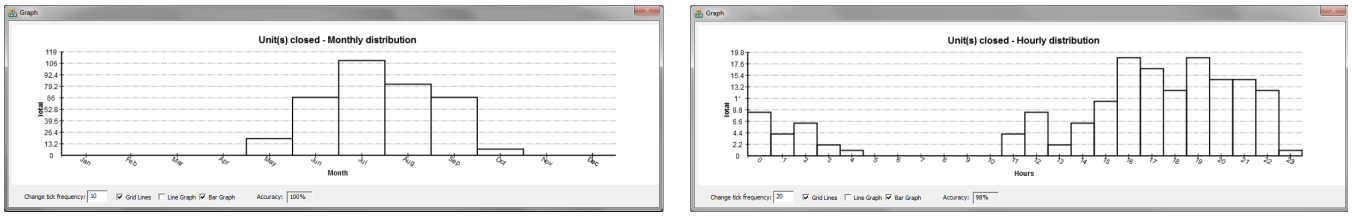


Figure 7: The monthly (left) and hourly (right) distributions of SAR incidents handled by Auxiliary station X between Jan. 2006 and Sept. 2010.

analyst thus identifies the regions in station X's area of operation that take a greater time to respond.

The analyst also visualizes the monthly and hourly distributions of all SAR cases responded to by the closed Auxiliary station X between January 2006 and September 2010 (shown in Figure 7, where the left graph shows the monthly distribution and the right graph shows the hourly distribution). We see that all activity occurs in the summer months with a peak occurring in July and the station responds to most cases mainly during evening hours. The analyst also visualizes the temporal activity of stations A and B, and determines the potential work load increase for both stations. This helps the analyst determine how the available resources must be reallocated if Auxiliary station X is to be closed. Furthermore, the analyst chooses to visualize the case distribution of Auxiliary station X between January 2006 and September 2010 using the interactive calendar view widget. This generates a summary dialog box that provides the details of the SAR cases responded to by station X and includes details including the total number of lives assisted, saved and lost. This further helps the analyst understand the risks associated with closing this station by providing an overview of the cases occurring in this region. With all these results at hand, the analyst uses his domain knowledge to make an informed decision.

6 Domain Expert Feedback

Our system was assessed by an analyst at the U.S. Coast Guard's Ninth District who is currently using the system to determine the potential risks in the maritime domain associated with the hypothetical allocation of Coast Guard resources. The analyst emphasized the need of such systems in the maritime domain that allow users to quickly and easily process large datasets in order to derive actionable results. The analyst noted that processing the desired queries took him a fraction of the time when using our system as compared to using other software (e.g., Microsoft Excel) that he had been previously using in his analysis. He was impressed by the fact that the system is intuitive to use and requires little user training. He observed that the system's ability to process large datasets allows him to quickly filter the data into manageable subsets while providing interactive spatiotemporal displays that further aid him (and ultimately the senior level decision makers) in making a decision using the best information available.

7 Conclusions and Future Work

Our current work demonstrates the benefits of visual analytics in analyzing risk and historic resource allocation in the maritime domain. Our visual analytics system provides analysts with a suite of tools for analyzing risks and consequences of taking major decisions that translate into important measures including potential lives and property lost. Our results show how our system can be used as an effective risk assessment tool when examining various mitigation strategies to a known or emergent problem.

Before this system was developed, Coast Guard officials explored possible mitigation strategies, including the implementation of seasonal or weekend only Auxiliary duty stations, but the sheer volume of data and information inhibited the efficient processing of the data. However, using our system, the decision makers were quickly made aware that most response cases happened on Mondays/Tuesdays at some of the units. This further asserts the benefits of the use of visual analytics in the maritime domain.

In addition to performing risk analysis on the Coast Guard SAR cases, our system can also be used to conduct a thorough review of the operations (i.e. non-distress cases) conducted by different Coast Guard stations. Users may choose to visualize different datasets and analyze how each station performs in terms of factors including average response times, average distance to target, lives saved, lives assisted, lives affected, etc. Hence, the officials may analyze the efficiency of each Coast Guard station and identify problem areas that may require further attention.

Future work includes deploying our system to assist in the analysis and optimization of all operations conducted by the U.S. Coast Guard Ninth District and expanding the use of our system to other Coast Guard districts. We plan on implementing algorithms that factor the geography of the coast line in the risk assessment process in order to get accurate response times by the Coast Guard assets. We also plan on employing prediction algorithms in the temporal domain as well as spatiotemporal correlation algorithms that correlate different datasets (e.g., weather, water temperature) with the response dataset to provide insights into the operation of the Coast Guard stations. Furthermore, we plan on incorporating additional risk metrics to provide insights into different risk scenarios.

8 Appendix

In this section, we briefly provide some domain specific terms and definitions:

Coast Guard Auxiliary: Volunteers that support the Coast Guard.

Coast Guard Ninth District: The area of Coast Guard operations that encompasses the Great Lakes.

Atlantic Area Command: The area of Coast Guard operations East of the Rocky Mountains.

Captain of the Port (COTP) Zone: Further division of Coast Guard operations within a Coast Guard District.

Unit or Station: The operational execution arm of the Coast Guard. For example, the small boat station provides the boat and personnel to execute the assigned mission.

Coast Guard asset: A boat or an aircraft reserved to perform Coast Guard operations.

Coast Guard sortie: An asset that responds to an incident.

9 Acknowledgments

The authors would like to thank Capt. Eric Vogelbacher, Steffen Koch and Zichang Liu for their feedback. This work is supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0002.

10 Disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the U.S. Coast Guard.

REFERENCES

[1] Command Post Of The Future (CPOF). Internet: http://www.gdc4s.com/documents/cpof_datasheet_web.pdf, [June 20, 2011].

[2] ESRI, ArcView. Internet: http://www.esri.com/software/arcgis/arcview/index.html, [June 9, 2011].

[3] Oculus, GeoTime. Internet: http://www.oculusinfo.com/SoftwareProducts/GeoTime.html, [June 25, 2011].

[4] OpenStreetMap. Internet: http://www.openstreetmap.org, [June 15, 2011].

[5] R. Adler and J. Fuller. An integrated framework for assessing and mitigating risks to maritime critical infrastructure. In *IEEE Conference on Technologies for Homeland Security*, pages 252 –257, May 2007.

[6] C. A. Brewer. *Designing Better Maps: A Guide for GIS users*. ESRI Press, 2005.

[7] J. V. Carlis and J. A. Konstan. Interactive visualization of serial periodic data. In *Proceedings of the Symposium on User Interface Software and Technology*, pages 29–38, 1998.

[8] W. S. Cleveland and M. E. McGill, editors. *Dynamic Graphics for Statistics*. Wadsworth and Brooks/Cole, 1988.

[9] C. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 51 –58, Oct. 2009.

[10] S. G. Eick and B. S. Johnson. Interactive data visualization at AT&T Bell laboratories. In *Conference companion on human factors in computing systems*, pages 17–18, New York, NY, USA, 1995.

[11] S. G. Eick and G. J. Wills. Navigating large networks with hierarchies. In *Proceedings of the 4th conference on visualization*, Visualization 1993, pages 204–209, Washington, DC, USA, 1993. IEEE Computer Society.

[12] M. Feather, S. Cornford, J. Kiper, and T. Menzies. Experiences using visualization techniques to present requirements, risks to them, and options for risk mitigation. In *First International Workshop on Requirements Engineering Visualization*, Sept. 2006.

[13] R. A. Gandhi and S.-W. Lee. Visual analytics for requirements-driven risk assessment. In *Second International Workshop on Requirements Engineering Visualization*, October 2007.

[14] F. Hardisty and A. C. Robinson. The geoviz toolkit: using component-oriented coordination methods for geographic visualization and analysis. *International Journal Geographical Information Science*, 25:191–210, February 2011.

[15] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[16] A. Inselberg. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer, September 2009.

[17] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the information age: Solving problems with Visual Analytics*. EuroGraphics, 2010.

[18] R. Lane, D. Nevell, S. Hayward, and T. Beaney. Maritime anomaly detection and threat assessment. In *13th Conference on Information Fusion*, pages 1 –8, July 2010.

[19] R. Laxhammar. Anomaly detection for sea surveillance. In *11th International Conference on Information Fusion*, pages 1 –8, July 2008.

[20] M. Mansouri, R. Nilchiani, and A. Mostashari. A risk management-based decision analysis framework for resilience in maritime infrastructure and transportation systems. In *3rd Annual IEEE Systems Conference*, pages 35 –41, March 2009.

[21] M. Migut and M. Worring. Visual exploration of classification models for risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 11–18, October 2010.

[22] B. Ristic, B. La Scala, M. Morelande, and N. Gordon. Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. In *11th International Conference on Information Fusion*, pages 1 –7, July 2008.

[23] J. Roy and M. Davenport. Categorization of maritime anomalies for notification and alerting purpose. *NATO Workshop on Data Fusion and Anomaly Detection for Maritime Situational Awareness*, pages 15–17, September 2009.

[24] S. Rudolph, A. Savikhin, and D. Ebert. Finvis: Applied visual analytics for personal financial planning. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 195 –202, Oct. 2009.

[25] N. M. Sanusi and N. Mustafa. A visualization tool for risk assessment in software development. In *International Symposium on Information Technology*, pages 1–4, August 2008.

[26] A. Savikhin, R. Maciejewski, and D. Ebert. Applied visual analytics for economic decision-making. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114, October 2008.

[27] R. Scheepens, N. Willems, H. van de Watering, and J. J. van Wijk. Interactive visualization of multivariate trajectory data with density maps. In *Proceedings of the Pacific Visualization 2011*, pages 147–154, March 2011.

[28] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[29] C. G. Soares and A. P. Teixeira. Risk assessment in maritime transportation. *Reliability Engineering and System Safety*, 74(3):299 – 309, 2001.

[30] J. Stasko, C. Gorg, Z. Liu, and K. Singal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 131–138, 2007.

[31] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[32] U. S. C. G. U.S. Department of Homeland Security. *U.S. Coast Guard Addendum to the United States National Search and Rescue Supplement (NSS) to the International Aeronautical and Maritime Search and Rescue Manual (IAMSAR)*. September 2009.

[33] J. J. V. Wijk and E. R. V. Selow. Cluster and calendar based visualization of time series data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 4–9, San Francisco, CA, USA, 1999. IEEE Computer Society.

[34] N. Willems, H. van de Wetering, and J. J. van Wijk. Visualization of vessel movements. *Computer Graphics Forum*, 28(3):959–966, 2009.

Bristle Maps: A Multivariate Abstraction Technique for Geovisualization

SungYe Kim, Ross Maciejewski, *Member, IEEE,* Abish Malik, Yun Jang, *Member, IEEE,* David S. Ebert, *Fellow, IEEE,* and Tobias Isenberg, *Member, IEEE*

Abstract—We present Bristle Maps, a novel method for the aggregation, abstraction, and stylization of spatio-temporal data that enables multi-attribute visualization, exploration, and analysis. This visualization technique supports the display of multi-dimensional data by providing users with a multi-parameter encoding scheme within a single visual encoding paradigm. Given a set of geographically located spatio-temporal events, we approximate the data as a continuous function using kernel density estimation. The density estimation encodes the probability that an event will occur within the space over a given temporal aggregation. These probability values, for one or more set of events, are then encoded into a bristle map. A bristle map consists of a series of straight lines that extend from, and are connected to, linear map elements such as roads, train, subway lines, etc. These lines vary in length, density, color, orientation, and transparency—creating the multivariate attribute encoding scheme where event magnitude, change, and uncertainty can be mapped as various bristle parameters. This approach increases the amount of information displayed in a single plot and allows for unique designs for various information schemes. We show the application of our bristle map encoding scheme using categorical spatio-temporal police reports. Our examples demonstrate the use of our technique for visualizing data magnitude, variable comparisons, and a variety of multivariate attribute combinations. To evaluate the effectiveness of our bristle map, we have conducted quantitative and qualitative evaluations in which we compare our bristle map to conventional geovisualization techniques. Our results show that bristle maps are competitive in completion time and accuracy of tasks with various levels of complexity.

Index Terms—Data transformation and representation, data abstraction, illustrative visualization, geovisualization.

1 Introduction

As data dimensionality increases, the encoding of variables and their relationships is often abstracted down to a representative subset for analysis in a single display, or dispersed across a series of coordinated multiple views [1–3]. Moreover, many techniques have been developed to visually encode multiple data attributes/variables for each data sample to enable interactive analysis, ranging from discrete glyph attribute encoding [4] to more spatially continuous color, transparency, and shading encodings [5–7]. As the number of visualized variables increases, the amount of information that can be effectively displayed becomes limited due to over-plotting and cluttering [8]. This is especially a problem in geographical visualization as a key attribute of the data is the location within the two-dimensional map space.

In geographical visualization, data can be described at any given location on a map. The data being described can come from an aggregated measurement, a direct event occurrence, or various other means. In dense data sets, plotting events as symbols on the map (e. g., Fig. 1(a)) leads to cluttering and is often unable to convey a meaningful sense of event magnitude within the data. Aggregation of the data by defined boundaries, such as county or census tract boundaries (e. g., Fig. 1(b)), leads to a loss of specificity in data location and runs afoul of the Modifiable Areal Unit Problem [9]. Furthermore, it is known that the level of data aggregation can affect aspects of task complexity such as information load and the user's ability to recognize patterns within the data [10]. In order to combat problems associated with areal aggregation, dasymetric mapping focuses on using zonal boundaries that are based on sharp changes in the statistical surface being mapped [11]. However, even when grouping data into small spatial quadrats, data can either be over-aggregated or under-aggregated. A third option is to estimate the discrete event points as a continuous function (e. g., Fig. 1(c)); such a mapping, however, only allows for the use of color as a means of representing data variables. As an encoding based on underlying network data, Fig. 1(d) shows a traditional line map. However, its representation is still restrained by the color and thickness of the lines.

In order to increase the amount of information that can be visualized within the constraints of a thematic map, this paper explores a novel method of multivariate encoding. Inspired by ideas of symbolic encoding from Spence [12] and choices of visual encodings by Wilkinson [13], we have developed the bristle map (Fig. 1(e)), a novel method for the aggregation, abstraction, and stylization of geographically located spatio-temporal data. The bristle map consists of a series of straight lines extended from and connected to

- *S. Kim, A. Malik and D. S. Ebert are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. E-mail: {inside|amalik|ebertd}@purdue.edu.*
- *R. Maciejewski is with Arizona State University, Tempe, AZ, USA. E-mail: rmacieje@asu.edu.*
- *Y. Jang is with Sejong University, Seoul, South Korea. E-mail: jangy@sejong.edu.*
- *T. Isenberg is with INRIA, Saclay, France. E-mail: tobias.isenberg@inria.fr.*

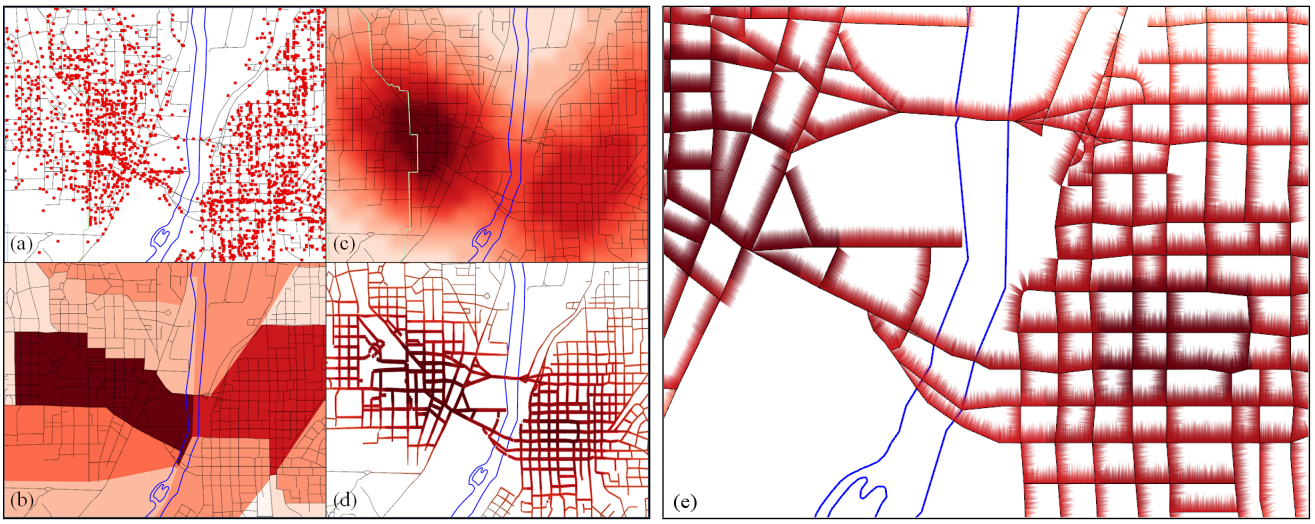


Fig. 1: Data abstraction in geovisualization. In this image, we show crimes in West Lafayette and Lafayette, Indiana where the blue line represents the Wabash River. (a) Plotting events as points. (b) Aggregation of points by areal units. (c) Approximation of a continuous domain from point sampling. (d) Approximation of a continuous domain using solid lines applied to roads. (e) Our abstraction using a series of bristle lines applied to roads.

linear map elements (roads, train lines, subway lines, etc.) that have some contextual relationship with the data being visualized. We vary these lines with respect to their color, length, density, and orientation to allow for a unique encoding scheme that can be used to create informative maps. With respect to the other representations shown in Fig. 1, our technique utilizes the underlying geographical context as a part of its symbology, thereby directly incorporating geographical elements within its encoding scheme. One of the major advantages of the bristle map technique is that the basis domain of the data (e. g., street network) remains highly visible regardless of the color scale being used. If one compares Fig. 1(c) and (e), the street network in Fig. 1(e) is clearly visible because the lines only 'bristle off' to one side, whereas in Fig. 1(c) some streets are hardly discernible due to the dark colors.

To demonstrate our technique, we focus on categorical spatio-temporal event data (e. g., emergency department logs, crime reports). In such data, events consist of locations in time and space where each event fits into a hierarchical categorization structure. These categories are typically processed as time series and snapshots of time are aggregated and typically visualized on a choropleth map [14]. Past work [6, 15] has shown that the use of kernel density estimation [16] is highly suitable in the spatial analysis of such data. Thus, our approach incorporates kernel density estimation as a means of estimating the underlying distribution of spatio-temporal events. Using the estimated distribution in an area for a given category (or categories) and temporal unit, we incorporate the underlying geographical network structure into the visual encoding. Bristles are extended from this underlying structure, and the color, length, density, transparency, and orientation of each bristle is mapped to a particular variable (or set of variables). Schemes presented in this paper include combinations of the following mappings:

- length, density, and color as data magnitude,
- orientation and coloring for bivariate mapping,
- color and length for bivariate mapping,
- color and density for bivariate mapping, and
- length and transparency for temporal variance.

Given the available parameters for visual encoding within the bristle map, other encodings also exist, which illustrate the flexibility and power of our technique. Our work focuses on showing how bristle maps can be used to show spatial and temporal correlations between variables, encode uncertainty in a unique way, and maintain geographical context through linking our visual encoding directly to geographical components. As such, the bristle map is a powerful multivariate encoding scheme that is adaptable to various attribute encodings to create richly informative visualizations.

2 Related Work

Many techniques in multivariate data visualization focus on a means of reducing clutter and highlighting information through a variety of approaches including filtering (e. g., [17]), clustering (e. g., [18]), and sampling (e. g., [19]). In this section, we focus particularly on techniques within geographical visualization for improving the understanding of thematic/statistical maps, as Wilkinson [13] noted that the problem of multivariate thematic symbology for maps is that they are not only challenging to make, but also challenging to read.

In geographical visualization, the most common means of data representation is the choropleth map in which areas are shaded or patterned in proportion to a measured variable. Such maps are typically used to display only one variable, which is mapped to a given color scale. Other research has focused on encoding multivariate information into choropleth maps (such as uncertainty) with textures

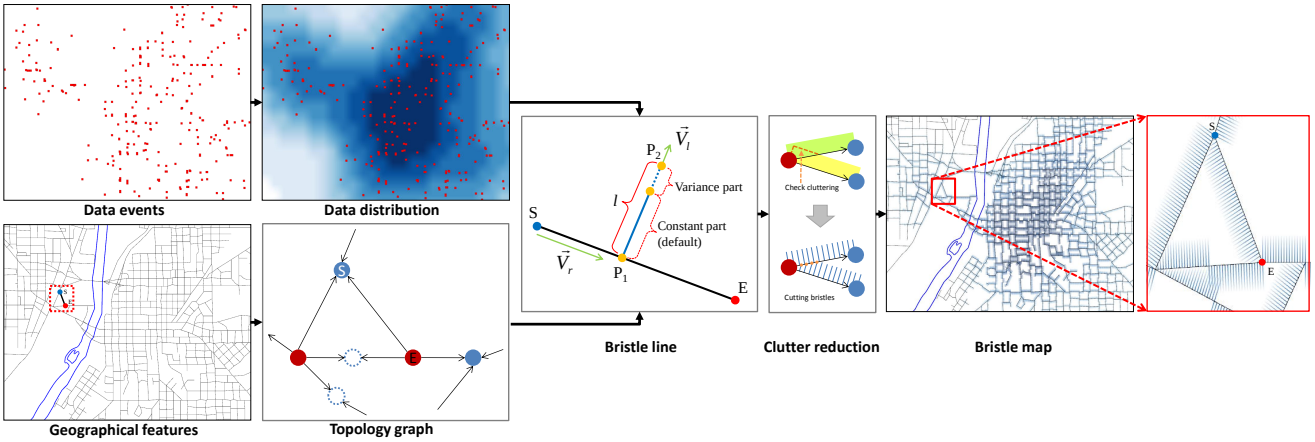


Fig. 2: The bristle map generation pipeline. Beginning with data events, a continuous abstraction is created. We also create a topology graph from contextually important linear features (in this case roads). Next, bristles are extended from these features based on the continuous abstraction and the topology. Clutter reduction is performed when generating each bristle, and finally the resultant bristle map is generated.

and patterns [20], creating bivariate color schemes for visualizing interactions between two variables [21, 22], or animating choropleth maps to enhance the exploration of temporal patterns and changes [23]. We present bristle maps as a robust alternative to these schemes in which multivariate attributes are instead mapped to a variety of graphical properties of a line (length, density, color and orientation), as opposed to utilizing a bivariate color scheme, texture overlays, or animation.

More recent geographical visualization techniques have included extensions to choropleth mapping ideas. Hagh-Shenas et al. [24] compared the effectiveness of visualizing geographically referenced data through the use color blending (in which a single composite color conveys the values of multiple color encoded quantities) and color weaving methods (in which colors of multiple variables are separately woven to form a fine grained texture pattern). The results from their study indicate color weaving to be more effective than color blending for conveying individual distributions in a multivariate setting. Saito et al. [25] proposed a two-tone pseudo coloring method for visualizing precise details in an overview display. Under this scheme, each scalar value is represented by two discrete colors. Sips et al. [26] focused on revealing clusters and other relationships between geo-spatial data points by their statistical values through the over-plotting of points. This work was later extended [27] to combine a cartogram-based layout to provide users with insight to the relative geo-spatial positioning of the dataset while preserving cluster information and avoiding over-plotting. Other cartogram techniques include the WorldMapper Project [28] which is used to represent social and economic data of the countries of the world. In each of these, novel data visualization techniques are created; however, the distortion of spatial features (country boundaries, roads) is often undesirable. While these techniques focus on displaying large amounts of aggregate data on small screens, our technique focuses on enhancing details of geographical context within the data. A similar concept of preserving data context is found in Wong et al.'s [29] GreenGrid in which they visualize both the physics of the power grids in conjunction with the geographical relationships using graph based techniques.

Along with the previously described map schemes and cartogram distortions, there has been work in the use of heatmaps based on spatial data. Fisher [30] applied heatmaps to visualize the trends of the interactions of users with interactive maps that are based on their view of the geographic areas. Maciejewski et al. [6] used heatmaps as one of the tools to find aberrations or hotspots that facilitate the exploration of geo-spatial temporal datasets. Work by Chainey et al. [15] illustrated a number of different mapping techniques for identifying hotspots of crime and demonstrated that kernel density estimation provides analysts with an excellent means of predicting future criminal activities.

In conjunction with previous visualizations, other research has focused on expanding the dimensionality of the data being displayed by utilizing three-dimensional visuals. Van Wijk and Telea [7] utilized color and heightfields to visualize scalar functions of two variables. Tominski et al. [31] explored embedding 3D icons into a map display as a means of representing spatio-temporal data. In contrast, our work focuses on a two-dimensional encoding scheme that incorporates a variety of the visual variables described by Bertin [32] and Wilkinson [13] as a means of representing multivariate data.

Finally, it is important to note that our technique is akin to traditional traffic flow maps (e. g., Fig. 1(d)) seen in a variety of atlases; however, provides more generalized schemes. In traffic flow maps, the amount of data that can be displayed is restrained by the color and the width of the line representing linear elements (i. e., roads) on the map. Our work is similar to that of the traffic flow maps in that we utilize width (specifically, matched to the length in our bristle maps) and color as underlying visual variables of our encoding. However, our work also incorporates bristle density as a means of further encoding parameters. In the following sections, we compare our encodings to a variety of methods including the point, color, and flow line maps.

3 BRISTLE MAP GENERATION

In Fig. 1, we developed our motivation for the need to directly incorporate geographic features to the underlying data in order to better preserve contextual information. It is clear that the aggregation of data into arbitrary geographical areas obscures data, while the continuous approximation of an underlying data source can lead to incongruent mappings with respect to geographic features. Furthermore, both these mappings are limited in the fact that only color and texture are available for variable encoding, limiting the amount of data that can be displayed to either a single variable or possibly two variables in the case of a bivariate color map. The goal of this work is to create visual encodings for higher order structures.

The bristle map was inspired by the Substrate simulation of Tarbell [33] and abstract renderings of map scenes in work by Isenberg [34]. Given these images, our work focuses on using the underlying visual properties to intelligently encode information for display. In *The Grammar of Graphics* [13], Wilkinson discusses the combination of several perceptual scales into a single display. Here, he notes the idea of separable dimensions of the data is a key issue, where discriminations between stimuli are of key importance in the visualization. The Substrate aesthetic directly lends itself to this approach as color, line length, and orientation are distinct classes within Wilkinson's table of aesthetic attributes and each of these visual parameters directly contributes to the substrate aesthetic.

Fig. 2 illustrates the bristle map generation pipeline. Given underlying data events, we compute a continuous distribution. We also create a topology graph from given geographically relevant linear content for clutter reduction described in Section 5. As an example of geographical content, if the underlying data was water pollution we could use a city sewage map for the geographic components, for our crime data examples we use roadways. Each linear geographic component consists of a series of line segments, and we extend *bristle lines* from these line segments. These bristle lines emerge perpendicularly from the underlying geographical line segment and are allowed to vary in length, density, color, transparency, and orientation, to facilitate multivariate data encoding. The third stage of the bristle map generation pipeline (Fig. 2) illustrates the bristle line concept for each geographical line segment, $\overline{SE}$, and $\overline{P_1P_2}$ defines our generated bristle line. Each bristle line is created using the vector equation of a line as shown in Equation 1.

$$P_2 = P_1 + \overrightarrow{V_l} L_l = P_1 + \overrightarrow{V_l}(t \times L_{lmax}) \tag{1}$$

Here, P_1 is a point on the contextually relevant geographic line segment, $\overline{SE}$, $\overrightarrow{V_l}$ is a unit vector perpendicular to the line $\overline{SE}$, and L_{lmax} is the maximum length of the $\overline{P_1P_2}$. L_l is the length of the $\overline{P_1P_2}$ determined by a parameter t.

Each line from P_1 to P_2 is drawn in such a manner that it will either encode different properties of a multivariate data set, or use a data reinforcement technique where properties are encoded to the same variable to provide redundant cues. We utilize three encoding properties for each bristle; length, color, and orientation. The length of a line $\overline{P_1P_2}$ is separated into two portions; a constant component, which is proportional to the magnitude of the variable being encoded, and a variance component. It captures temporal variance or other properties such as level of certainty. The color of a bristle $\overline{P_1P_2}$ is proportional to the underlying variable distribution to be encoded at point P_1. When the variance component is used, its transparency is adjusted as a means of visually distinguishing it from the constant component. Orientation of the bristle line is always perpendicular to $\overline{SE}$ and is utilized for bivariate comparison (i. e., day/night, two data types) and/or clutter reduction. To summarize, length and color represent a local data magnitude property at point P_1. We also choose to encode redundant information into the density of the number of bristles placed on a given line segment where the density of the bristles along $\overline{SE}$ is decided by an average data value on a line segment $\overline{SE}$.

TABLE 1: Parameters, corresponding variables, and ranges.

Parameters	*Potential variables*	*Range*
Base position (P_1)	Geographic location	(Double, Double)
Length 1 (constant portion)	Data magnitude	Double
Length 2 (variance portion)	Temporal variance, accuracy	e. g., monthly/yearly
Color	Data magnitude	Discrete, Continous
Transparency	Temporal variance, accuracy	Double [0.0, 1.0]
Orientation ($\overrightarrow{V_l}$)	Temporal difference, data type	Clock-wise, Counter clock-wise
Density	Average data magnitude on an area ($\overline{SE}$)	Double

For each visual encoding, the underlying data is assumed to be continuous over a given geographical segment, such that for all points between any two nodes on the underlying contextual geographic structure, a data distribution value is associated with the point. In the case of a discrete data set (e. g., crime locations), the choice of an appropriate means of data interpolation with regards to the underlying geographic information is dependent on the data analysis being performed. Based on the recommendations of Chainey et al. [15], we apply a kernel density estimation (KDE) [16] to approximate the underlying distribution of crimes over the geographic features. The kernel density estimation procedure used is defined by the following equation:

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{\mathbf{x}-X_i}{h}\right) \tag{2}$$

Here, the window width of the kernel placed on point x is proportional to a window bandwidth, h, and the total number of samples, N. We utilize the Epanechnikov kernel [16], Equation 3:

$$K(\mathbf{u}) = \frac{3}{4}(1-\mathbf{u}^2)1_{(||\mathbf{u}||\leq 1)} \tag{3}$$

where the function $1_{(||\mathbf{u}||\leq 1)}$ evaluates to 1 if the inequality is true and zero for all other cases.

Thus, given a multivariate dataset where locations in space and time correspond to a series of categorized events,

we can create bristle maps that encode various properties of the data. Note that this technique relies on the data being contextually relevant to an underlying geographical network. For example, crime event data with its 2D geographical coordinates is recorded and hence defined by addresses on streets; thus it is contextually relevant to a street network. Data sets in which this contextual relationship does not exist should utilize other visual encoding schemes. Table 1 shows the parameters in our bristle map and their corresponding potential variables being encoded to each parameter. In the following section, we present a series of potential parameter combinations for various bristle map encodings and discuss the various results.

4 Encoding Schemes

The bristle map is a powerful visual encoding scheme that lends itself to a variety of data encodings, examples of which we present next. For demonstration purposes we employ categorical spatio-temporal police reports collected in Tippecanoe County (specifically West Lafayette and Lafayette, IN, USA), from 1999 to 2010. The data set contains the date, time, crime type (e. g., armed/unarmed aggravated assault, armed robbery, burglary, homicide, noise, other assaults, rape, rape attempted, residential entry, robbery, theft, vandalism, and vehicle theft), and the address of each recorded criminal event. Note that other datasets can be easily encoded with bristle maps, and our choice of data was only made to illustrate the technique.

Utilizing this multivariate crime data set we discuss potential encoding schemes for multivariate spatio-temporal data. We then provide illustrations of each described encoding scheme with respect to our crime data set. Encoding schemes presented in this section include the use of bristle color, length, and density to encode data magnitude, the use of bristle orientation to inform temporal comparison, and the encoding of temporal variance in the bristle lengths.

4.1 Color, Length, and Density as Data Magnitude

Here, we discuss our technique for encoding the color, length, and density of the bristles into two separate variable groups. As both color and length (size) fall into two distinct categories of aesthetics according to Wilkinson [13], the use of separate variables for both categories allows for a distinguishable visual data encoding. In both cases, we assign data magnitude to both a color scale and a length scale. We note that such an encoding scheme has the potential to portray data more effectively than visualizations that map each data variable to a single display parameter. As noted in the arguments for the use of redundant color scales by Rheingans [35], the use of different display parameters is able to convey different types of information. Furthermore, by combining encodings in a redundant manner, it is possible to reinforce the encoding scheme. The utility of redundant color scales was confirmed by Ware [36].

In our encoding scheme, each bristle line's length, L_l, is calculated using Equation 4 based on a parameter, t, and the maximum length, L_{lmax}.

$$L_l = t \times L_{lmax} = (\alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1}) \times L_{lmax} \qquad (4)$$

For this visual encoding of the bristles, the parameter t is defined by the ratio of the data value at P_1, which we call κ_{P_1}, the ratio of the temporal variance at P_1, υ_{P_1}, and a set of tuning parameters (α and β) which provide weights to the constant and variance components as shown in Fig. 2. In this work, we use α=1.0 and β=0.3. Note that the choice of encoding the variance at a 30% value was chosen through trial and error by generating visualizations that the authors found to be the most useful and aesthetically pleasing. For problems where determining exact data values from the visual encoding is required (as opposed to approximating high and low rates), the variance portion is removed from the equation entirely by using β=0.0. As such, by creating the encoding scheme with diverse parameters, we are able to generate more aesthetic choices and visualizations. It is important to note that not all encodings will be appropriate and are most likely task dependent.

The L_{lmax} portion of Equation 4 is defined in Equation 5.

$$L_{lmax} = \rho \times log_b \left(\frac{1}{N_r} \sum_{i=0}^{N_r - 1} L_{\overline{SE}} \right) \qquad (5)$$

In this equation, we take the average length of all line segments (where N_r is the total number of line segments in the map) and calculate L_{lmax} using a non-linear function such that the length of bristle lines does not grow in an unbounded manner when zooming in. Moreover, L_{lmax} is modified by the parameter ρ, where ρ is the ratio of the current zoom level to the initial zoom level, to decouple our technique with the zoom level. In this work, we use b=15 for the base of a log function.

Next, we determine the number (or density) of bristles, N_l, to be drawn on each line segment $\overline{SE}$ using Equation 6.

$$N_l = \rho \left(\frac{\zeta}{\lambda} L_{\overline{SE}} \right) \kappa_{\overline{SE}} \qquad (6)$$

Here, N_l is calculated using two user-defined constants λ and ζ, where λ is the unit geographical length (distance) and ζ is the number of bristle lines per unit geographical length. We use λ=0.0009 and ζ=3-15 in our current visualization. As the bristle density may also be used to encode data magnitude parameters in bristle map generation, N_l should be proportional to the ratio of average data value on $\overline{SE}$, $\kappa_{\overline{SE}}$. Moreover, we also apply ρ such that N_l will be independent of the zoom level to preserve the extent of density.

For color, we allow users to choose either a continuous or a sequential color scheme from Color Brewer [37]. Then, data is linearly mapped to a probability that a crime of type A will occur at geographic point B, where the probability is estimated from the underlying data distribution using kernel density estimation as described in Section 3.

Fig. 3 illustrates our length, density, and color encoding using the previously described crime data set. Burglary is

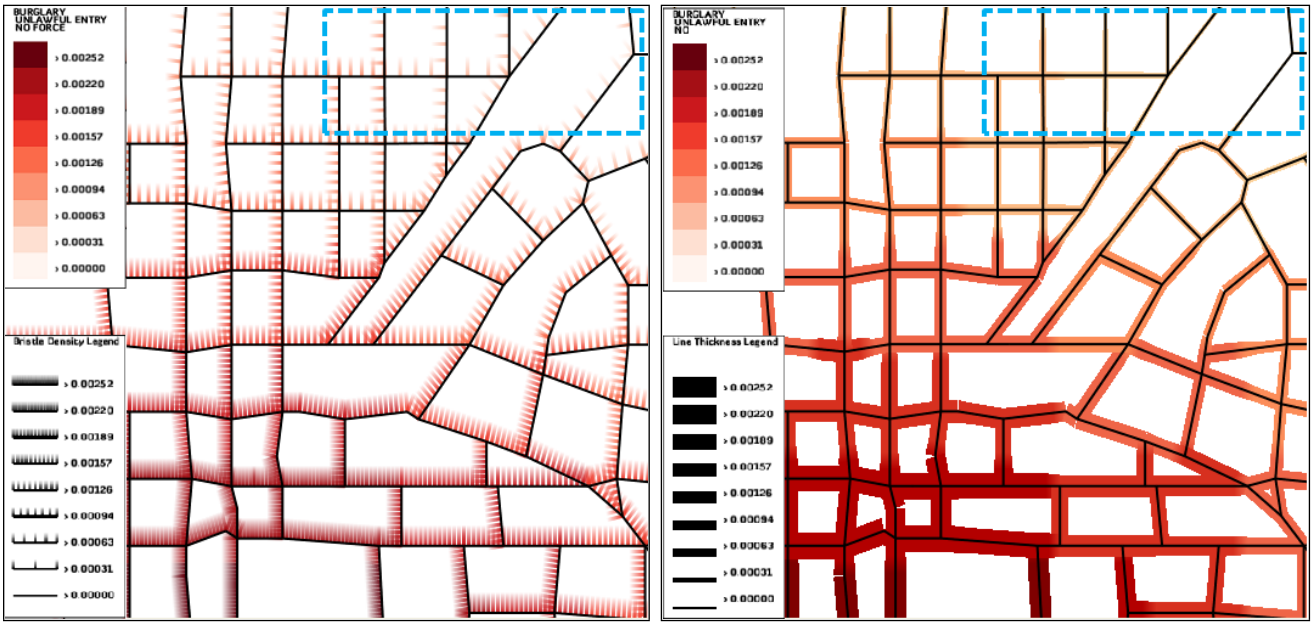


Fig. 3: (Left) Our bristle map encodes burglary rates with both bristle color and bristle density. (Right) A line map encoding burglary rates as both line color and line thickness. Compared to the line map, our bristle map provides a distinguishable visualization by incorporating bristle density. For example, bristle lines on the right top area are easily identified, whereas thickness in the line map on the same area is too small to clearly be perceived.

encoded with the red color scheme, and color is proportional to the probability (calculated from the underlying point distribution using kernel density estimation) that a burglary occurred at a given location. Fig. 3 (left) shows our bristle map encoding for burglary rates with a color scheme and bristle density, and Fig. 3 (right) shows a line map encoding the same information with a color scheme and line thickness for comparison to our bristle map. Compared to this line map, our bristle map provides the advantages of additional dimensionality through the density of bristle lines. In this scheme, one is able to easily encode two variables in different combinations of bristle map parameters (i.e., color and density with a constant length, color and length with a constant density), and provide users with distinguishable visual parameters that seem to focus attention to various details.

4.2 Multivariate Encoding: Separating Length, Density and Color, and using Orientation

In the previous section, we illustrated how our method can be utilized for univariate encoding by using a redundant encoding scheme. However, a major benefit of bristle maps is the ability to encode multivariate attributes. One example of this is seen in day versus night time comparison.

Here, one can utilize the orientation to separate two temporal components of a single variable by mapping the temporal components to different orientations of the geographic feature. For instance, it is likely that the rates of data variable will be different with respect to day and night occurrences. We illustrate this visual encoding in Fig. 4. We separate the events into day (6:00 am–6:00 pm) and night (6:01 pm–5:59 am) and map the daytime rates to red and one orientation, and nighttime rates to blue and the other orientation. In Fig. 4 (right) we illustrate a bristle map encoding of one variable (burglary) during 2009 where length, density and color represent the magnitude of the burglary as well as the encoding of day and night parameters is explored as line orientation.

In Fig. 4 (right), we show areas of high/low nighttime crime, high/low daytime crime, and combinations there within. In contrast, a traditional heatmap using a univariate color scheme can only show either daytime crime (Fig. 4 (left top)), or nighttime crime (Fig. 4 (left middle)). Hence, several heatmaps are needed to see day and night variations as shown in Fig. 4 (left column). Viewers must mentally combine the images to locate regions of the map that have high crime levels at daytime and nighttime, thereby increasing their cognitive load.

Another means of reducing the cognitive burden would be to create a heatmap of the difference between night and day. Fig. 4 (left bottom) shows the difference of day and night data, and the divergent color scheme shows where high daytime or high nighttime crimes occur. For instance, in Fig. 4 (left bottom) the right area indicates higher rates during day, the left area shows higher rates during night, and the border area between the blue and red color schemes only indicates that day and night rates were approximately equal, regardless of them being low or high. Moreover, you need other color maps to explore areas where one occurs similarly high or low during day and night time.

Bristle map encodings have benefits in this situation. When we explore a daytime versus nighttime bristle map in Fig. 4 (right), we see that there exists distinct temporal profiles along the road lines where we see exclusively dominant areas during either day or night. For instance,

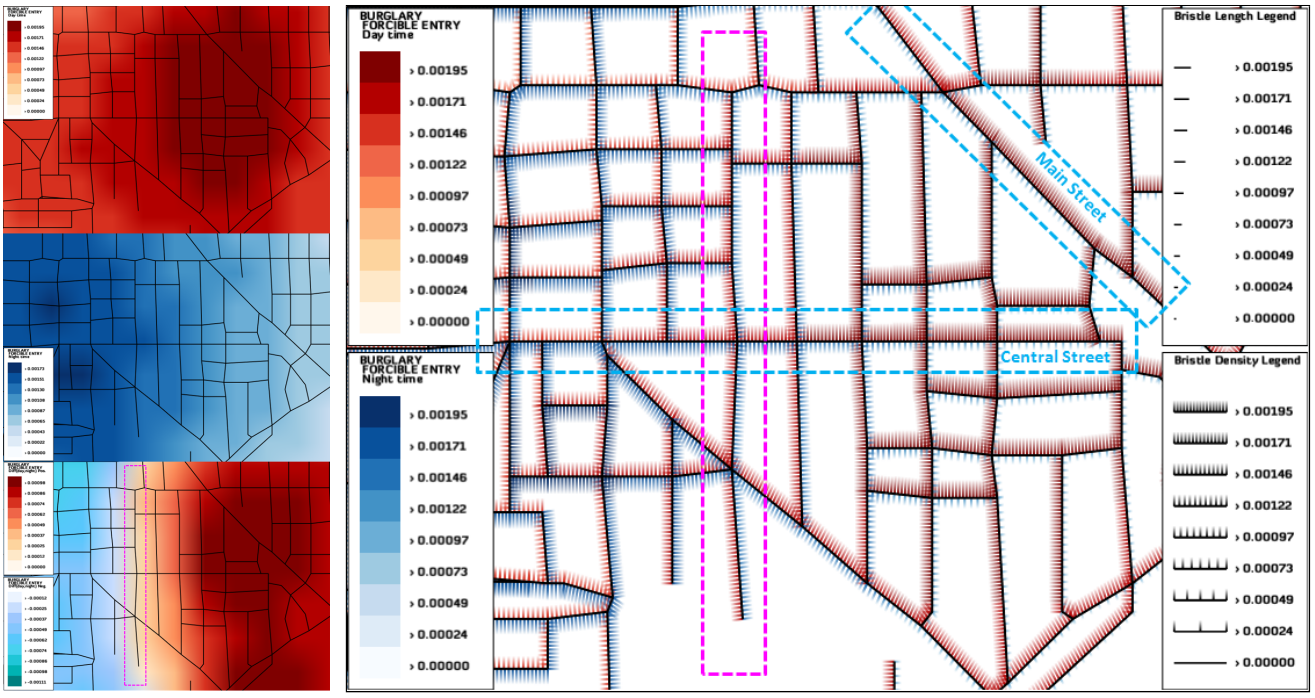


Fig. 4: Encoding daytime versus nighttime variations. (Left column) From top to bottom, color maps showing day, night and the difference of day and night burglary rates. (Right) Our bristle map separating the burglary rates into their day and night components with opposite orientation along roads. Note that a color map cannot present two components (i. e., both daytime and nighttime burglary rates) at the same location, hence three color maps are needed to see day and night variations simultaneously. Our bristle map can present such information within one bristle map by using different orientations of bristle lines.

see the diagonal road from the top center to the right center (Main Street, Lafayette, IN) showing that daytime burglary dominates along this road. Another observation is made on the horizontal road at the center of the map (Central Street, Lafayette, IN). Along this road, daytime burglary rates increase from left (west) to right (east), whereas nighttime burglary rates decrease from left to right. For the center area in Fig. 4 (left bottom), where the blue and red color schemes meet, we also see in Fig. 4 (right) that it has relatively equally high rates during both day and night. Such a comparison allows people to understand the differences between the data; however, when subtracting, areas of nearly equal daytime and nighttime crimes will be colored the same. Thus, areas that are safe during both day and night, and areas that are highly dangerous during both day and night will appear the same in the difference color map. In contrast, bristle maps allow viewers to quickly observe trends related to both day and night.

Another example of multivariate encoding using our bristle map is done by separating and/or combining bristle parameters. For instance, bristle density (or length) encodes a variable A, and color encodes a variable B while being presented on one orientation. Similarly, another two variables (C and D) could be encoded and presented on the other orientation. However, this type of parameter combination should be determined carefully so as not to increase viewers' cognitive load. Its effectiveness would depend on several factors such as data type and analysis purpose. In Section 6, we conduct experiments to explore the effectiveness of different parameter combinations.

4.3 Encoding Data Variance

As introduced in Fig. 2, each bristle can include a portion generated for temporal variance of data, see Equation 4. To present the temporal variance of the data over time, we compute both the monthly and yearly mean and variance values. For a given discrete data set during time periods N_T, we first calculate continuous distributions over time. Then, we determine mean and standard deviation values with respect to the underlying data distribution for the entire data set over a given temporal aggregation. Thus, we calculate the mean μ and variance σ values from time varying data K_i, where $i \in [0, N_T - 1]$. Note that μ and σ are computed only once as they represent constant values for a given dataset. Mean and variance values for each grid point j are calculated using Equation 7 and 8, respectively. Variance is then used to weight the parameter β in Equation 4 such that given the data magnitude at the current time K_{cur}, we compute the ratio of variance at the current time, $\tilde{\sigma}$ as shown in Equation 9. As such, the parameter t in Equation 4 can be detailed as shown in Equation 10 to represent the length of bristle lines with respect to temporal variance.

$$\mu[j] = \frac{1}{N_T} \sum_{i=0}^{N_T-1} K_i[j] \tag{7}$$

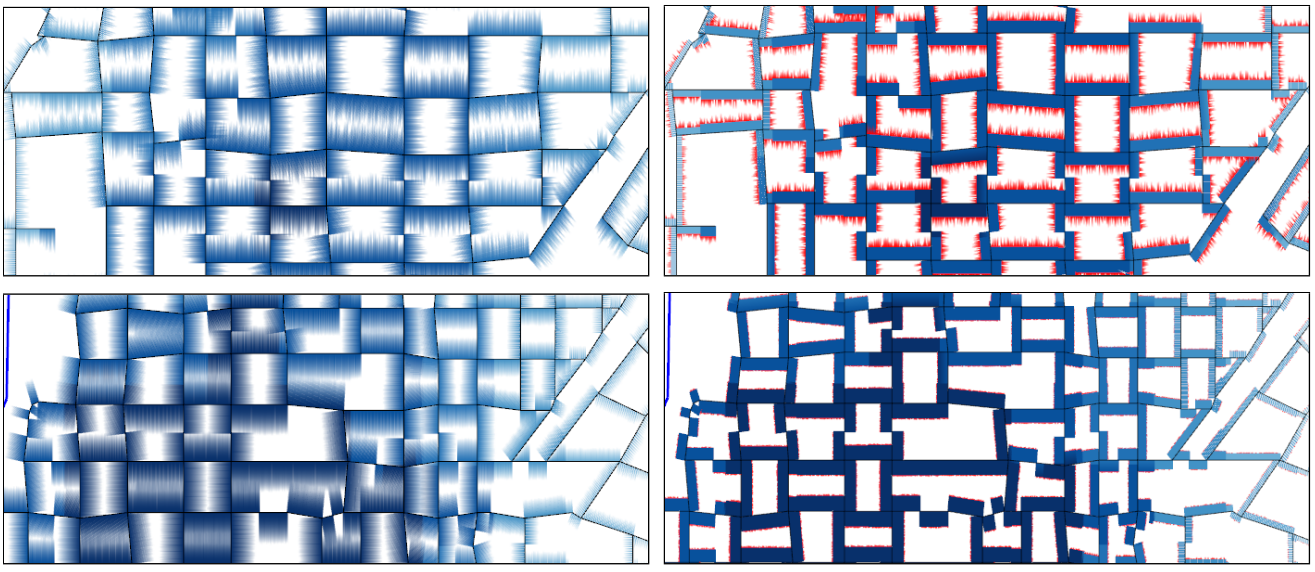


Fig. 5: Encoding data variance of vandalism-graffiti in Lafayette, IN, USA, in 2010 creating an uncertainty aesthetic. Yearly variance of vandalism-graffiti is represented in (a) a residential area and (c) a commercial area without distinguishing the variance component in the bristle length. (b) and (d) show the results using a highlight color for the variance portion and full alpha values for the constant portion of bristle lines. Here, we clearly see that our bristle map can encode the temporal variance and create an uncertainty aesthetic using the variance component.

$$\sigma[j] = \sqrt{\frac{1}{N_T}\sum_{i=0}^{N_T-1}(\mu[j]-K_i[j])^2} \tag{8}$$

$$\tilde{\sigma}[j] = \frac{1}{\sigma[j]}|\mu[j]-K_{cur}[j]| \tag{9}$$

$$t = \alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1} = \alpha \times \kappa_{P_1} + \beta \times \left(\frac{\tilde{\sigma}_{P_1}}{\tilde{\sigma}_{max}}\right) \tag{10}$$

Furthermore, the variance term, υ_{P_1}, in parameter t in Equation 4 can also be revised to encode an uncertainty factor by using randomness. We may also encode an uncertainty factor by using color and transparency to enhance the variance component. When using color and transparency, we use a highlight color for the variance component, and then fade out the variance component over the bristle length with a full alpha value for one end point and an alpha value weighted by the variance for the other end point. The constant portion of the bristle is assigned an alpha value of 1 to both end points as it represents an exact data value. Hence, according to the data type and analysis purpose, the encoding of parameter t and the use of the variance portion can be different and should be assessed with respect to the visual message trying to be conveyed. Fig. 5 illustrates the application of encoding the data variance of vandalism with the uncertainty factor. In Fig. 5(a, c), we use the same color scheme for the constant and variance portions of bristle lines. To enhance the variance component in Fig. 5(b, d) we highlighted the variance portion in a different color and assigned full alpha values for the constant portion of bristle lines. Fig. 5(a, b) show the same area. In this area, the bristle length shows large fluctuations, indicating a high yearly variance Fig. 5(c, d) show another area. In this area, the bristles are of a nearly constant length, indicating low yearly variance. When considering that the area in Fig. 5(a, b) includes residential areas, while the area in Fig. 5(c, d) includes the downtown Main street, an art theater, and the City Hall in Lafayette, IN, our bristle map shows that the residential areas have higher yearly variance of vandalism (graffiti) when compared to commercial areas.

5 Bristle Clutter Reduction

Although our bristle map can encode various characteristics from multivariate data, it often suffers from clutter around the intersections of road lines. In order to minimize cluttering, we employ two strategies in our bristle map generation pipeline (Fig. 2); 1) using topology among road lines to determine bristle orientation to minimize clutter and 2) cutting bristle lines crossing neighbor road lines.

5.1 Using Topology

Each bristle map contains an underlying topology of the contextual geographic network that the data is mapped to. In the topology graph, each node is defined as either 'outward' or 'inward' as illustrated in Fig. 6. Using the topology graph, we choose each segment's bristle line orientation such that the overlap of the bristles at intersections will be minimized, thereby reducing the clutter. If the encoding scheme requires both sides of the edge to contain bristles, then clutter at each intersection is inevitable. However, in cases where bristles map to only one side of an edge, we use the right-hand rule to decide the orientation. Hence bristle lines on edges connected to neighboring outward and inward nodes are generated in a manner that provides a reasonable reduction in clutter (Fig. 6).

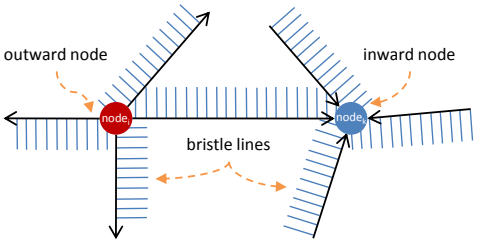


Fig. 6: To minimize clutter, a topology graph consisting of directed edges as road lines and outward (red) and inward (blue) nodes on the intersection of lines is used to decide the bristle line orientation.

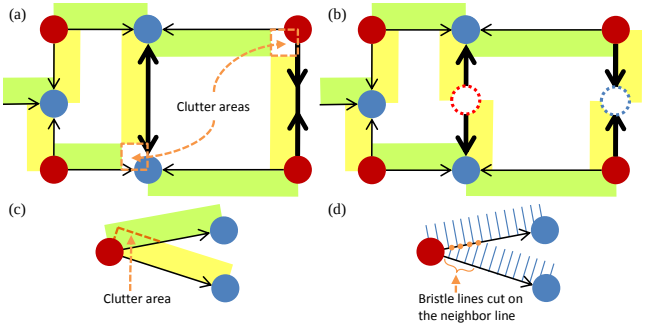


Fig. 7: Two pairs of the cluttering cases and our methods to minimize clutter. Colored box areas on a side of each edge line indicate the orientation for bristle lines. (a) Case 1: bad topology, where two inward nodes (blue) share a line and two outward nodes (red) share a line, generates inevitable clutter. (b) Virtual nodes (dotted circles) are added to split an edge line. (c) Case 2: a small angle between edge lines causes a clutter area. (d) Bristle lines crossing a neighbor edge line are cut on the neighbor line.

Choosing the orientation of bristle lines in order to minimize overlap can be considered as a 2-coloring problem in vertex coloring; one color presents 'outward' while the other presents 'inward.' Vertex coloring is a well known graph problem, where no two adjacent nodes share the same color. Moreover, coloring a general graph with the minimum number of colors is known to be an NP-complete problem. In our case, the minimum number of colors should always be 2 but such 2-colorability is not guaranteed for general road lines. While deciding the orientation of bristle lines, we often have undesirable topology generating inevitable overlap of bristle lines. Fig. 7 (upper row) shows such a bad topology example and our strategy to solve this issue. In Fig. 7(a), we see two clutter areas caused by an undesirable configuration of neighbor nodes which guarantee bristle overlap. To solve this, we consider the addition of a virtual node in a topology graph as shown in Fig. 7(b), thereby allowing for an orientation switch midway across the edge and reducing the clutter. For neighboring two inward nodes (blue), we add a virtual outward node (red dotted circle) at the road line connecting two inward nodes resulting in splitting bristle lines on the road line. Similarly, a virtual inward node (blue dotted circle) is added for neighboring two outward nodes (red).

5.2 Avoid Crossing Neighbors

Another cluttering case is illustrated in Fig. 7(c). When two road lines intersecting with less than a 90° angle have bristle lines, some of the bristle lines overlap as illustrated in Fig. 7(c). For this case, we forbid bristle lines to cross neighbor road lines by placing the end point of a bristle line on the neighbor road line as shown in Fig. 7(d). We first check the intersection of bristle blocks (colored boxes in Fig. 7) for the current road line on which we are generating bristle lines and its neighboring road lines by using the topology graph. If the blocks are intersected, we then check if a bristle line crosses the neighbor road lines by utilizing the intersection algorithm of 2D line segments [38]. This idea is based on the theory of amodal completion (or amodal perception) [39] in psychology that describes how the human visual system completes parts of an object even when it is only partially visible. Although the length of a bristle line represents data magnitude, benefits from cutting the length to avoid clutter dominate the side effects from data misunderstanding that could be caused by clutter. Moreover, when using redundant encoding utilizing bristle length and density as data magnitude, bristle density could help viewers complete parts of the bristle lines. Fig. 8 shows four image pairs before and after applying our clutter reduction strategies. Some improvements could also be considered in the future. For instance, our strategies still generate cluttered bristle lines in cases where road lines are very dense or close to others. We perform experiments in Section 6 to see how people understand the differences before and after clutter reduction. Here we note that the experiments performed were for comparison and identification tasks. In these task types, line direction (as will be shown in the experiments) had little impact on the user results. However, in a cluster/delineate task in which users are asked to segment the data, the splitting of direction may influence the user's perception of cluster boundaries. As such, we recommend that map designers take caution in employing this scheme and use it only in appropriate map contexts. Future work will explore other schemes and design issues to handle neighbor crossings and influence on map design.

6 Evaluation

To evaluate the effectiveness of our bristle maps, we conducted two quantitative controlled experiments. These studies are both comprised of an introductory session, and a training session. In the first study, five tasks were conducted to evaluate the efficiency of bristle maps compared to existing visualization methods (point, color (kernel density estimated, KDE), and line maps as shown in Fig. 1(a), (c), and (d)) and post-task questionnaires for qualitative feedback. In the second study, two tasks were conducted to evaluate the accuracy of users in estimating values from each of the map types (point, KDE, bristle and line) as well as evaluating the perceived aesthetics of each image. Prior to each study, a pilot study was also conducted to

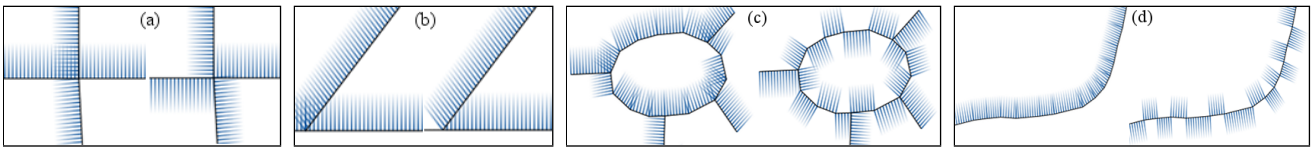


Fig. 8: Before/after image pairs of our clutter reduction. Each pair shows a case of (a) changing bristle orientation using topology, (b) cutting bristle lines crossing neighbor road lines, (c) circular roads, and (d) curved roads.

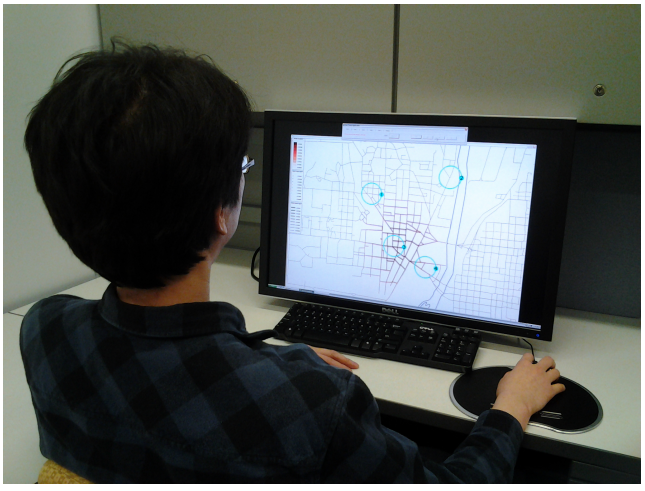


Fig. 9: Example setup for our experiment.

ensure that each task contains a fair comparison among the techniques.

***Participants*:** In the first study, thirty graduate students (23 males, 7 females) in engineering, science, and statistics from our university participated in the study. All participants reported that they had experience in visualizing data on geographical maps using colors or icons (e. g., paper maps, online map services). The experience varied from almost daily (11 participants), 1–2 times a week (17 participants) to 1–3 times a month (2 participants). For the identification/accuracy tasks and aesthetic comparisons (Tasks 6 and 7), a secondary study was run on twenty-six undergraduate students in engineering from our university.

***Apparatus*:** The experiment was performed on a 30" monitor using our experimental application running on Windows XP, as shown in Fig. 9, where all visualizations were generated with 2228×1478 resolution. Each visualization was overlayed with numbered circles as shown in Fig. 9. Participants selected one of the numbers to answer the question in each trial using buttons in the interface panel on the top of the screen. Criminal incident reports collected in West Lafayette and Lafayette, Indiana, USA from 1999 to 2010 were used in each trial, but different types of crimes were selected to generate visualizations in the training phase and in the actual study.

***Design*:** We employed a repeated measure design of tasks incorporating variations of the images shown in Fig. 1(a), (c), and (d) and line maps similar to those of Fig. 3 (right). Table 2 shows the number of data sets, techniques (cases as shown in Fig. 8 for Task 5), and trials in each task. For example, in Task 1 we utilized 18 different data sets to compare 5 different techniques (i. e., point map, color map, line map, and bristle maps using two different encoding schemes). Hence, each participant performed $18 \times 5 = 90$ trials in Task 1. In Task 3, we compared 6 different techniques (i. e., point map, bivariate color map, line maps in two different encoding schemes, and bristle maps in two different encoding schemes) with 15 data sets, resulting in 90 trials. Due to the difficulty of creating good examples to be used from our real crime data, we used fewer data sets in Tasks 4 and 5. In summary, each participant performed a total of 374 trials in Tasks 1 to 5, and it generally took 90 minutes.

Since the design of Tasks 1–5 focused on questions of comparing regions, a secondary study was also conducted. This study was again a repeated measure design of tasks incorporating variations of the images shown in Fig. 1(a), (c), and (d) and line maps similar to those of Fig. 3 (right). However, here the subjects were asked to identify the values of regions in the image. Areas of homogeneous visual variables were circled in each image and the subjects were asked to approximate the amount of crime per region. As a final task, the subjects were simultaneously presented with a point map, color map, bristle map and line map and asked to rank order the images based on their aesthetic values.

For all Tasks, trial order was varied using a magic square method [40] in each task. Completion time and participants' answers were recorded for a quantitative metric. The collected data from each task was subjected to an analysis of variance (ANOVA) test to determine if the average time and accuracy of task completion were significantly different among techniques. A post-hoc Tukey HSD test was then performed to determine significance between the techniques. P-values reported in this study come from the resultant Tukey HSD test. Before the study, participants were introduced to our experiment application and the techniques through an introductory session and a training session. During the training session, participants could ask questions and receive guidance in the use of the experiment application and analysis of each visualization. Once the training was completed, participants moved to the actual study. After completing each task (Tasks 1 to 4) participants were asked to answer the questionnaire to rate the efficiency of the techniques using a five-point Likert scale [41]. After completing Task 5, participants were also asked to describe their impression with regards to visual complexity for before and after image pairs applying our clutter reduction. In the questionnaire, we stated that the visual complexity is high if a participant felt any kind of difficulty or confusion in understanding the density, length

TABLE 2: The number of data sets, techniques (cases in Fig. 8 for Task 5), and trials.

	Data sets	Techniques (or cases)	Trials
Task 1	18	5	90
Task 2	18	5	90
Task 3	15	6	90
Task 4	12	4	48
Task 5	7	8	56
Task 6	2	4	24
Task 7	2	4	2

and color of bristle lines that encode the underlying data. Finally, after finishing all tasks, participants were asked to rate the overall efficiency among techniques.

***Hypotheses*:** In this experiment, we hypothesized that our bristle maps would be better than or equally as good when compared to the other techniques in terms of task completion time and accuracy. Specifically, we hypothesized that our bristle maps would be better than other techniques as the complexity level of tasks increased from univariate to multivariate. The rationale of this assumption is that the line map and bivariate color map use at most two variables, whereas the several encoding parameters in our bristle map have the potential to create effective encoding combinations. We also hypothesized that our clutter reduction strategies would be useful to minimize cluttering on areas where a large number of bristle lines are created. In our follow-on experiment exploring identification of values, we hypothesized that bristle maps would be as accurate as all other representations in determining values. We also hypothesized that bristle maps would be ranked higher in terms of their aesthetics.

***Tasks*:** We tested seven tasks: three for univariate, bivariate, and multivariate data encoding, respectively, one for temporal variance encoding, one for the clutter reduction, one for accuracy comparisons among the rendering styles and one for aesthetic comparisons.

In Task 1, when given four regions highlighted in circles on the map, participants were asked to "find the region with the highest crime rate" in different visualizations representing spatio-temporal crime data using point, color, line-T (data encoded in the line (T)hickness), bristle-CLD (a redundant data encoding using (C)olor, (L)ength, and (D)ensity), and bristle-LD (a redundant data encoding using (L)ength and (D)ensity).

In Task 2, four regions were highlighted in circles on the map. Participants were asked to "find the region with the highest crime rates at both (or either) day and night time," using point (encoding day/night time crime rates in different colors), color, line-TO (data encoded as line (T)hickness and using (O)rientation for day/night crime rates), bristle-CLDO (redundant data encoding using (C)olor, (L)ength and (D)ensity, and using (O)rientation to indicate day/night crime rates), and bristle-LDO (data encoded using (L)length and (D)ensity, but in a constant color, using (O)rientation to indicate day/night crime rates). The point map had differently colored points for day and night time crime rates, and two maps (day and night time color maps) were given in different colors for the color map.

In Task 3, four regions were highlighted in circles on the map. Participants were asked to "find the region with the highest crime rates for both (or either) two crimes (crime 1 and 2)," using point map (encoding two crimes in different colors), bivariate color map (Color-B), line-TO (a data encoding using (T)hickness in different colors, and using (O)rientation to indicate crime types), line-CT (encoding crime 1 using (C)olor and crime 2 using (T)hickness), bristle-LDO (a redundant data encoding using (L)ength and (D)ensity, and using (O)rientation to indicate crime types), and bristle-CD (an encoding using (C)olor to indicate crime 1 and (D)ensity to indicate crime 2, with constant length).

In Task 4, participants were given two regions highlighted in circles on the map. Then, they were asked to "find the region with the highest temporal variance" in different visualizations using point maps, color maps, line maps, and bristle-LDV (a redundant data encoding using (L)ength and (D)ensity, and representing (V)ariance in the variance part of a bristle line). For the point, color, and line maps, multiple images were displayed on the screen to provide visualizations during several years. Our bristle map embedded the variance in the variance part of the bristle length as shown in Fig. 2 (third stage) and 5 (right column).

In Task 5, given two regions predefined in circles on bristle maps, participants were asked to "answer if crime rates on this given two regions look either different or the same as each other." Fig. 8 shows representative image pairs before and after applying our clutter reduction method. In trials, participants compared each case in Fig. 8 to a base case (i. e., bristle lines on a single straight road).

In Task 6, subjects were presented with a series of images with a single predefined circle which covered an area consisting of homogeneous visual variables (i. e., identical color, bristle length, thickness, etc.). A univariate encoding was explored, and the Bristle-CLD settings were utilized for the bristle map. Participants were asked to estimate the amount of crime in the area using the provided scale (or scales in the case of bristle and line maps). Time and accuracy of the results were measured.

In Task 7, subjects were presented simultaneously with four images representing the same data set. These images consisted of a point map, a color map, a bristle map and a line map. Subjects were asked to rank order the images in order of most to least aesthetically pleasing.

7 Results and Discussion

After all tasks were completed, times and answers collected during the study were analyzed using a single-factor ANOVA. A post-hoc Tukey HSD test was then performed to determine significance between the techniques. P-values reported in this study come from the resultant Tukey HSD test. For accuracy, the percentage of correct answers was computed.

***Task 1*:** A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining

TABLE 3: Tukey HSD results for Task 1.

	p-value <	Point map	KDE map	Line-T
Time	Bristle-CLD	**.00001**	**.00001**	**.00042**
	Bristle-LD	**.00001**	**.00001**	**.01811**
	p-value <	Point map	KDE map	Line-T
Accuracy	Bristle-CLD	**.00001**	.1554	**.01851**
	Bristle-LD	**.00001**	.3214	**.00602**

TABLE 4: Tukey HSD results for Task 2.

	p-value <	Point map	KDE map	Line-TO
Time	Bristle-CLDO	**.01713**	.70091	.05943
	Bristle-LDO	**.02024**	.81621	.07166
	p-value <	Point map	KDE map	Line-TO
Accuracy	Bristle-CLDO	**.00001**	.07062	.36692
	Bristle-LDO	**.00001**	.99999	**.01283**

areas with highest crime rates within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was a significant effect of visualization type on time at the $p < .05$ level for the conditions $[F(4,145) = 35.366, p = .0000001]$ and a significant effect of visualization type on accuracy at the $p < .05$ level for the conditions $[F(4,145) = 3266.782, p = .0000000006]$. Because statistically significant results were found, we computed a Tukey post-hoc test with results reported in Table 3. In Table 3 p-values $< .05$ indicate that groups were statistically different from one another.

The result showed that the bristle maps groups were both significantly different than the point, color and line maps in terms of speed (at the $p < .05$ level). Specifically, the bristle map groups average times were 50.7 seconds and 56.6 seconds for the CLD and LD conditions respectively, which was slightly faster than the Line-T condition at 69 seconds and much faster than the point map condition at 102.6 seconds. However, the color map group was the fastest at 34.6 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group in terms of accuracy (at the $p < .05$ level). Specifically, the bristle map groups accuracy ratings were 99.6% and 99.8% for the CLD and LD conditions respectively, which was much higher than the point map condition with accuracy of 41.4%. No accuracy differences were found when compared to the other groups. See Table 8 for more specific results.

The comparison between color maps and bristle maps showed that color maps were better than the bristle map in terms of average time, and were not significantly different in terms of accuracy. This shows that bristle maps as a redundant encoding scheme has the same potential to convey data as single parameter encoding schemes; however, traditional schemes such as color maps may allow for a quicker comparison in the univariate case.

Comparing Bristle-LD and Line-T, we saw that the length of the bristle map matches the thickness of the line map. Hence, the bristle density was useful to find answers in Task 1 in terms of completion time and accuracy. Some participants also mentioned bristle density in their qualitative feedback as *"Bristle map is especially good when density of the bristles is also used"* and *"In bristle map, length and density were more noticeable than color difference."* In this univariate encoding test, the point map showed the worst results and the color map was the best results in terms of time and accuracy as shown in Table 8.

***Task 2*:** A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with highest crime rates at both day and nighttime within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was a significant effect of visualization type on time at the $p < .05$ level for the conditions $[F(4,145) = 2.717, p = .032]$ and a significant effect of visualization type on accuracy at the $p < .05$ level for the conditions $[F(4,145) = 89.89, p = .0000002]$. Because statistically significant results were found, we computed a Tukey post-hoc test with results reported in Table 4. In Table 4 p-values $< .05$ indicate that groups were statistically different from one another.

As we hypothesized, the result showed that the bristle maps groups were both significantly different than the point maps in terms of speed (at the $p < .05$ level). Specifically, the bristle map groups average times were 86.3 seconds and 87.2 seconds for the CLDO and LDO conditions respectively, which was slightly faster than the point map condition at 106.2 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group in terms of accuracy (at the $p < .05$ level). Specifically, the bristle map groups accuracy ratings were 90.5% and 93.3% for the CLDO and LDO conditions respectively, which was much higher than the point map condition with accuracy of 63.1%. See Table 8 for more specific results.

The comparison between color maps and bristle maps showed that color maps were better than the bristle map in terms of average time, and were not significantly different in terms of accuracy. This shows that bristle maps as a redundant encoding scheme has the same potential to convey data as single parameter encoding schemes; however, traditional schemes such as color maps may allow for a quicker comparison in the univariate case.

Findings also indicated that Bristle-LDO was better than Line-TO in terms of accuracy, whereas Bristle-CLDO was not significantly different from Line-TO in terms of accuracy. This indicated that the bristle density seems to be useful in finding correct answers in Bristle-LDO, but it was not in Bristle-CLDO. Further testing in combinations of visual variables and the ability to determine levels of sparseness will be done in the future.

***Task 3*:** A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas

TABLE 5: Tukey HSD results for Task 3.

	p-value <	Point map	KDE-B	Line-TO	Line-CT
Time	Bristle-LDO	**.00009**	**.00515**	.58128	.73239
	Bristle-CD	**.01131**	**.03469**	.15506	.20693
	p-value <	Point map	KDE-B	Line-TO	Line-CT
Accuracy	Bristle-LDO	**.00001**	**.00001**	.27189	**.02771**
	Bristle-CD	**.00001**	**.00001**	.07002	.41194

TABLE 6: Tukey HSD results for Task 4.

	p-value <	Point maps	KDE maps	Line maps
Time	Bristle-LDV	**.00001**	**.00001**	**.00001**
	p-value <	Point maps	KDE maps	Line maps
Accuracy	Bristle-LDV	**.00001**	**.00001**	**.00001**

with highest crime rates in two types of crimes within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was a significant effect of visualization type on time at the $p < .05$ level for the conditions $[F(5,174) = 6.655, p = .00001]$ and a significant effect of visualization type on accuracy at the $p < .05$ level for the conditions $[F(5,175) = 144.24, p = .00000001]$. Because statistically significant results were found, we computed a Tukey post-hoc test with results reported in Table 5. In Table 5 p-values $< .05$ indicate that groups were statistically different from one another.

The result showed that the bristle maps groups were both significantly different than the point maps and color maps in terms of speed (at the $p < .05$ level). Specifically, the bristle map groups average times were 88.2 seconds and 94.5 seconds for the LDO and CD conditions respectively, which was faster than the point map condition at 118.3 seconds and the color map condition at 115.3 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group and the color map group in terms of accuracy (at the $p < .05$ level). Specifically, the bristle map groups accuracy ratings were 94.4% and 90.4% for the LDO and CD conditions respectively, which was much higher than the point map condition with accuracy of 26.6% and the color map condition with accuracy of 72.6%. See Table 8 for more specific results.

Note that we separated parameters for different crime types in Bristle-CD; (C)olor encodes crime 1 and (D)ensity encodes crime 2. Bristle-CD showed a significant effect compared to the bivariate color map as shown in Table 5. However, generation on this type of bristle maps should be selected carefully since one parameter could dominate the other. For instance, when we use color and length to separate two crime data, short bristle length for low crime rates in crime 2 removes bristle lines in dark color for high crime rates in crime 1. In our experiment, we selected color and density for two crimes, with constant length of bristles.

***Task 4*:** A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with high temporal variance within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was a significant effect of visualization type on time at the $p < .05$ level for the conditions $[F(3,116) = 42.051, p = .00001]$ and a significant effect of visualization type on accuracy at the $p < .05$ level for the conditions $[F(3,116) = 42.33, p = .00001]$. Because statistically significant results were found, we computed a Tukey post-hoc test with results reported in Table 6. In Table 6 p-values $< .05$ indicate that groups were statistically different from one another.

The result showed that the bristle maps groups were both significantly different than the point maps, line maps and color maps in terms of speed (at the $p < .05$ level). Specifically, the bristle map groups average time was 48.4 seconds for the LDV condition, which was faster than the point map condition at 194 seconds, the color map condition at 171.8 seconds, and the line map condition at 178.9 seconds.

For accuracy, the bristle maps groups were both significantly different than the point maps, line maps and color maps in terms of speed (at the $p < .05$ level). Specifically, the bristle map groups accuracy rating was 94.7% for the LDV condition, which was much higher than the point map condition with accuracy of 53.6%, the color map condition with accuracy of 72.6% and the line map condition with accuracy of 75.5%. See Table 8 for more specific results.

As we hypothesized, we found that the representation of temporal variance in bristle maps was significantly faster and accurate in terms of both average time and accuracy compared to providing several images of the point, color and line maps. Moreover, we found that techniques showed the increasing pattern from the point maps to Bristle-LDV as shown in Table 8. This indicates that changes among several images would be better perceived in line patterns than in points or colors.

***Task 5*:** A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with high temporal variance within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was no significant effect of visualization type on time at the $p < .05$ level for the conditions $[F(1,56) = .328, p = .569]$ and no significant effect of visualization type on accuracy at the $p < .05$ level for the conditions $[F(1,56) = .315, p = .315]$. In Task 5, we found that bristle lines with and without clutter reduction did not differ significantly w.r.t. both average time and accuracy for all cases (Fig. 8). This means that the base bristle lines and bristle lines before applying clutter reduction and the base and bristle lines after applying our clutter reduction are perceived similarly by participants. Moreover, when told that the bristle line orientation does not encode data, the opposite orientations of bristle lines on a single straight

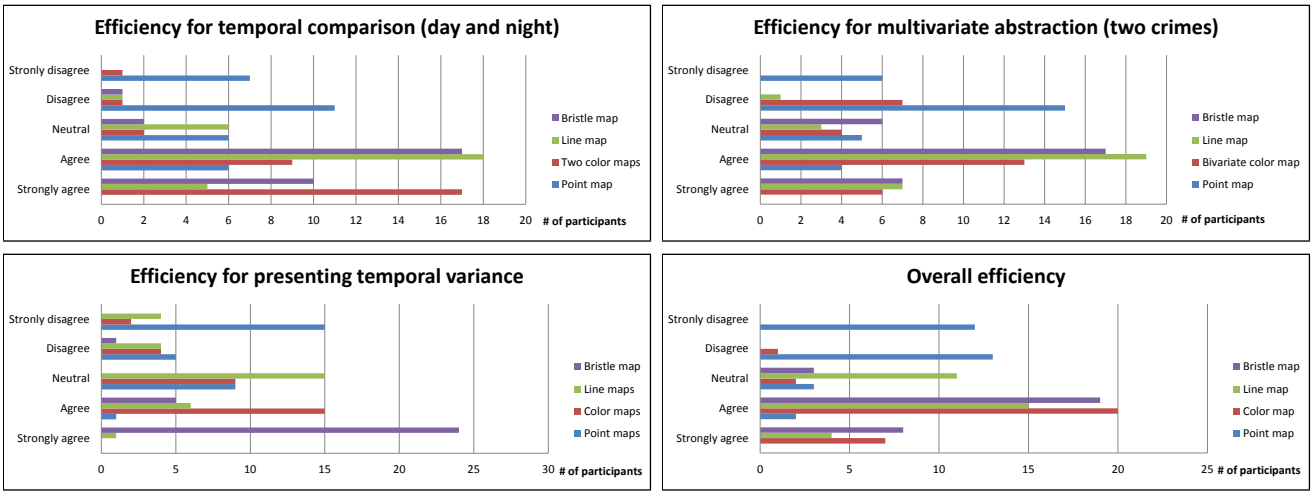


Fig. 10: Results from qualitative feedback for Tasks 2, 3, and 4 as well as overall efficiency.

TABLE 7: Average Rank Ordering by Aesthetics

	Point	KDE Map	Bristle	Line
Average	2.26	2.6	2.79	2.34
Std Dev	1.18	.97	1.19	1.10

road caused by virtual nodes (Fig. 7(b)) did not affect accuracy (87.7%). Other cases showed 42-58% of accuracy.

***Task 6*:** For Task 6, we hypothesized that subjects would be as accurate as all other representations in determining values In Task 6, we found that bristle maps did not differ significantly w.r.t. accuracy when compared with point map, color map and line map identification (ANOVA results of p-value=.18093, F=1.63). However, we found that bristle maps did differ significantly w.r.t. time when compared with point map, color map and line map (ANOVA results of p-value=.0314, F=2.622). Particularly, we found line maps and heat maps to both be significantly faster than point maps and bristle maps in identifying values (Tukey HSD test value of $p < .05$). Overall, these results indicate that in terms of accuracy, all geographical representations were equally useful; however, participants were (on average) over 1 second quicker in value judgments on both line maps and colors maps. This is most likely due to the fact that participants were quicker at making color judgments as compared to counting points and mentally linking multiple variables for the bristle maps.

***Task 7*:** In Task 7, we found that users had a highly variable rating of which image appeared to be more aesthetically pleasing. The average positions and standard deviations are summarized in Table 7. Here we find that while bristle maps have a slightly higher average ranking, there is no significant difference between the aesthetic ordering. A one-way between-subjects ANOVA was conducted to compare the rankings of map visualizations by subject in determining which visualization was ranked highest in aesthetics. There was no significant effect of visualization type on aesthetics at the $p < .05$ level for the conditions $[F(3,183) = 1.79, p = .149]$.

***Qualitative evaluation*:** Fig. 10 shows the results from qualitative feedback. Among the 30 participants, 27 participants (90%) agreed or strongly agreed that the bristle map was efficient for day and night time comparison in Task 2, 26 for two color maps and 23 for line map. 24 participants (80%) agreed or strongly agreed that the bristle map was efficient for the comparison of two crimes in Task 3, 26 for the line map and 19 for the bivariate map. 29 participants (96.6%) agreed or strongly agreed that the bristle map was efficient for temporal variance representation. In the question for overall efficiency, 27 participants (90%) agreed or strongly agreed that bristle maps and color maps were overall efficient, and 19 (63.3%) for line maps. For point maps, 25 participants (83.3%) disagreed or strongly disagreed.

Participants were also asked to answer visual complexity and preference questions regarding the before (NCR) and after (CR) image pairs applying our clutter reduction. For the circular case (Fig. 8(c)), 96.5% of participants felt that NCR has higher visual complexity and 78.5% preferred CR. For the curved case (Fig. 8(d)), 65.5% of participants answered that CR has a higher visual complexity and 64% preferred NCR. While both cases use a technically identical clutter reduction algorithm, participants reported different visual complexity and preference for them. This indicates that our clutter reduction could be improved by considering the complexity of the underlying network structure.

***Summary and Limitations*:** As a univariate encoding, the bristle maps were significantly different (in terms of speed and accuracy) than the point, color and line maps. In the case of the point and line maps, bristle maps use resulted in a higher average correctness and speed; however, the color map for the univariate case had the fastest response and accuracy totals. This seems to indicate that the redundant encoding scheme is actually not beneficial in these cases. As such, use of bristle maps for single variable encoding is not recommended.

With regards to bivariate and multivariate encoding,

TABLE 8: Average time and accuracy.

	Technique	Average time (seconds)	Accuracy (%)
Task 1	Point map	102.6 ± 40.9	41.4 ± 4.3
	KDE map	**34.6 ± 7.8**	**100 ± 0**
	Line-T	69 ± 21.9	98.1 ± 3
	Bristle-CLD	50.7 ± 15.3	99.6 ± 1.4
	Bristle-LD	56.6 ± 17.1	99.8 ± 1
Task 2	Point map	106.2 ± 34.7	63.1 ± 6.7
	KDE map	90 ± 30.9	93.1 ± 5.15
	Line-TO	100.5 ± 28.5	87.9 ± 10.4
	Bristle-CLDO	**86.3 ± 30.6**	90.5 ± 8.5
	Bristle-LDO	87.2 ± 28	**93.3 ± 5.2**
Task 3	Point map	118.3 ± 41.1	26.6 ± 12.5
	KDE-B	115.3 ± 47.9	72.6 ± 7.9
	Line-TO	**84 ± 22.8**	**96.4 ± 6.2**
	Line-CT	86.1 ± 22	86.8 ± 16.7
	Bristle-LDO	88.2 ± 23.7	94.4 ± 7.6
	Bristle-CD	94.5 ± 28.3	90.4 ± 16.7
Task 4	Point maps	194 ± 73.5	53.6 ± 17.6
	KDE maps	171.8 ± 58.8	61.9 ± 15.6
	Line maps	178.9 ± 61.8	75.5 ± 13.5
	Bristle-LDV	**48.4 ± 14.8**	**94.7 ± 13.4**

bristle maps and line maps outperformed color and point maps. This is not surprising as bristle and line maps are able to combine variables into a single image, where as in the case of point and color maps, the user must mentally combine the two images together. Bristle-(C)LD also showed a significant effect of the bristle density compared to Line-T. As a bivariate encoding, using orientation in bristle maps was not significant compared to two color maps. However, in the comparison with the bivariate color map, Bristle-LDO showed a significant effect in terms of average time and accuracy. As such, we have that Bristle-(C)LDO as a bivariate encoding scheme created a middle level of cognitive load in-between two color maps and a bivariate color map. Bristle maps also showed potential as a multivariate encoding technique in a single view. Based on the results in Task 3, a point map using various colors and a multivariate color map would considerably increase users' cognitive load. In Tasks 1-3, we also observed that there is no significant effect between the bristle maps using the different encodings. The representation of temporal variance in the bristle map was significantly different from other methods. Our results also showed the differences among point, color and line maps. Participants could better find the region with higher temporal variance when using line maps than using point and color maps. In the qualitative evaluation, 90% of the participants agreed or strongly agreed the overall efficiency of bristle maps to find answers. However, users also strongly preferred the color map in these cases as well.

Finally, we found that with regards to accuracy in identifying values, no technique outperformed any others. However, users were significantly faster in identifying values in both the color and line map scenarios. We hypothesize that in both cases the user focused only on the color, where as in the point map case they needed to count the points and in the bristle map case they needed to reconfirm the univariate value by double checking several of the encoding legends.

Overall, this technique would be recommended when encoding large amounts of multivariate spatio-temporal point data. As the number of point samples increase, aggregation techniques are need to allow for quick summaries of the data, and, as is evidenced by our studies, pure spatial location representation by glyphs results in too much overlap for accurate measurement and evaluation. As the number of variables increase, color map representations allow for the encoding of variables only along a single visual variable (resulting in bivariate color maps or small multiple plots).

In using multivariate encodings, it is extremely important to understand the interaction effects that the visual variables will introduce in one another. Research into the perceptual interactions among different visual variables was performed by Acevedo and Laidlaw [42]. They measured the perceptual interfernce of icon size, spacing and brightness, noting that brightness outperforms spacing and size while being subject to interferences from both spacing and size. Acevedo and Laidlaw also noted that spacing also outperformed size, which contradicted some previous results; however, this result seems to align with our participants noting that the bristle spacing was a useful cue. Their results were reportedly due to the spacing sampling along a sinusoidal curve. The sampling of our bristles follow a uniform pattern within classification bins. Thus, there seems to be sufficient scientific evidence to justify using sparsity as a discriminating variable in the case of the bristle maps; however, further studies on this are warranted. Stone [43] has also studies the effect of size in color perception, noting that color appearance changed dramatically with the size being viewed. As such, it may be better to utilize fewer map classifications (color bins) when using bristle maps in order to increase the perceptual distance between each color being visualized.

The main limitations of the bristle map technique is that the combinations of data encoding can potentially prove overwhelming for the designer, and a poor choice on variable encoding can result in a suboptimal visualization. In particular, previous studies have provided results that can be used to predict that certain combinations of visual variables will either enhance or impede map reading. For example, the combination of length and density form an emergent property akin to Bertin's definition of grain. Such effects cannot be ignored; however, bristle maps can be encoded to take advantage of such combinations, as shown in Tasks 3 and 4.

Finally, with regards to scalability of the bristle map technique, in areas of dense roadways, different aggregation methods would need to be considered. As the roads become dense, the ability to plot lines of perceptually different length would become untenable. However, a solution to this would be to draw only the most important roads, thereby removing smaller roads from the analysis, or utilizing bristle maps in a focus+context manner.

To summarize, we posited several different hypotheses. First, we hypothesized that our bristle maps would be

better than other techniques as the complexity level of tasks increased from univariate to multivariate. As the number of variables under analysis increased, the bristle maps outperformed more traditional analyses in terms of speed. However, at lower levels of complexity, traditional techniques such as density estimated heatmaps were found to be best in terms of speed. Second, we hypothesized that our bristle maps would be as accurate as all other representations in determining values. This hypothesis was verified through our user study where subjects estimated the magnitude of crimes on a map using points, density estimated heat maps and bristle maps. Results showed that there were no significant differences between map styles. Finally, we also hypothesized that bristle maps would be ranked higher in terms of their aesthetics. This hypothesis was refuted as our results did not demonstrate any significantly different rating for any of the different mapping techniques.

8 Conclusions

In this work, we have described our novel multivariate data encoding scheme, the Bristle Map. This scheme provides a novel approach for encoding color, length, density, and orientation as data variables and allowing the user to explore correlations within and between variables on a single view. Given the number of parameters available within this encoding, this article has presented only a subset of potential encodings and examples. Here, we have shown the use of encoding bristle lines with redundant information, multivariate attributes for variable comparison, and temporal variance. We also showed a means of potentially encoding data uncertainty. To minimize overlap of bristle lines, we generated a topology graph from underlying geographical line features and employed strategies for clutter reduction. Then, to evaluate the effectiveness of bristle maps, we performed an evaluation study, where we explored different visual encoding combinations within the bristle maps and compared with existing techniques in several tasks. Based on our experiment results, we believe that our bristle map technique has much potential to increase the amount of information that can be visualized on a single map for geovisualization.

Acknowledgments

The authors would like to thank Ahmad M. Razip for his help in setting up a web-based user study environment. This work was supported by the US Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. Jang's work was supported in part by the Industrial Strategic technology development program, 10041772, funded by the Ministry of Knowledge Economy (MKE, Korea). Isenberg's work was supported in part by a French DIGITEO chair of excellence.

References

[1] A. MacEachren, D. Xiping, F. Hardisty, D. Guo, and G. Lengerich, "Exploring High-D Spaces with Multiform Matrices and Small Multiples," in *Proc. InfoVis*. Los Alamitos: IEEE Computer Society, 2003, pp. 31–38. doi> 10.1109/INFVIS.2003.1249006

[2] C. North and B. Shneiderman, "Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata," in *Proc. AVI*. New York: ACM, 2000, pp. 128–135. doi> 10.1145/345513.345282

[3] C. Weaver, "Cross-Filtered Views for Multidimensional Visual Analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 192–204, Mar./Apr. 2010. doi> 10.1109/TVCG.2009.94

[4] D. S. Ebert, R. M. Rohrer, C. D. Shaw, P. Panda, J. M. Kukla, and D. A. Roberts, "Procedural Shape Generation for Multi-Dimensional Data Visualization," *Computers & Graphics*, vol. 24, no. 3, pp. 375–384, Jun. 2000. doi> 10.1016/S0097-8493(00)00033-9

[5] S. Bachthaler and D. Weiskopf, "Continuous Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1428–1435, Nov./Dec. 2008. doi> 10.1109/TVCG.2008.119

[6] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205–220, Mar./Apr. 2010. doi> 10.1109/TVCG.2009.100

[7] J. J. van Wijk and A. Telea, "Enridged Contour Maps," in *Proc. VIS*. Los Alamitos: IEEE Computer Society, 2001, pp. 69–74. doi> 10.1109/VISUAL.2001.964495

[8] R. J. Phillips and L. Noyes, "An Investigation of Visual Clutter in the Topographic Base of a Geological Map," *Cartographic Journal*, vol. 19, no. 2, pp. 122–132, Dec. 1982. doi> 10.1179/000870482787073225

[9] S. Openshaw, "The Modifiable Areal Unit Problem," in *Concepts and Techniques in Modern Geography*. Norwich, UK: Geo Books, 1984, vol. 38.

[10] M. Swink and C. Speier, "Presenting Geographic Information: Effects of Data Aggregation, Dispersion, and Users' Spatial Orientation," *Decision Sciences*, vol. 30, no. 1, pp. 169–195, Jan. 1999. doi> 10.1111/j.1540-5915.1999.tb01605.x

[11] C. L. Eicher and C. A. Brewer, "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation," *Cartography and Geographic Information Science*, vol. 28, no. 2, pp. 125–138, Apr. 2001. doi> 10.1559/152304001782173727

[12] R. Spence, *Information Visualization*. Reading, MA, USA: Addison-Wesley, 2001.

[13] L. Wilkinson, *The Grammar of Graphics*, 2nd ed. Heidelberg/Berlin: Springer-Verlag, 2005.

[14] A. M. MacEachren, *How Maps Work: Representation, Visualization, and Design*. Guilford Press, 1995.

[15] S. Chainey, L. Tompson, and S. Uhlig, "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security Journal*, vol. 21, no. 1–2, pp. 4–28, Feb.–Apr. 2008. doi> 10.1057/palgrave.sj.8350066

[16] B. W. Silverman, "Density Estimation for Statistics and Data Analysis," in *Monographs on Statistics and Applied Probability*. New York: Chapman & Hall, 1986, no. 26.

[17] C. Ahlberg and B. Shneiderman, "Visual Information Seeking using the FilmFinder," in *Proc. CHI*. New York: ACM, 1994, pp. 433–434. doi> 10.1145/259963.260431

[18] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Structure-Based Brushes: A Mechanism for Navigating Hierarchically Organized Data and Information Spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 2, pp. 150–159, Apr.–Jun. 2000. doi> 10.1109/2945.856996

[19] A. Dix and G. Ellis, "*By Chance*: Enhancing Interaction with Large Data Sets Through Statistical Sampling," in *Proc. AVI*. New York: ACM, 2002, pp. 167–176. doi> 10.1145/1556262.1556289

[20] A. MacEachren, "Visualizing Uncertain Information," *Cartographic Perspectives*, no. 13, pp. 10–19, Fall 1992.

[21] R. Dunn, "A Dynamic Approach to Two-Variable Color Mapping," *The American Statistician*, vol. 43, no. 4, pp. 245–252, Nov. 1989. doi> 10.1080/00031305.1989.10475669

[22] J. Olson, "Spectrally Encoded Two-Variable Maps," *Annals of the Association of American Geographers*, vol. 71, no. 2, pp. 259–276, Jun. 1981. doi> 10.1111/j.1467-8306.1981.tb01352.x

[23] A. MacEachren and D. DiBiase, "Animated Maps of Aggregate Data: Conceptual and Practical Problems," *Cartography and Geographic Information Systems*, vol. 18, no. 4, pp. 221–229, Oct. 1991. doi> 10.1559/152304091783786790

[24] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey, "Weaving Versus Blending: A Quantitative Assessment of the Information Carrying Capacities of two Alternative Methods for Conveying Multivariate Data with Color," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1270–1277, Nov./Dec. 2007. doi> 10.1109/TVCG.2007.70623

[25] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda, "Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data," in *Proc. InfoVis*. Los Alamitos: IEEE Computer Society, 2005, pp. 173–180. doi> 10.1109/INFOVIS.2005.35

[26] M. Sips, J. Schneidewind, D. A. Keim, and H. Schumann, "Scalable Pixel-Based Visual Interfaces: Challenges and Solutions," in *Proc. IV*. Los Alamitos: IEEE Computer Society, 2006, pp. 32–38. doi> 10.1109/IV.2006.95

[27] C. Panse, M. Sips, D. Keim, and S. North, "Visualization of Geo-spatial Point Sets via Global Shape Transformation and Local Pixel Placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 749–756, Sep./Oct. 2006. doi> 10.1109/TVCG.2006.198

[28] D. Dorling, A. Barford, and M. Newman, "Worldmapper: The World as You've Never Seen it Before," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 757–764, Sep./Oct. 2006. doi> 10.1109/TVCG.2006.202

[29] P. C. Wong, K. Schneider, P. Mackey, H. Foote, G. Chin, R. Guttromson, and J. Thomas, "A Novel Visualization Technique for Electric Power Grid Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 3, pp. 410 –423, May/Jun. 2009. doi> 10.1109/TVCG.2008.197

[30] D. Fisher, "Hotmap: Looking at Geographic Attention," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1184–1191, Nov./Dec. 2007. doi> 10.1109/TVCG.2007.70561

[31] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D Information Visualization for Time Dependent Data on Maps," in *Proc. InfoVis*. Los Alamitos: IEEE Computer Society, 2005, pp. 175–181. doi> 10.1109/IV.2005.3

[32] J. Bertin, *Semiology of Graphics*. Redlands, California: ESRI Press, 2011.

[33] J. Tarbell, "Substrate," Web site & simulation: http://www.complexification.net/gallery/machines/substrate/, 2003, accessed February 2012.

[34] T. Isenberg, "Visual Abstraction and Stylisation of Maps," *The Cartographic Journal*, vol. 50, no. 1, pp. 8–18, Feb. 2013. doi> 10.1179/1743277412Y.0000000007

[35] P. Rheingans, "Task-Based Color Scale Design," in *Proc. SPIE*, vol. 3905. SPIE, 2000, pp. 35–43. doi> 10.1117/12.384882

[36] C. Ware, "Color Sequences for Univariate Maps: Theory, Experiments and Principles," *IEEE Computer Graphics and Applications*, vol. 19, no. 5, pp. 41–49, Sep./Oct. 1988. doi> 10.1109/38.7760

[37] C. A. Brewer, *Designing Better Maps: A Guide for GIS Users*. Redlands, CA, USA: ESRI Press, 2005.

[38] M. Prasad, "Intersection of Line Segments," in *Graphics Gems II*, J. Arvo, Ed. Boston: Academic Press, 1991, pp. 7–9.

[39] A. Michotte, G. Thinès, and G. Crabbé, *Les Complements Amodeux des Structures Perceptives (Amodal Completions of Perceptual Structures)*. Louvain: Institut de Psychologie del'Université de Louvain, France: Studia Psychologica, 1964.

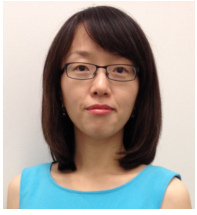
[40] M. S. Farrar, *Magic Squares*. Charleston, SC, USA: BookSurge Publishing, 1996.

[41] R. A. Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 5–55, 1932.

[42] D. Acevedo and D. Laidlaw, "Subjective Quantifi-

cation of Perceptual Interactions among some 2D Scientific Visualization Methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1133–1140, Sep. 2006. doi> 10.1109/TVCG.2006.180

[43] M. Stone, "In Color Perception, Size Matters," *IEEE Computer Graphics & Applications*, vol. 32, no. 2, pp. 8–13, Mar./Apr. 2012. doi> 10.1109/MCG.2012.37



SungYe Kim received her Ph.D. in Electrical and Computer Engineering from Purdue University in May, 2012. She also received her masters degree in Computer Science and Engineering from Chung-Ang University, South Korea in 2000. She is currently a graphics software engineer at Intel Corporation. Prior to this, she was employed as a research engineer at the Electronics and Telecommunications Research Institute from 2000 to 2006. Her research interests are computer graphics, illustrative visualization, visual analytics and information visualization.

Ross Maciejewski received his Ph.D. in Electrical and Computer Engineering from Purdue University in December, 2009. He is currently an Assistant Professor at Arizona State University in the School of Computing, Informatics & Decision Systems Engineering. Prior to this, he served as a Visiting Assistant Professor at Purdue University and worked at the Department of Homeland Security Center of Excellence for Command Control and Interoperability in the Visual Analytics for Command, Control, and Interoperability Environments (VACCINE) group. His research interests are geovisualization, visual analytics and non-photorealistic rendering.

Abish Malik is a Ph.D. student in the School of Electrical and Computer Engineering at Purdue University and a research assistant at the Purdue University Rendering and Perception Lab. He received his B.S. degree in Electrical Engineering from Purdue University in 2009. His research interests include visual analytics, correlation and predictive data analytics.

Yun Jang is an assistant professor of computer engineering at Sejong University, Seoul, South Korea. He received the masters and doctoral degree in electrical and computer engineering from Purdue University in 2002 and 2007, respectively, and received the bachelors degree in electrical engineering from Seoul National University, South Korea in 2000. He was a postdoctoral researcher at CSCS and ETH Zürich, Switzerland from 2007-2011. His research interests include interactive visualization, volume rendering, visual analytics, and data representations with functions. He is a member of the IEEE.

David S. Ebert is a professor in the School of Electrical and Computer Engineering at Purdue University, a University Faculty Scholar, director of the Purdue University Rendering and Perceptualization Lab, and director of the Purdue University Regional Visualization and Analytics Center. His research interests include novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually based visualization, illustrative visualization, and procedural abstraction of complex, massive data. Ebert has a PhD in computer science from Ohio State University and is a fellow of the IEEE and member of the IEEE Computer Society's Publications Board.

Tobias Isenberg is a senior research scientist with INRIA in France. He received is doctoral degree from the University of Magdeburg, Germany. Previously, he held positions as assistant professor for computer graphics and interactive systems at the University of Groningen, the Netherlands, and as postdoctoral fellow at the University of Calgary, Canada. He works on topics in interactive non-photorealistic and illustrative rendering as well as computational aesthetics and explores applications in scientific visualization. He is a member of the IEEE.

Business Intelligence from Social Media: A Study from the VAST Box Office Challenge

Yafeng Lu, Feng Wang, and Ross Maciejewski, *Member, IEEE*

Abstract—With over 16 million Tweets per hour, 600 new blogs posts per minute and 400 million active users on Facebook, businesses have begun searching for ways to turn real-time consumer based posts into actionable intelligence. The goal is to extract information from this noisy, unstructured data and use it for trend analysis and prediction. Current practices support the notion visual analytics can play a large role in enabling the effective analysis of such data. However, empirical evidence demonstrating the effectiveness of a visual analytics solution is still lacking. This paper presents a visual analytics system which extracts data from Bitly and Twitter to use for box office revenue and user rating predictions. Results from the VAST Box Office Challenge 2013 demonstrate the benefit of an interactive environment for predictive analysis compared to a purely statistical modeling approach. These visual analysis method used in our system can be generalized to other domain where social media data is involved, such as sales forecasting, advertisement analysis, etc.

Index Terms—social media, box office, visualization, prediction

1 Introduction

Social media data presents a promising, albeit challenging, source of data for business intelligence. Customers voluntarily discuss products and companies, giving a real-time pulse of brand sentiment and adoption. Unfortunately, such data is noisy and unstructured, making it difficult to easily extract real-time intelligence. Thus, the use of such data can be time-consuming and cost prohibitive for businesses. One promising current direction is the application of visual analytics. Recently, the visual analytics community has begun focusing on the extraction of knowledge from unstructured social media data [12]. Studies have ranged from geo-temporal anomaly detection [3], [4] to topic extraction [14] to customer sentiment analysis [5]. The development of such tools now enables end-users to explore this rich source of information and mine it for business intelligence.

One key area for business intelligence is revenue prediction. One means of revenue prediction is utilizing social media to understand product adoption and sentiment. Currently, very few tools exist that effectively enable the exploration of social media (such as Twitter) in conjunction with traditional business intelligence analytics (such as linear regression). Due to the abundance of social media discussions on movies, movie revenue prediction has drawn much attention from both the movie industry and academic field. Movie meta-data, social media data and google search volumes have all been explored in various prediction methods. For example, an early study by Simonoff et al. [13] predicted box office revenue with a logged response regression model using meta data features (e.g., time of year , genre, MPAA rating) as categorical regressors. Zhang et al., [15] demonstrated that regression models based on meta data features can be enhanced by utilizing variables extracted from news sources, and Joshi et al. [6] explored the relationship between film critic reviews and box office performance. Further work by Asur et al. [1] found that the rate of Tweets per day could explain nearly 80% of the variance in movie revenue prediction, and recent work from Google [10] claimed a 94% prediction accuracy in box office prediction by utilizing the volume of internet trailer searches for a given movie title.

While such methods have demonstrated the benefits of social media for extracting business intelligence for box office revenue prediction, they have relied solely on automated extraction and knowledge prediction. This paper presents our visual analytics toolkit for movie box office prediction. Our toolkit consists of a web-deployable series of linked visualization views that combine statistical techniques (multiple linear regression and time series modeling) with data mining (sentiment analysis) for predicting the opening weekend gross and viewer rating scores of upcoming movies. This type of visual analytics approach for social media analysis and forecasting can be directly applied to a wide range of business intelligence problems. Understanding how information is spread as well as the underlying sentiment of the messages being spread can give analysts critical insight into the general "pulse" of their brand or product. Developing a set of quick look visualization tools for an overview of such social media data along and linking this to models that business analysts generate for deploying new products, advertising campaigns and

- *Yafeng Lu, Feng Wang, and Ross Maciejewski, are with the School of Computing, Informatics and Decision Systems Engineering at Arizona State Unviersity.*
  E-mail: {lyafeng, fwang49, rmacieje}@asu.edu.
- *Visual Analytics and Data Exploration Research (VADER) Lab - http://vader.lab.asu.edu*

sales forecasts can be critical. Our toolkit can also be used to explore other business related social media data, for example, to see how well an ads campaign did and the pattern of information spreading. Some exploration can help adjust business decisions.

In order to demonstrate the effectiveness of our system, this paper reports on the results of the Visual Analytics Science and Technology (VAST) Box Office Challenge 2013. Results from this challenge also allowed us to explore the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution. Our results demonstrate that our analytics team was able to outperform the purely statistical model solution during the course of this contest; however, results from this study merely support the hypothesis that visual analytics can improve an end-user's analytic capabilities. More studies are required to create further convincing evidence.

2 Data Extraction, Analysis and Visualization Tools for Box Office Predictions

In order to explore the impact that visual analytics can have on generating insight into social media data, our work focused on box-office predictions using Twitter indices, bitly links, and access to the Internet Movie Database. This system is a web-enabled visual analytics toolkit that allows analysts to quickly extract, visualize and clean information from social media sources. These tools were combined with linear regression and temporal modeling for movie box office prediction and sentiment analysis for movie review rating prediction. In this section, we will discuss the various tools developed as well as lessons learned from the contest.

2.1 Tweet Mining – Overview, Sentiment and Cleaning Tools

While the tools developed are applicable to a variety of social media analysis problems, our specific application focused on structured data from the internet movie database (e.g., genre, budget, rating), and unstructured data from social media (e.g., Tweets, blog posts). While structured data is relatively straightforward to extract, unstructured data requires a large amount of pre-processing and manipulation. Unstructured data collected from social media revolved around movie related Tweets and bitly URLS. Tweets were collected for the two-week period prior to the release date based off the hashtag provided by a movie's official Twitter account. Our goal was to develop tools that could extract a variety of metrics from Twitter and IMDB (see the summary in Table 1 of the metrics we found most useful). Several of the extracted metrics required data mining and cleaning. To facilitate this, we developed tools that could present the volume of Tweets at various levels of temporal aggregation(Figure 1 (a)), enable users to remove unrelated Tweets from the aggregate metrics, and allow users to extract and manually adjust the sentiment of a Tweet (Figure 1 (b-d)).

TABLE 1: Variables Description

Variable	Description
OW	3-day Opening Weekend Gross
Budget	Approximate movie budget from IMDB. (unit is "million" of dollars)
Genre(category)	The movie's genre(s) according to IMDB
TUser	Number of unique users who tweeted about a movie
TBD	The average daily number of Tweets over the 2 weeks prior to release
TSS	Tweet Sentiment Score - A summation of each individual word's sentiment polarity as calculated via SentiWordNet [2]
MSS	Movie Sentiment Score - A derivation of the overall sentiment of a movie
MSP	Movie Star Power - A summation of the Twitter followers of the three highest billed movie stars (as listed by IMDB)

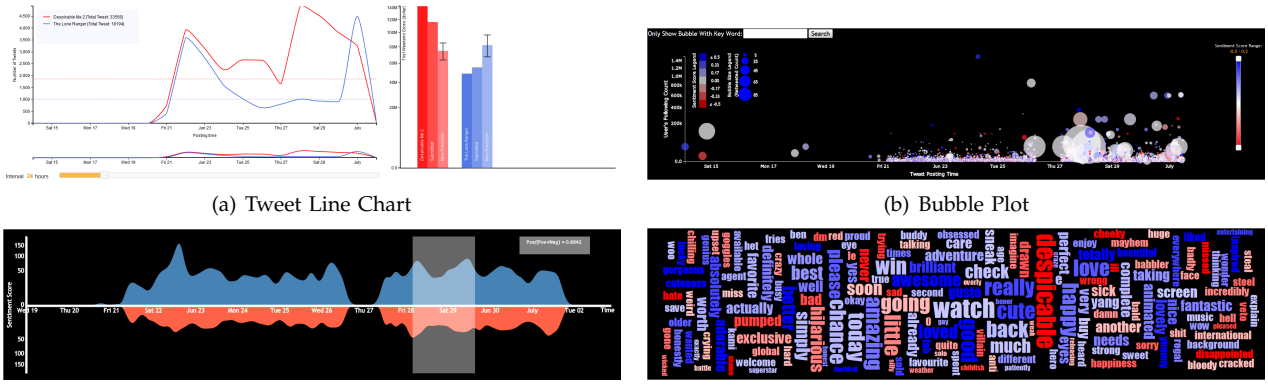
In order to approximate the popular sentiment of a movie, we processed each Tweet using a dictionary based classifier, SentiWordNet [2]. This process assigns each word in the Tweet with a score from -1 to 1 with -1 being the highest negative sentiment score and 1 being the highest positive sentiment score. Next, each Tweet is assigned a sentiment score by summing the sentiment score of all words in the Tweet and scaling the range from $-.5$ to $.5$ (TSS in Table 1). Finally, the movie sentiment score (MSS in Table 1) is calculated as

$$MSS = \frac{Positive\ Score}{Positive\ Score + Negative\ Score} \quad (1)$$

where *Positive Score* is the sum of all Tweets for a given movie with a TSS greater than zero and *Negative Score* is the absolute value of the sum of all Tweets for a given movie with a TSS less than zero.

Once the sentiment scores for Tweets were extracted, these values were then visualized to the end user. Figure 1 (b-d) shows the bubble plot view, the sentiment river view, and the sentiment wordle view. In the sentiment wordle view (Figure 1 (d)), the 200 most frequently mentioned words are extracted and visualized.

Both the bubble plot and the wordle plot enabled interactive searching and filtering by keywords and users. Users posting irrelevant messages could be removed from the Tweet count and mismatched sentiment could be modified by the end user. The primary use we found for the views in Figure 1 were for data cleaning. The primary lesson learned was that visualization tools are a necessity for data cleaning due to the noisiness of social media data and the problems inherent in sentiment matching using a sentiment dictionary (e.g., phrases such as "I want to see this movie so bad" are marked as negative due to the word "bad", and words such as "Despicable" give negative sentiment when they are merely references to a movie title). While the wordle view provided a quick way to assess the sentiment of popular words, it was necessary to hover over the bubble



(a) Tweet Line Chart

(b) Bubble Plot

(c) Sentiment River

(d) Sentiment Wordle

Fig. 1: Tweet trend and sentiment views for Despicable Me 2. (a) Line charts and bar graphs showing the number of Tweets per day and the predictions. (b) A Tweet bubble plot where blue represents positive sentiment and red represents negative. The size of the bubble represents the number of times a Tweet has been retweeted, the x-axis is time, and the y-axis is the number of followers that the user who submitted the Tweet has. (c) A sentiment river view where sentiment is aggregated over four hour intervals. Positive sentiment is plotted in red above the x-axis, negative in blue below. A user can select an area on the river to see the ratio of positive to negative sentiment. (d)A sentiment wordle where the size of the word represents the number of times it was used in a Tweet and color represents sentiment. By clicking a word, the bubble chart view will be filtered to only Tweets containing that word.

plot or open a Tweet list view through the search bar in order to fully explore the context of a Tweet. While such views were useful for data cleaning, our analysis approach (see Section 3) demonstrated to us that these views were more effective for cleaning and overview than for use in the model analysis. The critical need for tools to extract the correct metrics for regression modeling is a major hurdle that needs to be overcome in utilizing social media data for business intelligence. The bubble plot and wordle plot helped us to deal with the challenge of sentiment analysis and cleaning of noise from social media data.

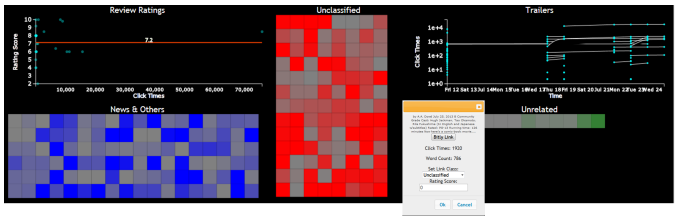


Fig. 2: Our interactive bitly classification widget. In the center are the unclassified links which the user can click and classify as seen in the floating window. The upper left is a plot of review score by click counts with a line for the average review score value.

2.2 Bitly Mining

While Tweets could be reasonably processed via the SentiWordNet dictionary, blog posts required a different approach. As part of this work, we explored long-form text by extracting bitly links containing movie keywords. These links typically consisted of review articles or news reports about the movies (or in many cases unrelated news, for example when the movie "The Heat" was released, the Miami basketball team, The Heat, had just won the NBA championship). For our review score prediction, we relied on prescreening review scores that were embedded in bitly links and developed an interactive tool for extracting these scores as shown in Figure 2. Initially, each bitly link starts as unclassified and is represented in a pixel matrix (color saturation corresponds to the number of times a link was clicked). By clicking on an unclassified square, a pop-up box appears with a brief bit of text from the article. The user can then choose to follow the link to scan the article for review scores and then manually assign a review score to an article or classify it as news or unrelated. A plot of review scores from articles versus the number of times an article was accessed is provided for analysis (see the upper left quadrant of Figure 2). This tool allows for quick data filtering and extraction, for example, reviews for the Star Trek video game can easily be separated from the Star Trek movie which would be difficult to automatically encode. Furthermore, the color coding from the pixel matrix can be used as a metric for classifying only those articles that had a substantial amount of views.

Similar to our lessons learned in Tweet mining, extracting information from bitly can be difficult to fully automate. As in the Star Trek example, multiple products for a movie may be released and reviewed at the same time. Furthermore, review scores may range from "two thumbs up" to "4 out of 5 stars" to "6 out of 10". With the analyst in the loop, these scores can be mapped to a user's own base system (in this case our metric was out

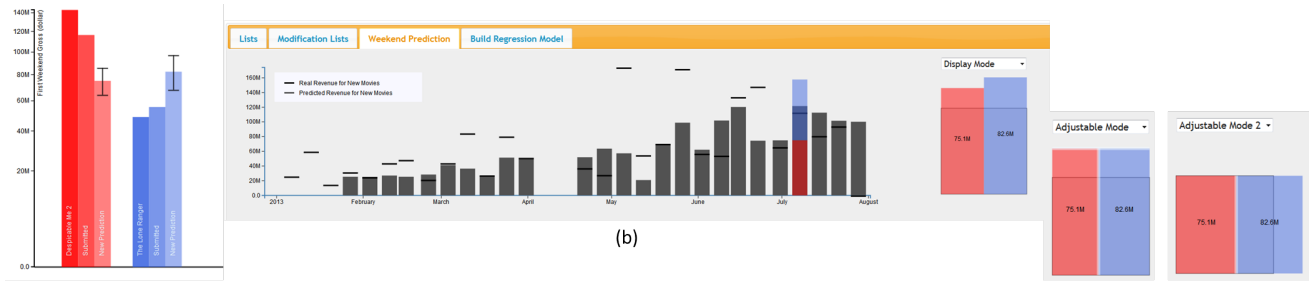


Fig. 3: The weekend prediction view for newly released movies and the prediction adjustment widget. This is the weekend when Despicable Me 2 and The Lone Ranger were released. (a) The bar graph view showing the actual value, submitted prediction and model prediction. (b) The stacked graph view showing the predicted weekend gross overlaid with the upcoming movie's regression model prediction. (c) The adjustment widget where users can modify the gross prediction; however, the predicted values for the new movies remain proportional. (d) The adjustment widget for changing individual predictions. The gray box represents the total weekend gross.

of 10).

2.3 Regression Modeling

Once data cleaning and variable extraction was complete, the next task was to use the social media metrics to develop a model for predicting box office revenue and review scores. Traditional variables used in these box office prediction models include structured variables (e.g., MPAA rating,movie budget) and derived measures (e.g., popularity of the movie stars, popular sentiment regarding the movie). Based on our initial literature search, we chose to utilize multiple linear regression for an initial prediction range for the opening weekend box office revenue (see the sidebar for a brief introduction to multiple linear regression modeling). We explored a variety of different variables that could be mined from the contest, see Table 1. After initial model fitting and evaluation using R [9], we found our best fit to be of the form:

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget + \varepsilon \qquad (2)$$

The model is updated weekly as new movies are entered into the data set. Parameters are fit using movie data beginning in January, 2013. Our first prediction was for the May 17th weekend and used data from 39 movies for training. Our weekly models reported an $R\text{-}adj^2 \approx 0.60$ with $p < .05$. Our final parameters were $\beta_0 \approx 4.9 \times 10^3$, $\beta_1 \approx 4462$, and $\beta_2 \approx 2.3 \times 10^5$.

The drawback of this model is that it does not fit the data overly well and predictions have a large variance. For comparison, a linear regression model using google search volumes was reported to explain more than 90% of the variance on box office performance [10], and models by Asur et al. [1]also report an $R\text{-}adj^2$ of over 90% when the number of theaters was used as a regressor. Our hypothesis was that a visual analytics toolkit could partially enable analysts to overcome poor data (partially due to the noise in social media data and partially due to the closed world nature of the contest). In order to facilitate better model prediction, we created a simple bar graph view (Figure 3(a)) which, for historical movies, showed the model prediction and its 95% confidence interval error range, our submitted prediction, and the actual box office gross. For new movies, only the model prediction and user submission was shown. This view was critical in our analysis process, and the primary view into the data consists of an overview of the Tweets per day and the model predictions of the movies under analysis as shown in Figure 1(a).

2.4 Temporal Modeling

While the regression model is able to provide one point for analysis, our goal was to also provide a big picture overview. For any given weekend, there is likely a maximum amount of money available in the market. In order to approximate the total amount of money available in the market, we employed a simple moving average model. Limitations here included access to data (historical weekend grosses were not available, and after a movie opens, further weekend takes were no longer reported in the contest). To compensate for this, we approximated subsequent weekend grosses for movies under the assumption that movies would run for three weeks following their opening weekend, and each weekend their box office take would be reduced by 50%. Thus, for any given weekend, we approximated the gross as:

$$\textit{Weekend Gross}(t) = \sum_{\forall_i} \mathrm{OW}_i(t) + \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t-j),$$

where t is the current weekend and i is the index to a movie that exists at time t. Then, for the weekend gross prediction, we use a moving average:

$$\textit{Weekend Gross}(t+1) = \frac{1}{3} \sum_{j=0}^{j=2} \textit{Weekend Gross}(t-j).$$

Finally, we approximate the available revenue for new movies as:

$$\textit{New Movie Gross}(t+1) = \textit{Weekend Gross}(t+1) - \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t+1-j).$$

While this prediction is crude, it provided the analysts with a valuable bound in which to explore the revenue predictions.

Results from the temporal weekend prediction and the linear regression models were then visualized in two different views as shown in Figure 3. The first view consists of a linked bar graph combined with stacked bars as shown in Figure 3 (b). The primary portion of the bar graph consists of light gray bars indicating the predicted total weekend market for the new movies and the dark gray short line indicates the actual weekend market for each calendar week whose date is shown on the x-axis. The stacked color bar graph is visualized only for the weekend under analysis, and the color design is the same as the movie's color in the prediction bar graph.

The second view, Figure 3 (c) and (d), is used to enable users to interactively adjust predictions while also visualizing the bounds of the total weekend prediction. In this view, a gray square is drawn, the area of which is scaled linearly to the total weekend prediction. Colored rectangles are superimposed onto the gray square, where the area of each colored rectangle represents the linear regression prediction for each movie being released on that weekend. If the sum of the individual predictions is equal to the total prediction, the colored rectangles will fit exactly into the gray square in both Figure 3 (c) and Figure 3 (d). The color design is the same as those of the bar graph, and modifying the size of a bar in any view will modify the size across all views.

Our system was designed to allow for three types of prediction adjustments.

1) Users are allowed to change the amount of the total gross prediction but the ratio between the movies will remain consistent.
2) Users are allowed to change the amount of an individual prediction but the total weekend prediction is kept consistent.
3) Users are allowed to arbitrarily change each movie's prediction and ignore the weekend gross.

By implementing and integrating multiple comparison methods, we found that we were able to quickly bound our analysis. While flexible, these bounds provided us with an early estimate of the total expected weekend gross in which to compare the predictions of our linear regression models. This multiple model comparison was a critical step for our overall box office prediction and was regularly used for all movie analyses.

While the results of our temporal predictions were of low quality, the combination of predictions and bounding of the problem space provided critical information for comparison and analysis. We will further discuss in Section 3 how the combination of both models was critical for successful predictions. Overall, the addition of multiple models predicting similar information can help guide analysts to a better ground truth. Similar to principles employed in the delphi method [11], where predictions are solicited from multiple experts and used to come to a common conclusion, in our system, we allow users to solicit predictions from multiple models to aid in their analysis. This bounded adjustment widget can be used in other hierarchical predictions which have both individual and total predictions, such as sub-topic trend prediction in a time period.

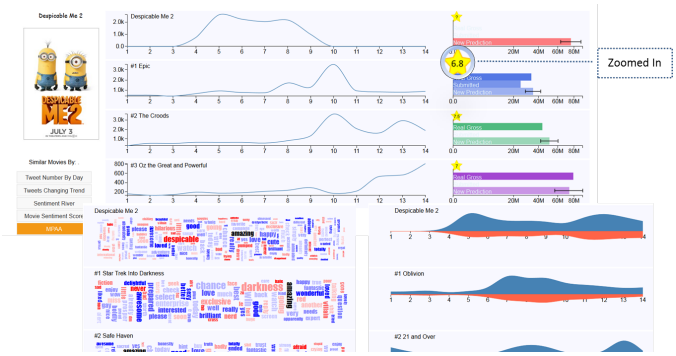


Fig. 4: A user defined similarity view cropped to show the topmost similar movies. In the center is the Tweets by day view, on the right is a graph of the opening weekend gross. There are bars for the actual gross, our final prediction, and the prediction range. The star in the upper left corner of the graph shows the review score.

2.5 Similarity Visualization

While bounding the movie predictions provided context for an overview of the total weekend, our other critical analytic view was the similarity widget. This widget enables analysts to quickly find and compare the accuracy of prediction based on various criteria of similarity. This allows analysts to determine if the given prediction model typically underestimates, overestimates or is relatively accurate with regards to movies that the analyst deems to be similar. In this manner, a user can further refine their final prediction value for both the box office gross and the review score. In this work, we have defined nine similarity criteria with distance calculation methods defined in Table 2. In all similarity matches, we show the top five most similar movies. These views allow users to directly compare Tweet trends and sentiment words between movies deemed to be similar in a category. Figure 4 contains snapshots from the Despicable Me 2 similarity page showing the line chart view with an MPAA similarity criterion, a wordle view with top word similarity criteria and a theme river view with sentiment similarity criteria.

While all of the variables used in our similarity metric could also be used in the linear regression model, results of the modeling indicated that these variables were not significant in altering the model. However, by providing an analyst with insight into these secondary variables,

TABLE 2: Calculations of Similarity Criteria

Similarity Criteria	Distance Measurement
Tweet Number by Day	$Dis(v,s) = \sum_{i=1}^{14} \lvert TBD_i(v) - TBD_i(s) \rvert$
Tweet Changing Trend	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{TBD\_i(v)}{Max(TBD\_j(v), j=1,2,...14)} - \frac{TBD\_i(s)}{Max(TBD\_j(s), j=1,2,...14)} \right\rvert$
Sentiment River	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{MSS\_i(v)}{Max(MSS\_j(v), j=1,2,...14)} - \frac{MSS\_i(s)}{Max(MSS\_j(s), j=1,2,...14)} \right\rvert$
MSS	$Dis(v,s) = \lvert MSS(v) - MSS(s) \rvert$
MPAA	same MPAA rating and close release date
Genre	$Dis(v,s) = 1 - \frac{card(Genre(v) \cap Genre(s)) \times 2}{card(Genre(v)) + card(Genre(s))}$
MSP	$Dis(v,s) = \lvert MSP(v) - MSP(s) \rvert$
Sentiment Wordle	$Dis(v,s) = 1 - \frac{card(SWordle(v) \cap SWordle(s))}{card(SWordle(v))}$

coupled with the temporal weekend modeling, further refinement of the prediction is made possible. For example, an analyst may compare the absolute difference between Tweets of two movies, or they can inspect the trend of the Tweets through line chart comparison using the Tweets Changing Trend similarity metric. This tool also allows users to quickly compare the current movies under analysis to recently released movies with the same Motion Picture Association of America rating, genre or movie stars popularity based on the number of Twitter followers a star has.

3 A Visual Analytics Process for Box Office Predictions

This system was used to predict 23 movies over the course of 3 months in the VAST 2013 Box Office Challenge. Our prediction process involved 3 steps. Our example prediction process focuses on the July 4th holiday in the United States when Despicable Me 2 and The Lone Ranger were released.

3.1 Movie Review Score Prediction Process

Our movie review score process centered around using the wisdom of the crowd for predicting an expected IMDB review score. For each movie, our process began by entering the bitly view and manually extracting review scores from bitly users who had done a prescreening of the movie (Figure 2). In the case of Despicable Me 2, the analysts manually classified the most clicked bitly reviews. The average value of all review scores extracted for Despicable Me 2 was 7.8. Once the average value is recorded we would then use the similarity view to compare to other movies. The movie review score is visualized as a star highlighting the review value in the corner of the bar graphs (Figure 4). Typically we would compare across genre, movie rating and sentiment to determine if we felt the average value extracted from bitly links was a reasonable prediction. In the case of Despicable Me 2, we compared to Monsters University as both movies were animated sequels. Monsters Universitys IMDB rating was 7.8 giving us confidence that our predicted value of 7.8 was reasonable. This same process was then performed for the Lone Ranger, and a viewer rating of 6.4 was predicted.

3.2 Movie Gross Prediction Process

Once the viewer rating was predicted, we then focused on determining the box office gross for the two movies. This weekend was challenging for two reasons. First, the data stream from the contest was broken, providing only 6 days worth of Tweets, and, second, the predictions were for a five-day weekend as opposed to the typical three-day weekend. Using the available data, we obtained a rough estimate for the Despicable Me 2 box office value in the range of $76M +/- $13M and $85M +/- $13M for The Lone Ranger. Next, we explore the expected three-day weekend total and see that our time series model approximates that $124M is available for the two movies for the three-day weekend. A quick look at Figure 3 shows that our regression predictions are well outside the bounds of the time series model prediction.

Given the misalignment between the two models, we begin exploring the similarity views to determine which movies The Lone Ranger and Despicable Me 2 are most similar to based on our predicted review score as well as various other metrics. We compare Despicable Me 2 to a variety of animated movies and we see that the predicted $73M is actually low when compared to animated movies such as Monsters University. Next, we explore various similarity views for The Lone Ranger and see that it is likely similar to World War Z, which had a weekend gross of $66M. However, World War Z's viewer rating was much higher at 7.4 than the predicted 6.4 for The Lone Ranger.

After looking at the available information, we determined that Despicable Me 2 should perform similarly to Monsters University, and we predicted a three-day gross of $85M. Based on our temporal prediction, this left only $39M for The Lone Ranger; however, given the other evidence, it seemed likely that The Lone Ranger would underperform. Finally, we took our three-day prediction values and linearly scaled them to be a five day prediction, resulting in a final five day prediction of $116.5M for Despicable Me 2 and $55.45M for The Lone Ranger. The actual three-day gross for Despicable Me 2 was $83.5M and $29M for The Lone Ranger. The actual five-day gross for Despicable Me 2 was $143M and $48.7M for The Lone Ranger, and the actual IMDB ratings were 7.9 for Despicable Me 2 and 6.8 for The Lone Ranger.

TABLE 3: Comparison with Peer Teams Predictions

Team	Gross Prediction				Viewer Rating			
	Entry	Average Error	STD	MRAE	Entry	Average Error	STD	MRAE
VADER(Interactive)	23	11.213	9.416	0.467	23	0.487	0.460	0.075
Team Prolix	23	16.466	15.195	0.424	20	0.82	0.640	0.129
Uni Konstanz Boxoffice	14	17.056	15.743	3.929	21	0.905	1.519	0.095
CinemAviz	21	17.219	17.677	1.970	21	0.738	0.559	0.114
Team Turboknopf	8	21.9	15.606	0.685	18	0.514	0.426	0.079
elvertoncf - UFMG	3	12.677	9.806	3.009	3	1.323	0.328	0.259
Philipp Omentisch	5	30.657	38.028	0.678	5	0.5	0.324	0.071
CDE IIIT	2	60.6	62.084	0.537	2	0	0	0

4 Results From VAST Challenge

Eight teams (our team being Team VADER) from various research institutes participated in the VAST Box Office Challenge. Data was also collected from 4 professional movie prediction websites. In this section, we compare our prediction performance with respect to peer teams from the VAST challenge and professional predictions.

4.1 Comparison with Peer Teams

Table 3 provides summary statistics of the performance of each team that participated in the VAST Box Office Challenge. For the gross prediction we report the average error (in terms of millions of dollars), the standard deviation (STD) of the average error term and the mean relative absolute error (MRAE), which is the percentage of bias deviating from the real value.

$$MRAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Prediction_i - RealValue_i|}{RealValue_i} \quad (3)$$

Similar values are reported with regards to predicting the IMDB rating (in the case of the IMDB rating, participants submitted a rating score from 1-10). These statistics can be interpreted by their magnitude, where smaller values indicate more accurate predictions. Data collected in Table 3 was provided to all challenge participants after the contest was closed.

In terms of average error and standard deviation, our team reported the lowest values in gross prediction across all teams. With respect to the MRAE for gross prediction and viewer rating, our results are slightly worse than Team Prolix (MRAE of .424 for Prolix compared with our .467), and similar in range to Philipp Omentisch, CDE IIIT and Team Turboknopf. While Team Prolix was able to achieve a smaller MRAE over the contest than our group, comparatively, they have a much larger average error and standard deviation indicating more inconsistency in their predictions.

With regards to the viewer rating prediction, our team had the lowest average error and MRAE of all teams with more than 5 submissions. CDE IIIT submitted two perfect predictions; however, those were CDE IIIT's only predictions making it difficult to determine if their methods would produce consistent results. With regards to the average error and standard deviation of the viewer rating, our team had similar results to Team Turboknopf, slightly besting them with regards to Average Error, but being slightly worse with regards to standard deviation.

TABLE 4: Comparison with Professional Predictions.

Prediction Source	Entry	Average Error	STD	Average MRAE
VADER (interactive)	21	12.729	9.425	0.285
VADER (No interaction)	21	23.051	22.011	0.501
boxoffice.com	21	8.538	7.466	0.191
filmgo.net	6	12.75	7.409	0.297
hsx	20	9.06	7.397	0.205
boxofficemojo	14	9.864	7.527	0.224

4.2 Comparison with Professional Predictions

In order to explore the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions we have also collected results from four professional prediction websites for comparison. For our comparison to the professional prediction websites, we again explore the results of the VAST Box Office challenge. Given that these results were collected and verified by the contest organizers, we feel this is an adequate means of justifying their validity. For the comparison in Table 4, only 21 movies are shown in the chart as two movies, The Bling Ring and The To Do List, were limited release movies which opened in only 5 and 591 theaters respectively and most expert prediction sites do not provide predictions for limited release movies. For each prediction, we followed the same general process as described in Section 3. As previously stated, the underlying linear regression model used in our system was significant with an R^2-$adj \approx .6$.

Results in terms of the MRAE are given in Figures 5 and 6 for the opening weekend gross and review score respectively. Figure 5 provides a comparison of our MRAE with that of several expert prediction websites. From Figure 5, it is clear that we outperformed the experts in the case of three movies (Epic, Hangover 3 and Fast and Furious 6), and in the case where we had the largest error (After Earth) we relied heavily on the analytical component with no interaction.

Table 4 gives the average error, standard deviation and MRAE for the predicted movies. What the results show is that for the model used, the predictions of our team utilizing an interactive tool were a dramatic

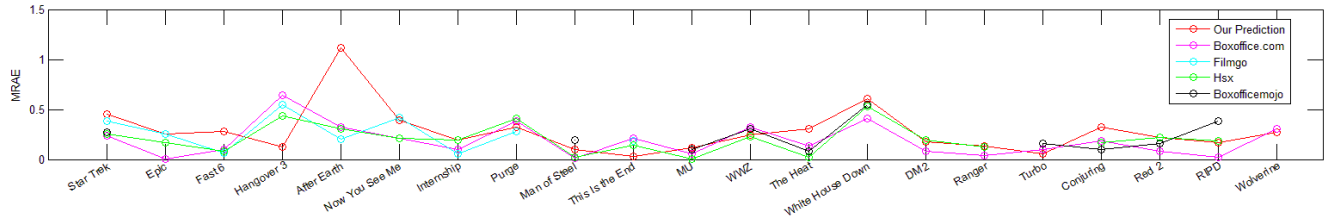


Fig. 5: The mean relative absolute error of box office weekend gross predictions, where the x-axis is the predicted movies.

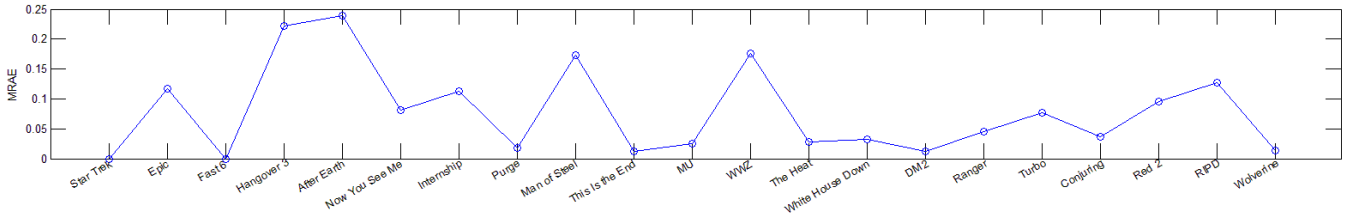


Fig. 6: The mean relative absolute error of our viewer rating predictions, where the x-axis is the predicted movies.

improvement over just the model itself (see Table 4 VADER (Interactive) versus VADER (No Interaction)). This provides a strong indication that the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution is valid. However, we do not wish to overstate our claims. This contest provides only a single data point for exploring how one group of analysts in a closed world setting were able to utilize a visual analytics toolkit for improved prediction. What this demonstrates is the need for further controlled studies in which a group of analysts perform similar model predictions and results are compared between analysts using a visual analytics platform and analysts using only results from a given regression model. However, results from the contest indicate that a visual analytics toolkit can enhance business intelligence.

Further analysis of the data also indicates that these tools enabled our team of novice box office analysts to quickly close the gap between the experts. Table 4 shows the average error and standard deviation for our predictions and compares them to four well known professional prediction websites. What we see is that both our average error and average MRAE are slightly lower than filmgo.net indicating that our methodology enabled our group of novice analysts to be competitive when compared to expert analysts. The significance of this relies on three major assumptions:

1) The professional prediction websites have more experience in box office prediction than our team.
2) The professional prediction websites have access to more data than our team was allowed in the closed world contest.
3) Access to more data can enable better predictive models as evidenced by [1], [6], [10], [15]

First, it seems reasonable that a professional prediction website would have much more experience than a computer science team who has never previously attempted to predict box office sales. Second, it is clear that utilizing data sources (specifically the number of theaters a movie is released in) will result in a better prediction model (a larger R^2). From these assumptions, it becomes clear that (in this instance) the application of a visual analytics toolkit can enable individuals that are knowledgeable with respect to data analysis to quickly understand information being presented to them in new domains and make predictions that are in line with expert predictions. Overall, our prediction error (.285) was slightly lower than that of filmgo (.297), but approximately 50% worse than boxoffice.com (.191). However, if we remove the After Earth and Now You See Me weekend (during which we relied heavily on the model and very little on the interactive visuals), our MRAE drops to .239 which puts us near the prediction range of boxofficemojo. Other sources of error can be accounted for in disrupted Twitter and bitly data feeds. These interruptions were pronounced for The Heat, White House Down, Monsters University and World War Z. However, even with those interruptions, our predictive analysis process was still quite robust with only The Heat being a significantly worse prediction than the professional sites.

5 Conclusions and Future Work

Overall, the application of visual analytics for social media analysis has proven relatively effective. However, there are still many challenges in applying this to all domains of business intelligence. First, social media data is extremely noisy. Movie predictions work well as one can track the effectiveness of ad campaigns by following the specific hashtags promoted by a brand. As

the analysis gets farther afield from Twitter (for example when trying to mine data from bitly) it becomes difficult to choose effective keywords. Second, due to the ever changing stream of social media sources and users, it is likely that any automated system for data collection and prediction will eventually be steered off course. As such, it is critical to link the human into the loop; however, as is evidenced by the issues in sentiment analysis, the data cleaning process should not overburden the analyst. The sentiment analysis and cleaning process employed in this work places an overly large burden on the end user. As such, integrating a system for having a user label a subset of tweets for sentiment model training could be a more effective solution. Third, it is imperative to link highly curated small datasets with this so call "big data". While social media data can be used as a proxy for many signals, we find that linking multiple data sources with varying levels of reliability (for example, total weekend take for all movies and regression modeling) can enhance the predictive abilities of a system. For example, doing focus groups and linking their data with results from social media could enhance the analysis of a proposed new product release. Finally, this paper demonstrates the need for interactive tools to mine social media data. From the examples of box office prediction, it is clear that such data contains a wealth of information. However, extracting knowledge from this data and effectively communicating this remains a challenge. There are clear needs for effective data cleaning tools to improve filtering of unrelated social media signals, as well as for improving the results of challenging analytical problems (such as sentiment analysis). Our results demonstrate that the use of visual analytics tools can have a significant impact on knowledge discovery for business intelligence.

While our results are able to only demonstrate a single data point, we feel this is significant in that the provisions of the contest allow us to directly compare a group of analysts using a visual analytics toolkit to experts in a particular modeling domain. However, we recognize that this is a far cry from definitively validating the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution. Overall, this work points for the need of better methods for evaluating the impact of visual analytics when used for complex problems such as prediction. There are a variety of factors and variables that need to be addressed and controlled, including the level of expertise and the types of visualizations provided. With our current system in place, we have been collecting streaming movie data in a manner similar to the VAST Box Office Challenge and plan to run a variety of controlled experiments. Of primary interest are exploring levels of expertise and the impact that visual analytics has on resultant predictions. We feel that results shown in this paper provide an important starting point for such explorations.

6 Sidebar: Linear Regression Model Construction and Evaluation

Regression analysis is one of the most widely used methods of pattern detection and multifactor analysis [7]. With a proper regression model, data can be better described, interpreted, and predicted.

6.1 Linear Regression Model

The basic form of a k-variable linear regression model is defined as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{4}$$

Variable y is known as the response, variables $x_i, i = 1, \ldots, k$ are the regressors and ε represents the error term. The goal is to define a relationship between the response term and the regressors by solving for the linear coefficients, β_i that best map the regressors to the response. The linear regression model is most often written in matrix form such that:

$$Y = X\boldsymbol{\beta} + \varepsilon$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

For multiple regression models, higher order terms may also be used to model the response (e.g., 2^{nd} order variables are of the form x_i^2 and $x_i x_j$). However, for this work our focus is on the simple linear regression model.

6.2 Parameter Estimation

In order to solve for the parameters β_i the ordinary least square (OLS) solution is most commonly employed. Note that this assumes normality for the data; however, if this assumption is not valid a maximum likelihood estimation would then be employed (which is equivalent to OLS under the assumption of normality).

For OLS, we wish to minimize

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

by satisfying

$$\frac{\partial S}{\partial \boldsymbol{\beta}} \mid_{\hat{\boldsymbol{\beta}}} = -2X^T y + 2X^T X\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

Under assumptions of normality, the solution takes the form of $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$ and the prediction function is $Y = HY$ where $H = X(X^T X)^{-1} X^T$. In one-order multiple linear regression, the predicted response is a linear combination of observations.

6.3 Model Selection

In a multiple variable dataset with a single response variable (such as in our box office gross prediction), analysts will traditionally be faced with a large set of potential linear regression models consisting of various regressors and orders. For example, in box office prediction, the response could be related to the number of Tweets per

day, or the number of theaters the movie is released in, or any combination of variables.

In order to decide which model should be used in prediction, there are several principles an analyst will typically consider.

- Do not violate the scientific principle, if there exists one, behind the dataset.
- Maintain a sense of parsimony to keep the order of the model as low as possible and the number of regressors as small as possible.
- Keep an eye on extrapolation. Regression fits data in a given regressor space but there is no guarantee that the same model also applies to other data outside this space.
- Always check evaluation plots more than the statistics. Residual plots and normal plots help show outliers and lack of fit.

In order to verify the efficacy of a model, analysts will typically rely on a variety of statistical graphics to determine the critical variables in the model, i.e., those that explain the most variation with the simplest form [8]. Several statistics are usually reported to evaluate the effective fit of a given model: p-value, R^2 and R^2-adj. The p-value shows the significance of a regression model, where $p < .05$ indicates the model is significant with a 95% confidence interval. R^2 and R^2-adj generally describe the percentage of variance explained by a given model. R^2-adj specifically takes the degree of freedom into consideration and should be used in multiple regression to compensate for the increased variance when adding regressors. A model is typically selected when it has a small p-value and a high R^2 or R^2-adj value and a relatively simple form with reasonable residual distributions.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. The authors would like to thank the VAST challenge organizers and participants for their help in data collection, evaluation and discussions.

REFERENCES

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499, 2010.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[3] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.

[4] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE Conference onVisual Analytics Science and Technology (VAST)*, pages 143–152, 2012.

[5] M. C. Hao, C. Rohrdantz, H. Janetzko, D. A. Keim, U. Dayal, L.-E. Haug, M. Hsu, and F. Stoffel. Visual sentiment analysis of customer feedback streams using geo-temporal term associations. *Information Visualization*, 2013.

[6] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296, 2010.

[7] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[8] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008.

[10] A. C. Reggie Panaligan. Quantifying movie magic with google search. *Google Whitepaper — Industry Perspectives + User Insights*, 2013.

[11] G. Rowe and G. Wright. The delphi technique as a forecasting tool: Issues and analysis. *International journal of forecasting*, 15(4):353–375, 1999.

[12] T. Schreck and D. Keim. Visual analysis of social media data. *Computer*, 46(5):68–75, 2013.

[13] J. S. Simonoff and I. R. Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.

[14] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-si: Scalable architecture for analyzing latent topical-level information from social media data. *Computer Graphics Forum*, 31(3pt4):1275–1284, June 2012.

[15] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 301–304, 2009.

Journal of Maps
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tjom20

Exploring geo-genealogy using internet surname search histories

Yifan Zhang [a], Muhammad Adnan [b], Paul Longley [b] & Ross Maciejewski [a]

[a] School of Computing, Informatics & Decision Systems Engineering, Arizona State University, Arizona

[b] Department of Geography, University College London, London

Published online: 30 Jul 2013.

To cite this article: Yifan Zhang, Muhammad Adnan, Paul Longley & Ross Maciejewski (2013) Exploring geo-genealogy using internet surname search histories, Journal of Maps, 9:4, 481-486, DOI: 10.1080/17445647.2013.824391

To link to this article: http://dx.doi.org/10.1080/17445647.2013.824391

Journal of Maps, 2013
Vol. 9, No. 4, 481–486, http://dx.doi.org/10.1080/17445647.2013.824391

SOCIAL SCIENCE

Exploring geo-genealogy using internet surname search histories

Yifan Zhang[a], Muhammad Adnan[b]*, Paul Longley[b] and Ross Maciejewski[a]

[a]*School of Computing, Informatics & Decision Systems Engineering, Arizona State University, Arizona;* [b]*Department of Geography, University College London, London*

(*Received 6 March 2013; resubmitted 3 June 2013; accepted 24 June 2013*)

We present an interactive flow map to visualize aspects of the ways in which surnames have dispersed and migrated around the globe. This work utilizes Internet search queries from the Worldnames Project and uses the density of search locations to determine the node and leaf structures of a flow map. The mapping technique utilized in this work is a variant of geometric minimal Steiner arborescences called the *spiral tree*. Our implementation is developed in JavaScript to allow for interactive online exploration. Nodes and flow lines can be interactively modified to allow for esthetic changes of color and layout. The results can provide interesting insight into the geography of amateur genealogy.

Keywords: surname; genealogy; flow map

1. Introduction

The study of family lineage and history is a well-established field of research, both amateur and professional, that uses historical records and genetic analysis to ascertain kinship. With the advent of digital encoding of historical records, and their wide availability through the Internet, more and more amateur genealogists are able to explore the global reach of their family lineage. A related innovation is that of 'geo-genealogy', in which the geographic origins and contemporary spatial distributions of surnames can be ascertained and mapped. For example, work by Cheshire et al. (2010) has explored the regional basis to surname distributions in Great Britain, using part of a database of names that is becoming increasingly global in coverage (http://gbnames.publicprofiler.org).

The focus of this map is the database that underpins the Worldnames website (worldnames.publicprofiler.org), which provides users with online maps of surname distributions at the global scale and a range of regional levels. Since its launch in August 2008, this website has been visited by more than 3.1 million unique users. In order to better measure the outreach and impact of the site, the location of the ISP address for every new search, along with the name searched for, has been recorded since July 2011. As of February 2013, a total of 318,838 paired names and locations have been stored.

The map accompanying this commentary extends the Worldnames mapping application by representing names searches as interactive flow maps. The information arising from name

*Corresponding author. Email: m.adnan@ucl.ac.uk

searches is likely to have been stimulated by historic migration of bearers of the searched for surnames across the globe. Users may access the online version of the map at http://www.uncertaintyofidentity.com/SurnameSearch.html in order to explore potential migration patterns of all surnames that have been searched for, in addition to those shown in our map. The results of surname searches are presented as flow maps in which the root of the flow identifies the highest share of the local population accounted for by the selected surname in any of the 26 countries for which surname data are available on the Worldnames site. These name 'origins' are represented as points as a way of summarizing the local and regional distribution of the name. In many cases, particularly when a name is rare and has a unique origin, this will identify the approximate location at which the name was first coined. In other cases, more common names (such as Smith or Brown) had multiple origins, and the point locations will thus summarize a wider distribution rather than providing an accurate point of origin. In such cases, the point location nevertheless provides an indication of the broader area in which a name originated. Further details on the origins of many established Anglo Saxon names in Great Britain may be found in the data pages of http://gbnames.publicprofiler.org, and similar information for other countries can be obtained from the tables accompanying the maps generated at http://worldnames.publicprofiler.org. The maps provide an indication of the Diaspora of many surnames and can stimulate users to suggest hypotheses about how their own family's history corresponds with the global pattern of searches.

Flow maps combine maps and flow charts as a means of showing the movement of objects from one location to another, making them ideal for communicating surname migration patterns from the location (or region) in which the name was first coined to the locations at which present day bearers of the name query the Worldnames database. One of the earliest examples of a flow map was Minard's map of French Wine exports in 1864 (Minard, 1864). Such maps proved informative and esthetically pleasing; however, the time needed draw them and the complexity in distributing flow lines often made the creation of such maps intractable. However, in recent years, thanks to increases in computational power, researchers have begun developing interactive computer-based tools for creating flow maps.

One of the earliest tools was Flow Mapper (Tobler, 1987, 2003) which allowed for the production of a total movement map shown by volume-scaled bands. Later work by Dodge and Kitchen (2004, 2007) utilized flow maps to explore the Internet infrastructures, and work by Cox, Eick, and He (1996) and Munzner, Hoffman, Claffy, and Fenner (1996) extended the concept of flow maps into 3D space. Guo (2009) developed an integrated interactive visualization framework utilizing flow maps to explore multivariate data, and Boyandin, Bertini, and Lalanne (2010) developed a tool for visualizing migration flows over time using animation.

Central to this approach is the development of algorithms for distributing flow lines in 2D geographic space such that they avoid (whenever possible) intersection and are esthetically pleasing. Work by Phan et al. (2005) developed an algorithm for defining the layout of flow maps utilizing hierarchical structures within the data. Cui et al. (2008) utilized control mesh methods for flow line layouts, and Holten and Van Wijk (2009) proposed a force-directed method for bundling edges. Our work utilizes the spiral-tree method introduced by Buchin, Speckmann, and Verbeek (2011) which utilizes *angle-restricted Steiner arborescences* for reducing visual clutter by bundling lines smoothly and avoiding self-intersection.

2. Methods

We have used two datasets for plotting surname searches made on the Worldnames site (http://worldnames.publicprofiler.org). Worldnames is a web service developed under a research project at Department of Geography, University College London, and holds surname data for

26 different countries of the world. These countries include a major part of Europe, USA, Canada, Argentina, India, Australia, and New-Zealand, and account for over a billion of the world's population. On each of the maps, the red dot identifies the location with the highest concentration of the surname in 26 countries that make up the Worldnames database.

Since July 2011, the Worldnames website has been collecting some additional information from users who search for a surname. This is the second database used in this study, which includes the IP addresses of the individuals who conducted the search, the name that they searched for, and the gender of the person conducting the search. Users are made aware that we collect email addresses and locations as evidence of the wide use of the service we provide, and to further our research into the geography of family names. The site terms and conditions also make clear that individual data are not passed on to third parties. Irrespective of gender, it is a reasonable assumption that the overwhelming numbers of users would be conducting searches that are related to their personal family histories. The archived IP addresses are converted to latitude and longitude values by using InfoDB's (http://www.ipinfodb.com/) IP address to Geo-location Application Programming Interface (API). For the maps that accompany this paper we have used the data archived between July 2011 and February, 2013. This database comprises 318,838 entries of surname searches and the corresponding IP address locations of the users. All of the remaining dots on the maps identify the locations at which searches originated for the relevant surname.

In order to create flow maps from these data, we have developed a modified JavaScript implementation of the spiral tree flow map. This method is a recent variant of geometric minimal Steiner arborescences, using a logarithmic spiral tree. The logarithmic spiral segment is an *angle-restricted* path assumed from a point p to its root node r, with the path's self-similar and self-approaching properties being defined by Aichholzer et al. (1998). Specifically, from point p to r, there exists two spirals defined as a right spiral S_p^+ and left spiral S_p^-. These two spirals can be given using parametric equation in polar coordinates assume $p = (R, \phi)$:

$$\text{Right spiral: } R(t) = \mathrm{Re}^{-t},\ \phi(t) = \phi + \tan(\alpha)t$$

$$\text{Left spiral: } R(t) = \mathrm{Re}^{-t},\ \phi(t) = \phi + \tan(-\alpha)t$$

where t is the parameter and α which less than $\pi/2$ is the angle of the spiral, when $0 \leq t \leq \pi\cot(\alpha)$ those two segments will construct a spiral region R_p.

If another point q lies within the region R_p denoted as $q \in R_p$, we will have $R_q \subseteq R_p$ as scaling a logarithmic spiral will result in another logarithmic spiral.

If another point $q \notin R_p$, then there will be an intersection denoted as join point J_{pq} between the spiral segments of p and q, which could be either S_p^+ with S_q^- or S_p^- with S_q^+.

The flow map in our application is then constructed using an enhanced greedy algorithm where each node (or terminal) represents a search for a specific surname on the Worldnames site. For esthetic purposes, we have modified the way in which the spiral tree algorithm creates terminals and joins points. The root of the flow map (colored red) is the location where the highest frequency of a given surname exists within the Worldnames database. An auxiliary circle AC_t is defined as the circle passing terminal t and centered at root r. There is also a set called wavefront W which is a list used for storing the active nodes. It is designed as a balanced binary search tree that organizes active nodes in counter-clockwise radial order around the root, in order to assist the procedure.

The algorithm iteratively joins terminals from the wavefront W until all terminals are connected into a single tree. All of the terminals are initialized as non-active and the wavefront W

is empty. A sweep circle C ranges from the outermost terminals inwards towards the center of root r in order to determine the next node to be processed. During the sweeping procedure, the sweep circle C will encounter two types of nodes: Terminal and Join Point, where a join point is only computed by two adjacent nodes in the wavefront W.

If C reaches a terminal t, t will be noted as an active terminal first and added to W. Then the algorithm examines whether or not t is in any of its neighbor's spiral region W. If there is a spiral region of its neighbor n, then W contains t, and our algorithm finds an auxiliary point x which marks the intersection of AC_t with the farthest spiral segment of n, and then connects x with n using the spiral segment. n is then removed from W and x is treated as a new terminal.

If C reaches a join point J_{pq} and its parent nodes p and q are active in W, then we join its parent nodes p and q to this join point using the spiral segments and remove p and q from W. Next J_{pq} will be noted as active and added to the wavefront W, and the sweeping process continues. At this step we have added a threshold function to position the parent node away from its join point such that if the distance of one parent node p to its join point J_{pq} of q is below a certain threshold, then we will utilize p's auxiliary circle AC_p to find the intersection point x with another spiral segment of node q. After that, q is connected to x using spiral segments and x is taken as a new active terminal point added to W. This is done in order to make terminals more visible to the viewer. In practice, the threshold is set to be 2 to the power of the map zoom level and the default values for the angle of the spiral α is 15 degrees. Users may interactively adjust these properties to create their desired map esthetics.

The JavaScript implementation of the spiral tree flow map is available online at the URL https://github.com/yifantastic/FlowMap.

3. Discussion

As previously stated, maps created using our tool represent the geography of interest in family genealogy, relative to the approximate locations in 26 countries at which different surnames were first coined. For names originating outside these countries, the origin point is defined at the location within the Worldnames' 26 countries at which the highest number of bearers of the name is concentrated. The precision with which the origin locations can be identified varies between names, although this is not an issue at the global scale at which our maps are produced. The maps also provide insights into the geography of interest in genealogy and the likely flows of information between bearers of the same name. This is potentially of use in the marketing of tourist destinations and heritage sites around the world.

Where the flow lines intersect, it is important to be aware that this is the result of the default parameters that are used by the algorithm, or those specified by the user. The curves depicted on the maps do not reflect actual migration flows but rather link a likely historic point in a family tree and the probable end destination of a migrant – ignoring any intervening points in family migration history. The maps may help suggest family migration histories based upon the spatial patterning of locations at which queries were submitted to the Worldnames database.

Our tool also allows users to map multiple flows simultaneously. In such cases, overlap between flows is not considered in the algorithmic layout. However, such overlays can allow users to search for multiple family branches and extrapolate information on potential encounters between family members.

4. Conclusion

This work presents an interactive tool for online exploration of interest in family genealogy. Our maps are illustrations of the outputs of our interactive website at http://www.uncertaintyofidentity.

com/SurnameSearch.html, which allows users to explore the geography of interest in the origins of approximately 130,000 surnames that were the subject of searches on the Worldnames website between July 2011 and February 2013. The branching algorithm that we have adapted presents this information in a clear and uncluttered way, even when large numbers of names searches have been conducted. This is particularly apparent in the online version of this map, which may be viewed at a full range of recursive levels.

Software

The online application that forms the basis to the maps was written in JavaScript. Our software is linked to a back-end database which stores the surname dataset in MySQL and an interactive front-end interface which can be accessed via web-browsers. The data are queried and transferred using AJAX (Asynchronous JavaScript and XML). The flow map representation consists of the background Google map overlain with the SVG network flows.

The Google Map JavaScript API is used for pre-processing the data from the geo-space into the 2D-rendering space and combining the flow layers with the interactive map, available at http://www.uncertaintyofidentity.com/SurnameSearch.html. We also utilized the D3 JavaScript library (http://d3js.org) to generate multiple SVG flow layers and jQuery User Interface library (http://jquery.com) for implementing the user interface. This makes it possible to create multiple tabs for searching different surnames. Two sliders are embedded into the interactive map frame so that users can manipulate the appearance of the flow by adjusting the stroke-width and angle of the spiral tree. Additional information is provided for clickable nodes linked with pop-up windows. Color palette widgets are provided using the JSColor JavaScript library (http://jscolor.com) so that different colors can be applied to distinguish multiple flows.

Acknowledgements

This work was completed as part of the EPSRC research Grant 'The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds' (EP/J005266/1).

References

Aichholzer, O., Aurenhammer, F., Icking, C., Klein, R., Langetepe, E., & Rote, G. (1998). Generalized self-approaching curves. *Algorithms and Computation*, *Springer*, 317–327. doi: 10.1007/3-540-49381-6_34

Boyandin, I., Bertini, E., & Lalanne, D. (2010). Using flow maps to explore migrations over time, In *Proceedings of Geospatial Visual Analytics Workshop in conjunction with the 13th AGILE International Conference on Geographic Information Science (GeoVA)*, Guimaraes (Portugal).

Buchin, K., Speckmann, B., & Verbeek, K. (2011). Angle-restricted steiner arborescences for flow map layout. In *Abstracts of the 27th European Workshop on Computational Geometry*, Springer, pp. 163–166.

Cheshire, J. A., Longley, P. A., & Singleton, A. D. (2010). The surname regions of Great Britain. *Journal of Maps*, *6*, 401–409. doi: 10.4113/jom.2010.1103. URL http://www.tandfonline.com/doi/abs/10.4113/jom.2010.1103

Cox, K. C., Eick, S. G., & He, T. (1996). 3D geographic network displays. *ACM Sigmod Record*, *25*, 50–54. doi: 10.1145/245882.245901

Cui, W., Zhou, H., Qu, H., Wong, P. C., & Li, X. (2008). Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*, 1277–1284. doi: 10.1109/TVCG.2008.135

Dodge, M., & Kitchin, R. (2004). Charting movement: mapping internet infrastructures, *Moving People, Goods and Information in the 21st Century: The Cutting Edge Infrastructures of Networked Cities*, Routledge, pp. 159–185.

Dodge, M., & Kitchin, R. (2007). *Atlas of cyberspace*. Download from www.kitchin.org/atlas/contents.html

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, *15*, 1041–1048. doi: 10.1109/TVCG.2009.143

Holten, D., & Van Wijk, J. J. (2009). Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, *28*, 983–990, URL http://dblp.uni-trier.de/db/journals/cgf/cgf28.html#HoltenW09

Minard, C. J. (1864). Carte figurative et approximative des quantités de vin français exportés par mer en 1864. *Lithograph* (835 × 547).

Munzner, T., Hoffman, E., Claffy, K., & Fenner, B. (1996). Visualizing the global topology of the MBone. In *Proc. IEEE Symposium on Information Visualization (INFOVIS '96)*, IEEE, USA, pp. 85–92. URL http://dblp.uni-trier.de/db/conf/infovis/infovis1996.html#MunznerHcF96

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., & Winograd, T. (2005). Flow map layout. *IEEE Symposium on Information Visualization (INFOVIS'05)*, 219–224, URL http://dx.doi.org/10.1109/INFOVIS.2005.13

Tobler, W. R. (1987). Experiments in migration mapping by computer. *Cartography and Geographic Information Science*, *14*, 155–163. doi: 10.1559/152304087783875273. URL http://www.ingentaconnect.com/content/cagis/cagis/1987/00000014/00000002/art00005

Tobler, W. R. (2003). *Movement mapping, center for spatially integrated social science*. URL http://www.csiss.org/clearinghouse/FlowMapper/

VAST 2013 Mini-Challenge 1: Box Office VAST - Team VADER

Yafeng Lu*
Arizona State University

Feng Wang
Arizona State university

Ross Maciejewski, *Member, IEEE*
Arizona State University

ABSTRACT

VAST 2013 Mini-Challenge 1 was a rolling competition in which participants would submit a set of weekly predictions for movie revenue and viewer rating. This was a closed world contest in which contestants were provided with a set of Twitter indices, bitly links, and access to the Internet Movie Database. In order to facilitate these predictions, we have created a web-deployable system that combines predictive techniques (multiple linear regression), data mining (sentiment analysis), and interactive visualizations for predicting the opening weekend gross and viewer rating scores of upcoming movies. Taking use this system with human in the loop, we got 4 times top predictor in 11 entries of submission.

Keywords: Multivarite regression, Box office prediction, Social media data, Twitter, Visual analytics

1 INTRODUCTION

Movie revenue prediction has drawn the most attention in both movie industry and academic field. Movies meta-data, social media data, text extraction and analysis, and google search volume have been explored in many prediction methods. In this work, we explore augmenting traditional box office modeling techniques through the use of visual analytics.

2 MULTIPLE LINEAR REGRESSION MODELING

Regression analysis is one of the most widely used methods for pattern detection, and a substantial body of literature exists on developing multiple linear regression models [2] for movie revenue prediction (e.g., [4, 5]). Traditional variables used in these box office prediction models include known variables (e.g., the MPAA rating, the number of screens) and derived measures (e.g., popularity of the movie stars, popular sentiment regarding to the movie).

Based on our initial literature search, we chose to utilize multiple linear regression for an initial prediction range for the opening weekend box office revenue. Due to the closed world nature of the contest, traditional variables used by other researchers were not always available (for example, theater count is not provided for every movie in IMDB). As such, we explored a variety of different variables available in the contest, see Table 1.

After initial model fitting using R [3], we found the best model to predict the opening weekend gross from our variables was:

$$OW = \beta_0 + \beta_1 \text{TBD} + \beta_2 \text{Budget}.$$

Parameters are fit using movie data released beginning in January of 2013. Our first submitted prediction was for the May 17th weekend and used data from 39 movies for training. Each week model parameters were updated with the newly collected data, and our weekly models reported an $adjusted - R^2 \approx 0.60$ with $p < .05$. Our final parameters were $\beta_0 \approx 4.9 \times 10^3$, $\beta_1 \approx 4462$, and $\beta_2 \approx .23$.

*email: lyafeng,fwang49,rmacieje@asu.edu

Variable	Description
OW	3-day Opening Weekend Gross
Budget	Approximate movie budget from IMDB
TBD	The average daily number of tweets over the 2 weeks prior to release
TSS	Tweet Sentiment Score - A summation of each individual word's sentiment polarity as calculated via SentiWordNet [1]
MSS	Movie Sentiment Score - A derivation of the overall sentiment of a movie
MSP	Movie Star Power - A summation of the twitter followers of the three highest billed movie stars (as listed by IMDB)

3 TWEET SENTIMENT MINING

While multiple regression was found to be a reasonable starting point for predicting the box office gross, predicting the review score required text analysis. We wanted to approximate tweet sentiment as it relates to upcoming movies. First, each tweet is processed according to the SentiWordNet dictionary [1] , where each word in the tweet is assigned a score from -1 to 1, where -1 is the highest negative sentiment score and 1 is the highest positive sentiment score. Next, each tweet is assigned a sentiment score by summing the sentiment score of all words in the tweet and normalizing the range from $-.5$ to .5 (TSS in Table 1). Finally, the movie sentiment score (MSS in Table 1) is calculated as

$$MSS = \frac{Positive\ Score}{Positive\ Score + Negative\ Score}$$

where *Positive Score* is the sum of all tweets for a given movie with a TSS greater than zero and *Negative Score* is the absolute value of the sum of all tweets for a given movie with a TSS less than zero.

4 VISUALIZATION TOOLS FOR FILTERING AND ANALYSIS

Part of the challenge in utilizing twitter and bitly data for prediction is the noise inherent in the data. In order to deal with this challenge, we have developed a variety of visualizations that allow for interactive analysis and filtering for noise reduction.

4.1 Tweet Bubble Chart

For sentiment analysis, some tweets follow a pattern of "I want to see this so bad." In this example, 'bad' will be classified as a negative sentiment word with the rest being neutral; however, this statement is actually quite positive about the movie. Thus, it is clear that a completely automatic analysis of tweet sentiment could result in large errors. To compensate for this, we have developed an interactive bubble chart, Figure 1, which allows users to modify the sentiment of any given tweet.

In Figure 1, the user has moused over a tweet and sees that the user is wondering if the movie will be good. From there, the user can interactively set this sentiment to neutral, which will update the database and all associated models using that variable. As part of the bubble chart widget, users can filter by keyword, thus in the example of "I want to see this so bad," if there is a common keyword that is causing sentiment misclassification, the user can quickly adjust the overall sentiment score. Tweets can also be filtered based on their overall sentiment score using the legend.

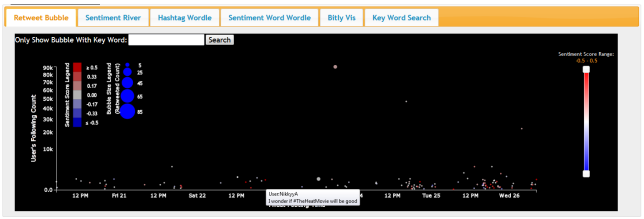


Figure 1: Our sentiment tweet bubble chart for The Heat. Each circle is a tweet. Red represents positive sentiment, blue negative. The size of the bubble represents the number of times a tweet has been retweeted, the x-axis is time, and the y-axis is the number of followers that the user who submitted the tweet has.

Along with sentiment noise, there is also noise in the underlying data collection itself. A specific example of this would be the movie, "The Heat". This movie shares its name with the popular professional basketball team, The Miami Heat. To further compound issues, this movie was released near the same time that the Miami Heat won the NBA championship, resulting in a large number of tweets referencing "The Heat" being unrelated to the movie. The bubble chart tool can also be used to mark tweets for removal from the dataset by right clicking on a tweet, thus updating the TBD variables from Table 1 which in turn updates the regression model. A list view of all modified tweets is also available to allow collaborative analysis in which multiple analysts can see modifications and adjust items based on their own expertise.

4.2 Similarity Visualization

The most used feature of our system was interactive similarity matching. The goal was to enable users to see which past movies are similar to the current film under prediction, then we can see if our model typically underestimates, overestimates or is relatively accurate. This allows us to refine our final prediction value for both the box office gross and the review score.

We have nine similarity criteria: TBD, tweet trend, tweet sentiment, MSS, MPAA rating, genre, MSP, sentiment wordle and predicted gross. The distance between two movies, when using any similarity criteria, is calculated using a Euclidean distance metric. In terms of categorical similarity (such as MPAA), we show the most recently released movies that have the same MPAA rating. In all similarity matches, we show the top five most similar movies (Figure 2 is cropped to show the top-most similar movie).

5 VADER Prediction Results

From May 17th through July 26th, we predicted the box office gross and review score for 23 movies (only 21 are shown in the chart as two movies (The Bling Ring and The To Do List) were limited release movies). Our typical process began with utilizing the bubble chart to adjust the data for noise and sentiment values as a means of data cleaning for the regression model. Next, the bitly widget was used to extract review scores from web articles. The average review score was analyzed and a mental model for the review score prediction was developed. Subsequently, the similarity widget was used to compare previous opening weekend predictions and review scores of similar movies. Finally, the opening weekend gross predictions were then summed and the total was compared to the total weekend prediction for final adjustments.

Average errors for all predictions are given in Table 2. Note that we provide the average error for using our regression model with no modification in Table 2 and we see that it is nearly double that of our final prediction informed through visual analytics. After all, the model only explains 60% of the variance of the data. Overall, our prediction error (.281) was only a slightly higher margin of

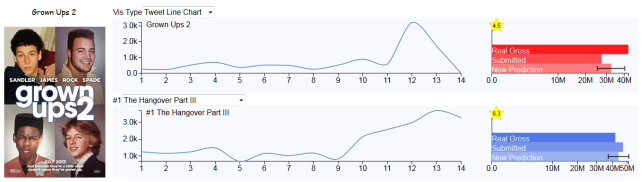


Figure 2: A user defined similarity view cropped to show the top-most similar movie. In the center is the tweets by day view, on the right is a graph of the opening weekend gross. There are bars for the actual gross, our final prediction, and the prediction range. The star in the upper left corner of the graph shows the review score.

Prediction Source	Average MRAE
Our Submitted OW Prediction	0.281
Regression Model OW prediction	0.447
boxoffice OW Prediction	0.183
filmgo OW Prediction	0.265
hsx OW Prediction	0.205
boxofficemojo OW Prediction	0.230
Our Submitted Viewer Rating prediction	0.078

error than filmgo (.265), but approximately 50% worse than boxoffice.com (.183). However, if we remove the After Earth and Now You See Me weekend which with prediction uses only the model, our MRAE drops to .239 which puts us solidly in the prediction range of several experts (HSX and boxofficemojo). Other sources of error can be accounted for in disrupted twitter and bitly data feeds. These interruptions were pronounced for The Heat, White House Down, Monsters University and World War Z. However, even with those interruptions, our predictive analysis process was still quite robust.

6 Summary

It is clear that visual analytics can be used to improve predictive models by bringing domain knowledge into play. While our systems visuals are relatively traditional, the combination of these with analytical methods has proven very effective. Of the 11 weeks in which contest entries were submitted, our team was the top predictor four of the weeks, our predictions had a lower MRAE than the pros in week 2 (as well as for multiple movies over the course of the contest, and we were consistently one of the top 3 teams for overall prediction. Currently no other team has been the top predictor more than twice. For a demonstration of the full system, please see the accompanying video at `http://www.youtube.com/watch?v=dQuti7aHvIw&fmt=22`.

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[3] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008.

[4] A. C. Reggie Panaligan. Quantifying movie magic with google search. *Google Whitepaper — Industry Perspectives + User Insights*, 2013.

[5] J. S. Simonoff and I. R. Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.

EUROGRAPHICS 2014 / M. Pohl and J. Roberts
(Editors)

Volume 0 (1981), Number 0

What's In a Name? Data Linkage, Demography and Visual Analytics

Feng Wang[1], Jose Ibarra[1], Muhammad Adnan[2], Paul Longley[2] and Ross Maciejewski[1]

[1]Arizona State University
[2]University College London

Abstract

This work explores the development of a visual analytics tool for geodemographic exploration in an online environment. We mine 78 million records from the United States public telephone directories, link the location data to demographic data (specifically income) from the United States Census Bureau, and allow users to interactively compare distributions of names with regards to spatial location similarity and income. In order to enable interactive similarity exploration, we explore methods of pre-processing the data as well as on-the-fly lookups. As data becomes larger and more complex, the development of appropriate data storage and analytics solutions has become even more critical when enabling online visualization. We discuss problems faced in implementation, design decisions and directions for future work.

1. Introduction

Family names (surnames) are a widely recorded marker for spatially-referenced population datasets. A surname can provide relevance to historical geography, genealogy and even population genetics. For example, work from Mateos et al. [MLO11] created global naming networks by generating linked forename-surname pairs revealing cultural naming practices for new and existing communities. Recent work from Cheshire and Longley [CL12] explored methodologies for identifying spatial concentrations of surnames. Their initial work focused on the development of an automated methodology for classifying the spatial distributions in surnames focusing on Great Britain [CLS10, LCM11]. Cheshire and Longley's work was later extended to 25 other countries (e.g., [CLYN13]), and an international surname mapping site (worldnames.publicprofiler.org) was created. This previous work in exploring demographics through names has primarily focused on classification methods and used visualization only as a means of displaying final results.

In this work, we extend the functionality of the worldnames profiler to explore not only the spatial distribution of names, but also linked demographic data. Our work focuses specifically on the United States, mining over 78 million records from the 2008 United States public telephone directories. Addresses are geocoded and then automatically linked to demographic data (specifically income distributions) from the United States Census bureau [U.S13]. Similar to the worldnames profiler, our tool (Figure 1) allows users to query surnames and see a density estimate distribution of the surname. Extensions include:

1. The ability to visualize and explore spatially similar names through a linked wordle of surnames where the size and color relates the spatial similarity of a surname;
2. The ability to visualize the estimated income distribution for a name based on census data, and;
3. The ability to explore the similarity between surnames based on income distributions through a linked wordle of surnames where the size and color relates the income distribution similarity of a surname.

While the visualizations provided are well known, the data linkage and integration of interactive analytic methods for comparing similarity is novel. Such a tool can provide unique insights into genealogy, demographics and social mobility. Furthermore, the challenge of distributing an online visual analytics tool for moderately large data provides an opportunity to explore the use of various data storage structures and distributed computing to enable interactive queries and visualization.

2. Names Profiler System

As georeferenced data has become increasingly available, more and more geographic visualization tools

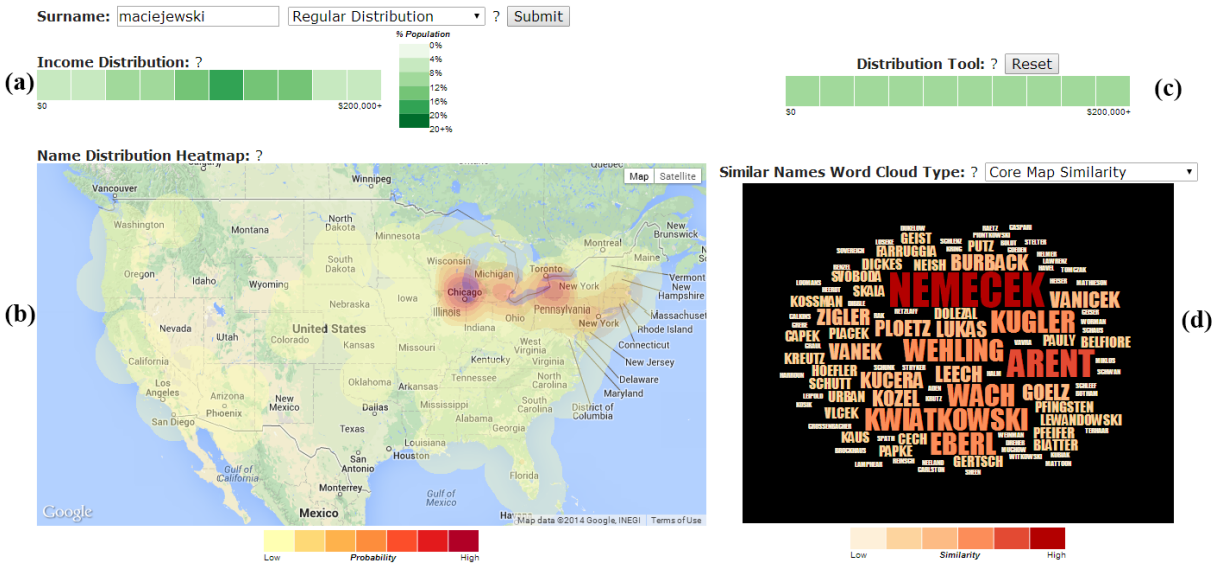


Figure 1: *The visual analytics interface to the United States name profiler. (a) A histogram encoded by color denoting the percentage of a given surname that is likely to map to an income range. (b) The spatial distribution of a surname. Users may look at a magnitude or probability distribution. (c) An income similarity toolbar. Users may search for names that are similar to a user defined income distribution. (d) The similarity wordle. The user may explore other surnames that have a similar spatial distribution or income distribution. Users can select a different similarity metric by changing the selected item in the dropdown.*

have been developed across a variety of domains (e.g., maritime analysis [MMME11, WvdWvW09], crime [MMCE10], healthcare [MBHP98, MHR*11], twitter analysis [MJR*11], movements [AAH*11] and various others [Wea09, GCML06, vLBA*12]). This work takes cues from Wood et al. [WDSC07] in developing a mashup for exploring surname distributions. We utilize publicly accessible telephone data that includes the geographic location of about 78 million people in the United States and link this data to the United States Census data. The goal of this work is to enable both novices and experts to explore name distributions and spatial relationships. We focus on three issues: aggregation, similarity and speed.

2.1. Density Estimation and Aggregation

This system estimates the probability density function of surnames to produce heatmap visualizations (Figure 1 (b)). We employ a fixed bandwidth kernel density estimation [Sil86] similar to other recent work [MRH*10, SWvdW*11]. Equation 1 defines the multivariate kernel density estimation.

$$\hat{f}_{\mathbf{h}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j=1}^{d}\frac{1}{\mathbf{h}_j}K\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\mathbf{h}_j}\right)\right). \quad (1)$$

Here, $\mathbf{h}$ represents the multi-dimensional smoothing parameter, N is the total number of samples, d is the data dimensionality, and K is a kernel function. In our system, we used the Epanechnikov kernel:

$$K(u) = \frac{2}{\pi}(1-u^2)1_{\{u\leq 1\}}, \quad (2)$$

where $1_{\{u\leq 1\}}$ evaluates to 1 if the inequality is true and 0 for all other cases.

We provide views for visualizing both the magnitude (count of a surname in a given region) and probability distribution of the data (count of a surname in a given region divided by the population estimate of that region). For names with less than 100 records in the database, no aggregation was made to ensure data privacy.

2.2. Linking With Secondary Data Sources

In order to link surnames to income, we utilize the household income in the 2008-2012 American Community Survey 5-Year Estimates [U.S13]. Each surname's address can be mapped to a given census tract. We then solve a system of linear equations to estimate the probability distribution associated with a given surname. For surnames with over 1000 records, we use three matrices to represent the distribution of name records and income histograms. In matrix D, D_{ij} is the number of surname records for the ith census tract and the jth surname. B contains the income histograms of the census tracts. Specifically, each census tract reports the per-

Figure 2: *Heatmap comparisons for surname Alvarado. Subfigure A represents the L^2-norm comparison and Subfigure B represents the core comparison. The left most images are heatmaps of the population distribution of Alvarado. The wordle displays the most spatially similar names to Alvarado with the larger and darker names being the most similar. The right most images show heatmaps of the similar names to Alvarado based on the comparison type.*

centage of the population that falls within one of ten given income ranges. B_{ik} is the percentage of the population within a given income range in the kth income bin in the ith census tract. The linear system is then defined as:

$$DX = B \tag{3}$$

Since D is not a full rank matrix, we used a non-negative least square solver [LH95] to obtain a solution. For surnames with less than 1000 records, we take a weighted average of the income distributions of all the census tracts a given surname falls within. Finally, the income distribution of a surname is mapped as a 1D histogram, where color represents the % of the surname that is likely to fall within that income range (Figure 1 (a)).

2.3. Similarity Exploration

The third component of our system consists of a wordle that is encoded to show similarity between names with respect to either spatial distribution or income (Figure 1 (d)). For the spatial similarity [Coe07, AFC10], we explored two distance metrics: the L^2-norm (Euclidean distance) and the core distance. In order to allow for interactive rates of similarity matching, we first precomputed the density estimates at a fixed zoom level and resolution (170×90). The distance between two names is then calculated as the L^2-norm between the 2D density estimate array.

While straight-forward to implement, the single-core CPU implementation on a computer with a 3.4GHz Core i7-2600 needs 40 minutes to calculate the pairwise similarity for a single surname (there are 1.4M unique surnames in the dataset). While all similarities can be precomputed, our goal was to also explore other potential designs. Previous work by Cheshire and Longley [CL12] looked at what they called the core distance between density distributions. This distance was related to the distance between the centroids of regions between two distributions that cover approximately 55% of the data. We extract the five largest local maxima from each density estimate as our cores, and then compute the similarity as the smallest pairwise distance between the cores of each surname. In this manner, all core distances can be fetched and fit into local memory and pairwise correlations can be calculated. We need no more than 3.5ms to compute the distance of a pair of names. The time to compare one name with all the other names in the database is reduced from 40 minutes to 30-50 seconds. The top five maxima were chosen based on performance and

Figure 2 compares the results of using the L^2-norm and the core distance metric. For the surname Alvarado, Marquez is the most similar heatmap using the L^2-norm comparison and Herrera is the most similar heatmap using the core comparison. The wordle can also be mapped to income similarity which is calculated as the L^2-norm between all sets of surnames in the dataset. The smaller the L^2-norm the more similar the income distribution. The wordle in Figure 3 shows the most similar surnames to Wang with respect to the income distribution, where the largest and darkest colored

Figure 3: *Income comparison for surname Wang with the most similar surname, Loh, presented. The larger and darker colored names are most similar to Wang.*

Figure 4: *A user defined income distribution looking for names that are predominately wealthy. The larger and darker colored names are most similar to the defined income.*

names representing the most similar surnames. Users may also define an income distribution using the tool shown in Figure 1 (d). The wordle in Figure 4 shows the most similar surnames with respect to the user defined income distribtuion.

3. Experiments

Finally, our main research interest was in enabling interactive exploration of this modestly large dataset in a web environment where both data aggregation and similarity searches are a priority. Previous work on BigData infoVis has focused primarily on enabling data aggregation techniques as they form the basis for creating interactive maps, scatterplots and parallel coordinate plots. For example, Liu et al. [LJH13] addressed interactive scalability of big data systems through data reduction methods such as brushing and linking. Lins et al. presented Nanocubes [LKS13] as a method for efficient storage and querying of large datasets. However, the current nanocubes implementation supports only single spatial dimensions and some datasets use large amounts of memory. Both works primarily focused on the use of data cubes as a means of modeling and viewing data in multiple dimensions.

While data cubes have been shown to be extremely effective for enabling information visualization, it is important to note that the data in a data cube has already been processed and aggregated. Their primary functions lie in summarization of trends and operational reports. In our case where we want to enable similarity searches, and such calculations are not well supported within a data cube. For our current implementation, we primarily focused on preprocessing the data. Map aggregates were saved as images to reduce the data overhead, and pairwise similarity comparisons were generated and surnames were linked to their 100 topmost similar surnames. We use a single-core CPU implementation with a 3.4GHz Core i7-2600. Our program uses approximately 2GB of memory for the 73283 census data records and 78 million surnames in the database. The database takes about 14 GB of space in a MySQL database. The precalculated similarities can be returned within 30 ms and took 14 days to precalculate the similarities.

4. Conclusions

Surnames in our system tend to follow expected ethnic distributions, discounting names with a large populations, such as Smith. Figure 3 hints to potential ethnic patterns within surnames of similar origins. Wang is an Asian surname and the most similar name to Wang (Loh) is also of Asian origin. Similar patterns occur within the spatial distributions (Figure 2) and the income distribution tool (Figure 4).

While the visualizations presented in this work are standard, the implementation of a web-enabled system for large scale visual analytics is still challenging. Our design of precomputing similarities for a large number of categories is effective only under the case of static data. What this shows is the need for using high-performance computing as a method of quickly processing analytical queries. In this way we can move from putting the burden of finding similar data items on the user to placing this burden on the computational side. With regards to the name profiler system, anecdotal evidence suggests that the data matches users' mental models, and system users typically engage in exploration for 10 minutes or more. The current implementation can be tested at: http://goo.gl/gOGEVJ. A video demonstration can be viewed at: http://youtu.be/pANl4YJ1C5I.

5. Acknowledgments

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and by the Engineering and Physical Sciences Research Council UK EPSRC grant EP/J005266/1.

References

[AAH*11] ANDRIENKO G. L., ANDRIENKO N. V., HURTER C., RINZIVILLO S., WROBEL S.: From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[AFC10] ANSARI M. H., FILLMORE N., COEN M. H.: Incorporating spatial similarity into ensemble clustering. In *MultiClust KDD* (2010). 3

[CL12] CHESHIRE J. A., LONGLEY P. A.: Identifying spatial concentrations of surnames. *International Journal of Geographics Information Science 26* (2012), 309–325. 1, 3

[CLS10] CHESHIRE J. A., LONGLEY P. A., SINGLETON A. D.: The surname regions of great britain. *Journal of Maps 6*, 1 (2010), 401–409. 1

[CLYN13] CHESHIRE J. A., LONGLEY P. A., YANO K., NAKAYA T.: Japanese surname regions. *Papers in Regional Science 92* (2013), In Press. 1

[Coe07] COEN M. H.: *A Similarity Metric for Spatial Probability Distributions*. Tech. rep., CSAIL MIT, 2007. 3

[GCML06] GUO D., CHEN J., MACEACHREN A. M., LIAO K.: A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1461–1474. 2

[LCM11] LONGLEY P. A., CHESHIRE J. A., MATEOS P.: Creating a regional geography of britain through the spatial analysis of surnames. *Geoforum 42* (2011), 506–516. 1

[LH95] LAWSON C. L., HANSON R. J.: *Solving Least Squares Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1995. 3

[LJH13] LIU Z., JIANG B., HEER J.: imMens: Real-time visual querying of big data. *Comput. Graph. Forum 32*, 3 (2013), 421–430. 4

[LKS13] LINS L., KLOSOWSKI J., SCHEIDEGGER C.: Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2456–2465. 4

[MBHP98] MACEACHREN A. M., BOSCOE F. P., HAUG D., PICKLE L.: Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Proceedings of the IEEE Symposium on Information Visualization* (1998). 2

[MHR*11] MACIEJEWSKI R., HAFEN R., RUDOLPH S., LAREW S., MITCHELL M., CLEVELAND W., EBERT D.: Forecasting hotspots - a predictive analytics approach. *IEEE Transactions on Visualization and Computer Graphics 17*, 4 (2011), 440–453. 2

[MJR*11] MACEACHREN A. M., JAISWAL A., ROBINSON A. C., PEZANOWSKI S., SAVELYEV A., MITRA P., ZHANG X., BLANFORD J.: Senseplace2: Geotwitter analytics support for situational awareness. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[MLO11] MATEOS P., LONGLEY P. A., O'SULLIVAN D.: Ethnicity and population structure in personal naming networks. *PLoS ONE (Public Library of Science) 6*, 9 (2011), 1–12. 1

[MMCE10] MALIK A., MACIEJEWSKI R., COLLINS T. F., EBERT D. S.: Visual analytics law enforcement toolkit. In *Proceedings of the IEEE Conference on Technologies for Homeland Security* (2010). 2

[MMME11] MALIK A., MACIEJEWSKI R., MAULE B., EBERT D. S.: A visual analytics process for maritime resource allocation and risk assessment. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology* (2011). 2

[MRH*10] MACIEJEWSKI R., RUDOLPH S., HAFEN R., ABUSALAH A. M., YAKOUT M., OUZZANI M., CLEVELAND W. S., GRANNIS S. J., EBERT D. S.: A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics 16*, 2 (2010), 205–220. 2

[Sil86] SILVERMAN B. W.: *Density Estimation for Statistica and Data Analysis*. Chapman & Hall/CRC, 1986. 2

[SWvdW*11] SCHEEPENS R., WILLEMS N., VAN DE WETERING H., ANDRIENKO G., ANDRIENKO N., VAN WIJK J. J.: Composite density maps for multivariate trajectories. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2518–2527. 2

[U.S13] U.S. CENSUS BUREAU: 2008-2012 American Community Survey 5-Year Estimates, 2013. 1, 2

[vLBA*12] VON LANDESBERGER T., BREMM S., ANDRIENKO N., ANDRIENKO G., TEKUSOVA M.: Visual analytics methods for categoric spatio-temporal data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct 2012), pp. 183–192. 2

[WDSC07] WOOD J., DYKES J., SLINGSBY A., CLARKE K.: Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1176–1183. 2

[Wea09] WEAVER C.: Cross-filtered views for multidimensional visual analysis. *IEEE Transactions on Visualization and Computer Graphics 16* (2009). 2

[WvdWvW09] WILLEMS N., VAN DE WETERING H., VAN WIJK J. J.: Visualization of vessel movements. *Computer Graphics Forum* (2009), 959–966. 2

A Bipartite-Graph Based Approach for Disaster Susceptibility Comparisons among Cities

Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, Ning Xie, Jinpeng Wei
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, U.S.A.
Email: {wzhou005,cshen001,taoli,chens,nxie,weijp}@cs.fiu.edu

***Abstract*—People are attracted to large cities because of more employment opportunities, convenient facilities, and rich cultural activities. However, large cities are also more vulnerable to natural disasters, which have caused widespread physical destructions, great loss of life and property, and immense havoc. "Which city is less susceptible to natural disasters?" is thus one of the most critical questions one faces when making decisions on travelling or job and business relocation. In this work, we propose a bipartite-graph based framework to compare the impacts of disasters on two cities by answering different queries using textual documents collected online. Besides intuitive simple comparison using statistics, our system also generates textual comparative summaries to better describe the differences between the two cities in terms of safety. Although a number of online services provide disaster events statistic information for cities, our framework compares the impacts of disasters on cities in a more straightforward and comprehensive way.**

Keywords: Disaster Susceptibility Comparison, Disaster-Impact Bipartite Graph, Comparative Summarization

I. Introduction

People are attracted to metropolitan areas due to more employment opportunities, convenient facilities, and rich cultural activies. However, large cities are also vulnerable to natural disasters, which tend to cause more damage in densely populated areas. For example, about 80% of New Orleans was flooded in Hurricane Katrina 2005; New York City was seriously affected by Hurricane Irene and Hurricane Sandy in 2011 and 2012; the winter storm 2011 left 21 inches of snow in Chicago; lots of earthquakes have happened in the two major cities on the west coast of U.S., Los Angeles and San Francisco; and frequent hurricane hits in Miami area. Therefore, before making decisions on traveling or job and business relocation, one of the most critical questions people face is: which city is safer?

For city safety comparison, a number of online services[1] provide statistic data about various aspects of cities or neighborhoods such as crime rates, races, living expenses and house prices. However, to the best of our knowledge, none of them considers the impacts of natural disasters. On the other hand, although current and historical disaster data can be easily obtained (e.g., through National Hurricane Center[2] for hurricanes and U.S. geological survey[3] for earthquakes), information about how a disaster event affects a specific city is not readily available. In most cases, data on impacts of disasters on cities is stored and archived by different government agencies or organizations. Extra efforts are often required to collect data or/and perform data integration into a unified database to support comparisons among different cities. Moreover, although statistics about damages and fatalities can provide direct evidences for the safety comparison, it is still quite challenging to obtain an overview on historically how severe a city was affected by disasters, since many types of impacts from disasters – e.g., road closure caused by a hurricane – are not reflected by the statistics.

In this paper, we tackle this problem by aggregating easily acquired textual documents available online and providing comprehensive descriptions of different impacts under natural diasters of a city. Instead of answering the question "which city is safer?" directly, we provide straightforward and descriptive information about a pair of cities for the following four types of queries to help users make their own decisions:

- What are the major impacts caused by a specific type of disasters for the two cities? For example, hurricanes in Miami are more likely to cause house damage, but more likely to cause rainfall and landslide in Los Angeles.
- What are the major types of disasters leading to a specific type of impact for the two cities? For example, "house damage" is mainly caused by hurricanes in Miami, but by earthquakes in Los Angeles.
- What are the most likely or frequent disasters affecting the two cities? For example, hurricanes occur more frequently in Miami, and earthquakes in Los Angeles.
- What are the overall impacts caused by disasters for the two cities? For example, in Miami, there is more flooding and house damage, and in Philadelphia it is more likely to have rainfall and death.

To answer these queries, we propose an interactive *weighted bipartite graph* to model the disaster impacts on cities. There are two types of nodes, **disaster** nodes and **impact** nodes, in the bipartite graph. **Disaster** nodes represent hazards to the city safety which can cause significant damages and destructions. The hazards can be decided by domain experts

[1]Examples include: http://www.neighborhoodscout.com, http://www.numbeo.com/, and http://www.city-data.com
[2]http://www.nhc.noaa.gov/data/
[3]http://earthquake.usgs.gov/earthquakes/map/

or using an ontology of disaster management. **Impact** nodes represent consequences caused by **disaster** nodes, and they are extracted from plain texts via a topic modeling approach [1]. A weighted edge from a **disaster** node to an **impact** node denotes that the source node is responsible for the target node and its weight specifies to what extent the responsibility is. Triggered by users' queries, various comparative summaries will be generated from the filtered text to provide detailed textual descriptions of the differences between the two cities. A demonstration system can be visited at http://bigdata-node01.cs.fiu.edu/CitySafetyComparison/.

In summary, our main contributions are listed below:

- We present a weighted bipartite graph based framework to model the problem of comparing city disaster susceptibilities, in which the casual relationship between different types of disasters and their impacts on a city is encoded in weighted edges;
- We apply topic modeling to extract topics from documents to represent different types of disaster impacts;
- We design a prototype system which provides textual summaries about two chosen cities for various comparative queries;
- We conduct a case study using Wikipedia documents on 4 different U.S. cities with 6 pairwise comparisons to demonstrate the effectiveness of our proposed framework.

The rest of the paper is organized as follows. After discussing related work in Section II, we first give a brief overview of our framework in Section III. Detailed descriptions of how to construct the bipartite graph and how to conduct city safety comparisons based on the bipartite graph are presented in Section IV and Section V, respectively. We present our case study results in Section VI and finally conclude with discussions and outlines for future extensions in Section VII.

II. Related Work

City safety study has attracted much attention recently in computer science. Classical prediction methods such as ARIMA models and artificial neural networks [2], [3], [4], [5] have been successfully applied in crime-related prediction, like drug market or other specially designed safety indices. Another direction is how to build up sensor networks that can quickly respond in an emergency event like fires and traffic accidents [6], [7], [8]. While most existing studies focus on the safety of an individual city, our work provides a comparative view between different cities in terms of their safety.

Many information systems and techniques have been proposed in disasters monitoring, relief and recovery. Commercial systems such as Web EOC and E-Team are usually used by Emergency Management departments located in urban areas [9], [10]. Recently many disaster situation-specific tools provide query interfaces, GIS and visualization capabilities to support user interactions and queries to improve situation awareness [11] in a specific disaster event.For example, Ushahidi [12] provides a platform with visualization and interactive maps to crowd source news stories and crisis information using multiple channels and GeoVISTA [13] monitors tweets to form situation alerts according to the geo-locations associated with the tweets. However, these tools do not answer the comparative queries about all disaster related data of different cities.

Multi-document summarization has been used to provide concise summaries about large document collections and many different approaches have been developed including centroid-based [14], graph-based [15], [16], clustering-based [17], [18], knowledge-based [19], [20], etc. Comparative summarization, as a special class of summarization tasks, helps people understand what are the connections and differences between two document collections and has been studied with different applications. Kim and Zhai [21] compare positive reviews and negative reviews for one product by extracting the most related and representative sentence pairs for the two review sets, while Huang et al. [22] compare related news topics by extracting sentences covering the most important related or representative concepts. Wang et al. [23] model the comparative summary as a sentence set including the most discriminative sentences from different document sets. Wan et al. [24] conduct comparative summarization on news from different regions (in different languages) on the same topic using random walk methods on a sentence graph. Instead of directly extracting sentences from different document sets, this work utilizes the weighted bipartite graph to model impacts of disasters and filter documents for comparative summarization.

Graph-based approaches have also been used to generate event storylines that describe how an event evolves over time. Wang et al. [25] developed a multi-view graph based framework for integrating text, image, and temporal information to generate storylines to reflect the evolution of the given topic. Wu et al. [26] proposed a two-layer storyline generation framework which provides global storylines for cross-location disaster events on the first layer and location-specific storylines for individual events on the second layer. Shahaf et al. [27] developed *metro map* for creating structural summaries of documents by optimizing several objectives (e.g., relevance, coherence, coverage and connectivity) simultaneously. Unlike existing studies, in this work, we utilize a weighted bipartite graph based framework to perform city safety comparison.

III. The Framework Overview and Notations

To capture the relationship between disasters and their impacts on a city, we propose a weighted bipartite graph based framework.

Definition 1: A *weighted bipartite graph* is a graph $G = (U, V, E, w)$ whose vertices can be divided into two disjoint set U and V such that every edge connects a vertex in U to a vertex in V, i.e. $E \subseteq U \times V$, and $w : E \to \mathcal{R}^+$ is a weight function which assigns a non-negative weight to each edge $e \in E$.

In our framework, U is the set of **disaster nodes**, V is the set of **impact nodes**, and every edge is associated with a triple (c, S, w), where c is the label of a city, S is a sentence set related to the edge, and w is the weight of the edge.

Definition 2: Disaster nodes are the (left) vertices in the bipartite graph that represent city hazards, such as hurricane, storm and tornado.

Definition 3: Impact nodes are the (right) vertices in the bipartite graph that represent consequences caused by the **disaster** nodes, such as death, house damage and economic loss.

Definition 4: An *impact topic of disasters* is a bag-of-words which are commonly used to describe a type of impacts of disasters. For example, *death, died, killed, fatalities, injuries* are commonly used words to describe the impact "human life loss" caused by disasters.

Figure 1 shows our framework architecture and Table I summarizes the notation used in this paper. The input of our framework is several sets of sentences, $S^c, c \in \{c_1, c_2, ..., c_n\}$, and the sentence set S^c for city c is collected from online disaster-related documents (e.g., Wikipedia pages of disaster events in our case study in Section VI). Every sentence $s \in S^c$ depicts some aspect of the city c in a disaster event.

The following is a sentence instance about *Chicago*:

Only two people died in the fire but 10,000 were made homeless and 1,800 buildings were burned to the ground.

In the above sentence, *fire* is a disaster type and its impacts include *death, homeless, building burned*.

To process the sentences, words describing disaster damages are extracted from sentences and grouped into *impact topics* in our framework. Then for each impact topic a, we assign a probability $p(a|s)$ for each sentence s (the details will be described in Section IV-A), indicating the weight of impact topic a discussed in the sentence. For instance, in the above example, "homeless, building, burned" will be assigned higher weights than "died" for the disaster fire in Chicago.

The vertex set of the bipartite graph includes *disaster nodes* and *impact nodes*, representing disasters and impact topics, respectively. Edges between *disaster nodes* and *impact nodes* indicate the causal relationship between them and the weight on an edge specifies the strength of the relationship. The bipartite graph encodes all the information about the queries mentioned in Section I for city safety comparison. Users can interact with this bipartite graph and submit a comparative query by clicking a node. The default query without clicking any nodes is: *what are the overall differences between city c_1 and c_2?* By clicking a disaster node d_i, the query becomes: *what are the differences between city c_1 and c_2 on disaster d_i?* By further clicking a impact node a_j, the query becomes: *what are the differences between city c_1 and c_2 on impact a_j caused by disaster d_i?*

IV. Bipartite Graph Construction

The weighted bipartite graph is constructed as follows. First, we pre-define some disaster types like *hurricane, tornado, storm and earthquake*. We then apply a domain ontology of disaster management [19], [20] to extract sentences from the input sentence sets which contains concepts belonging to those disasters. For instance, sentences containing the phrase "tropical cyclone" are extracted as sentences about "hurricane", since "tropical cyclone" is considered as a sub-concept of "hurricane".

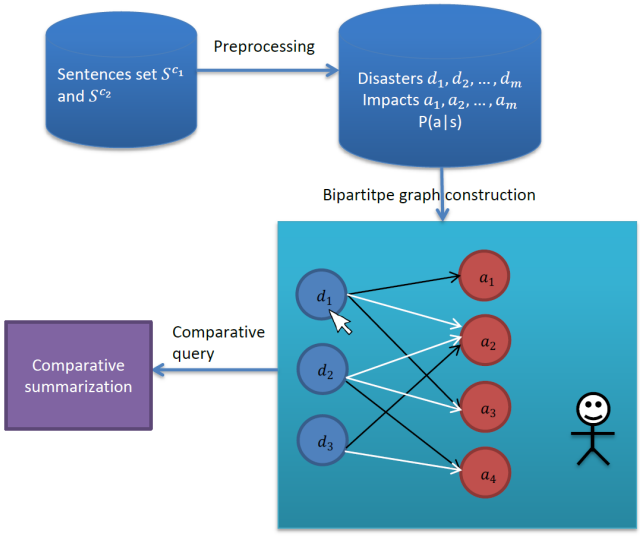


Fig. 1. An Overview of The System Framework.

TABLE I. A Summary of Notation.

c	city
d	**disaster** node j
a	**impact** node k
S^c	sentences set of city c
$\hat{S}^c$	sentences set of city c after removing **impact** unrelated words
S^c_i	sentences subset of city c filtered out on **disaster** node i
$S^c_{i,j}$	sentences subset of city c filtered out on **disaster** node i and **impact** node j
$p(a\|s)$	probability of an impact topic a in sentence s, which comes from output of LDA. Here document-topic distribution in LDA model represents sentence-impact distribution.
$R_n(s)$	the most likely top n impact topics according to sentence-impact distribution of s
$e^c_{i,j}$	edge between **disaster** node i and **impact** node j on city c
$w^c_{i,j}$	weight of edge $e^c_{i,j}$

A. Impact Node Extraction

According to Definition 3 and Definition 4, impact nodes encode negative consequences caused by disasters and are associated with a representative bag-of-words. However, unlike disaster nodes, it is difficult to enumerate or predefine all possible impact types and it is even more difficult to associate predefined impact types to the actual textual descriptions in given documents. To overcome this difficulty, we extract impacts directly from texts using information extraction and text mining techniques. Consider the ideal case in which the input sentence set is about disaster impacts on cities, then each sentence is a textual description of a tuple $(disaster, where, when, impact)$. Therefore, if each impact node represents an impact topic, we need to identify different impacts that have less overlap with each other. Based on this intuition, we use a topic modeling tool, *latent Dirichlet allocation* (LDA) [1] to cluster words about impacts into several groups, where each group corresponds to an impact topic. To exclude other unnecessary words in the sentences, we preprocess the original sentence set S^c as follows: (1) remove words related to **disaster** nodes; (2) remove words explaining

when, where, who using entity recognition techniques [28]; and (3) remove the stop words.

After the preprocessing, we obtain a sentences set $\hat{S}^c$ for every city c. To compare two cities c_1 and c_2, we apply LDA on the preprocessed sentence set $\hat{S}^{c_1} \cup \hat{S}^{c_2}$ together with the impact number k, which specifies the number of **impact** nodes. The LDA approach will generate k topic with words distribution respectively, as well as a conditional probability $p(a|s)$ for every impact topics a on a given sentence s, which is then used to calculate the weights of edges between **disaster** nodes and **impact** nodes.

B. Weight Calculation for Disaster-Impact Edges

We calculate the weight of an edge based on the sentence set of the city related to the disaster node and the impact node.

Let S_i^c be the set of sentences related to disaster i in S^c, which is extracted using a disaster ontology as

$$S_i^c = \{s \in S^c \mid s \text{ contains } d_i \text{ or a sub-type of } d_i\}. \quad (1)$$

Let $S_{i,j}^c \subset S_i^c$ be the sentence set about city c, **disaster** node i and **impact** node j, which, roughly speaking, is the set of sentences containing impact topic j:

$$S_{i,j}^c = \{s \in S_i^c \mid p(a_j|s) > \epsilon\}, \quad (2)$$

where ϵ is a threshold parameter. However, we find it is difficult in practice to choose a proper parameter ϵ value, as it is very sensitive to the input data set. A small ϵ will lead to too many connections, while a large ϵ will rule out too many sentences and result in very sparse bipartite graphs. Instead, in our framework, for every sentence s in S_i^c, we only consider its top n most likely impact topics $R_n(s)$ (n is set to 2 in our case study), and use the following to define $S_{i,j}^c$ in place of Eq.(2):

$$S_{i,j}^c = \{s \in S_i^c \mid a_j \in R_n(s)\} \quad (3)$$

Finally, the weight of edge $e_{i,j}^c$, $w_{i,j}^c$, is defined as

$$w_{i,j}^c = \sum_{s \in S_{i,j}^c} p(a_j|s). \quad (4)$$

If $w_{i,j}^c$ is 0, then we remove the edge between d_i and a_j and assume there is no connection between the disaster and the impact.

V. City Comparison Based on The Bipartite Graph

Our framework provides city comparisons through two perspective views: simple comparison and textual comparison. Simple comparison through bipartite graph gives general and direct discrepancies so that users can quickly grasp the differences between two cities but it does not provide detailed textual description. Textual comparison remedies this by providing comparative summaries according to users' comparative queries.

A. Simple Comparison

Figure 2 shows a simple comparison result of two cities, Miami and Los Angeles. From the thickness of edges between disaster nodes and impact nodes (used to denotes the weights of edges) in the bipartite graph, one can observe that earthquakes occur more frequently in Los Angles, while in Miami hurricanes happen much more often.

More generally, the four types of queries of city safety comparisons described in Section I can now be addressed using the information stored in the bipartite graph (in particular, the edge weight $w_{i,j}^c$ indicating the causal strength between disaster d_i and a_j in city c) as follows:

- *What are the impact differences caused by the specific disaster d_i for city c_1 and c_2 ?* Such a query can be answered by comparing two weight vectors $w_{i,1}^{c_1}, \ldots, w_{i,k}^{c_1}$ and $w_{i,1}^{c_2}, \ldots, w_{i,k}^{c_2}$, which are visualized in the bipartite graph as the line thickness of highlighted edges with different colors.
- *What are the disasters differences leading to specific impact a_j for city c_1 and c_2 ?* Such a query can be answered by comparing two vectors $e_{1,j}^{c_1}, \ldots, e_{m,j}^{c_1}$ and $e_{1,j}^{c_2}, \ldots, e_{m,j}^{c_2}$, which are visualized in the bipartite graph as the line thickness of highlighted edges with different colors as well.
- *What are the overall disaster differences for city c_1 and c_2?* To answer such a query, for a city c and a disaster d_i, we aggregate weights of edges from d_i to all impact nodes as $w_{i*}^c = \sum_j w_{i,j}^c$. Then we can compare the two aggregated weight vectors $w_{1*}^{c_1}, \ldots, w_{n*}^{c_1}$ and $w_{1*}^{c_2}, \ldots, w_{n*}^{c_2}$, which are visualized as the length of bars along with the disaster nodes.
- *What are the overall impact differences caused by disasters for city c_1 and c_2 ?* To answer such a query, for a city c and impact a_j, we accumulate weights of edges originating from all disaster nodes to a_j as $w_{*j}^c = \sum_i w_{i,j}^c$. Then we can compare two weight vectors $w_{*1}^{c^1}, \cdots, w_{*m}^{c^1}$ and $w_{*1}^{c^2}, \cdots, w_{*m}^{c^2}$, which are visualized as the length of bars along with the impact nodes.

B. Textual Summarization for Comparative Queries

The bipartite graph provides simple comparisons using weights induced from topic modeling, but it lacks detailed textual descriptions, which can be remedied by textual comparative summarization. In this work, we apply the comparative summarization method in [23] on two sentence sets according to different comparative queries.

For two cities c_1 and c_2, our framework performs comparative summarization on two sentence sets S^{c_1} and S^{c_2}. Different comparative queries (resulted from user interactions via clicking bipartite graph nodes) will generate differentS^{c_1} and S^{c_2} for comparative summarization. For example, S^{c_1} and S^{c_2} are set to be $S_i^{c_1}$ and $S_i^{c_2}$ when a user clicks the disaster node d_i; they are set to be $S_{i,j}^{c_1}$ and $S_{i,j}^{c_2}$ when the user sequentially clicks the disaster node d_i and the impact node a_j.

TABLE III. MOST LIKELY DISASTER TYPES AND IMPACT TYPES FOR THE CITIES IN PAIRWISE COMPARISON.

City Pair	$\text{argmax}_d\, p_1(d)$	$\text{argmax}_d\, p_2(d)$	$\text{argmax}_e\, p_1(e)$	$\text{argmax}_e\, p_2(e)$
Miami Chicago	storm	storm	landfall,fatalities,weather	damaged,struck,collapse
Miami Los Angeles	storm	earthquake	depression,inches,rain	ground,killed,dropped
Miami Philadelphia	storm	storm	killed,flooded,streets	destroyed,accident,fatalities
Chicago Los Angeles	storm	earthquake	rain,temperatures,flood	flight,killed,billion
Chicago Philadelphia	storm	storm	fire,flight,alarm	death,rain,attack
Los Angeles Philadelphia	earthquake	storm	adventures,destroyed,discovery	fire,killed,weather

TABLE IV. MOST LIKELY IMPACTS CAUSED BY EACH DISASTER TYPE FOR CITIES OF MIAMI AND LOS ANGELES.

City Pair	hurricane	storm	tornado	earthquake
Miami Chicago	crash,bodies,dropped landfall,fatalities,weather	landfall,fatalities,weather flooding,homes,killed	damage,million,buildings warning,pressure,tides	$\emptyset$ depression,quickly,evaluated
Miami Los Angeles	depression,inches,rain occurred,large,reached	rainfall,flooded,houses landfall,pressure,struck	depression,inches,rain $\emptyset$	$\emptyset$ warnings,destroyed,moved
Miami Philadelphia	killed,flooded,streets peak,inches,power	landfall death warnings damage,rainfall,million	reported,pressure,force $\emptyset$	$\emptyset$ reported,pressure,force

TABLE II. THE SIZE OF CITY SENTENCE SET

city	# of sentences
Miami	772
Chicago	618
Los Angeles	607
Philadelphia	685

In [23], the comparative summarization is modeled as a discriminative sentence selection process based on a multivariate normal generative model to extract sentences best describing the unique characteristics of each document group.

Problem 1. Suppose we have f sentences of the document collection, denoted by $\{X_i \mid i \in F\}$, where F is an index set of sentences with $|F| = f$. We are also given the group variable, Y, which is represented by multiple group indicator variables. The problem of *sentence selection* is to select a subset of sentences, $S \subset F$, to accurately discriminate a group of documents from other groups, i.e. to predict the group identity variable Y, given that the cardinality of S is m ($m < f$). Let us denote $\{X_i \mid i \in S\}$ by X_s, for any set S. The prediction capability of Y given X_s can be measured by the entropy of Y given X_s, which is defined as

$$H(Y|X_s) \stackrel{def}{=} -E_{p(Y,X_s)} \log p(Y|X_s), \tag{5}$$

where $E_p(\cdot)$ is the expectation given the distribution p, and p stands for the underlying document distribution, i.e. the joint distribution $p(Y, X_s)$. The sentence selection problem using the mutual information criterion is

$$\underset{S}{\text{argmin}}\, H(Y|X_S). \tag{6}$$

Selecting an optimal subset of sentences known to be an NP-hard problem. A greedy approach is proposed in [23], which sequentially selects sentences to obtain a sub-optimal solution.

VI. THE CASE STUDY

To demonstrate the effectiveness of the proposed framework, a case study is conducted to compare city safety among four U.S. cities (Miami, Chicago, Los Angeles, and Philadelphia) using the impacts of four types of disasters – hurricane, storm, tornado, and earthquake.

A. Dataset Description

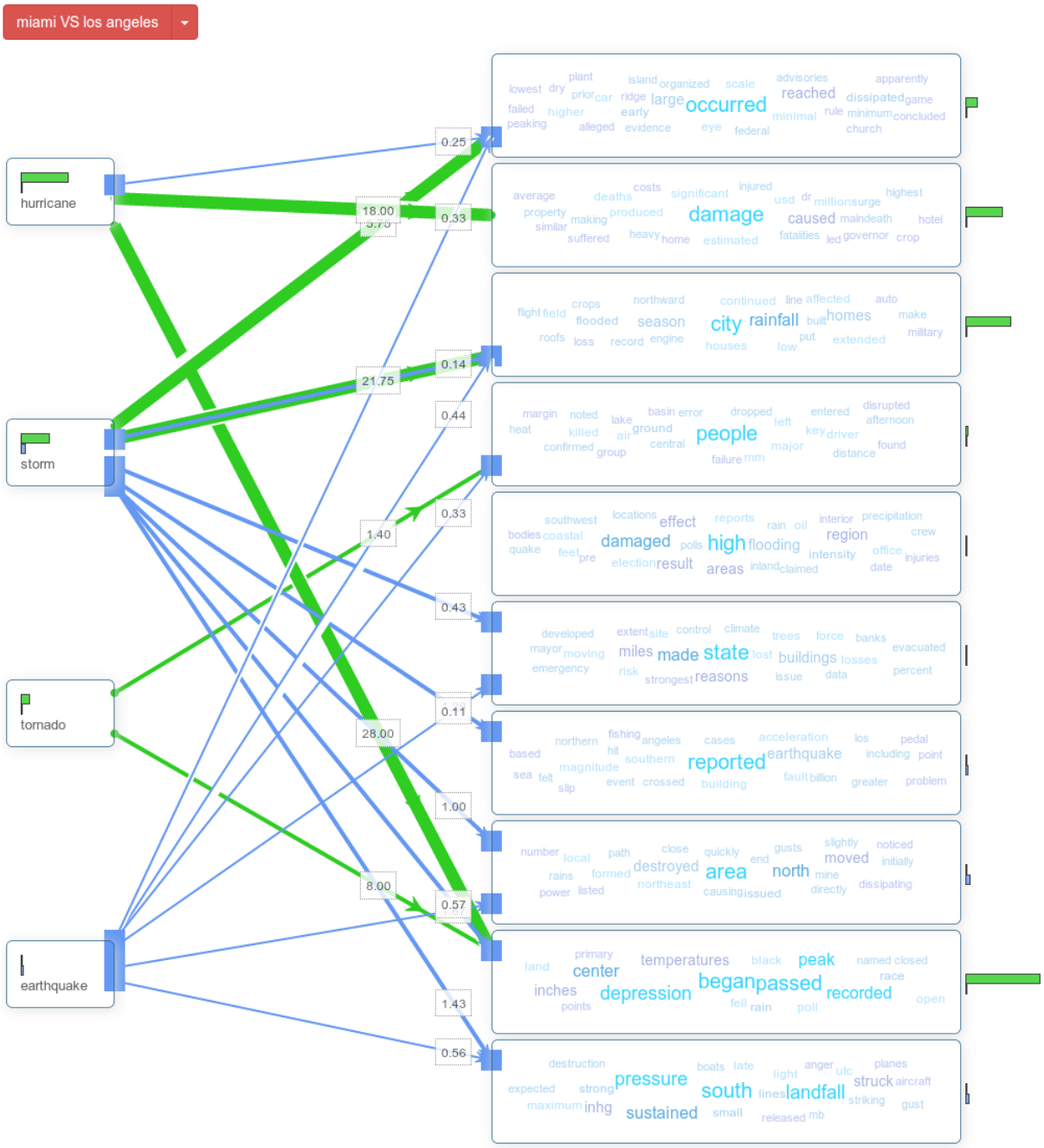
We collect the dataset from Wikipedia. For each city, we first extract all the paragraphs of Wikipedia page containing the city name, and further extract sentences containing phrases about one of the four disaster types. Table II shows the basic statistics of the dataset.

B. Results Analysis

Figure 2 demonstrates the comparative result of pairwise city comparison between Miami and Los Angeles, in which green components encode the information for city Miami and blue components encode the information for city Los Angeles. Furthermore, Table III shows the general differences in pairwise city comparison. The third column in Table III lists the most likely disaster types, and the fourth column in Table III lists the most likely effects/impacts. For each entry, 3 representative words are manually selected among the 15 top-ranked words, according to the word probability in the corresponding impact topic generated from LDA. Similar to Section V-A, we can answer the following queries in Section I from the case study results.

What are the overall disaster differences for city Miami and Los Angeles? From Table III, one can see that the most likely disaster for Miami is storm, and the most likely disaster for Los Angels is earthquake. This reflects the real difference between these two cities, since Miami is a city located on the Atlantic coast in south-eastern Florida which has a tropical monsoon climate and Los Angeles is subject to earthquakes due to its location on the Pacific Ring of Fire. In addition, from Figure 2, one can observe that tornadoes barely happen in Los Angeles.

What are the overall impact differences caused by disasters for city Miami and Los Angeles? Table III shows that the most likely impact types for city Miami are *depression, inches, rain*, which is regarded as rainfall, but for city Los Angeles they are *ground,kill,dropped*, which can be interpreted as life loss and house collapse. This observation can be easily explained since frequently occurred storms in Miami cause plentiful rainfall while earthquakes in Los Angeles cause life loss and house collapse. Here, we only illustrate results of pairwise



cityOne

At the time, the storm was nearly stationary. The storm caused significant property and crop damage along the . A rare hurricane also affected the — area. The death toll in was low because of well executed warnings and advisories. The cyclone turned south, under the influence of northerly winds from a high pressure system. It likely developed from a tropical wave several days before. During the storm, up to of rain in three hours were reported to have fallen on the city of Fort Lauderdale, and sections of Broward County were under of water. Minimal erosion occurred in some locales. By , made landfall near border and dissipated over on .

cityTwo

Farther north in Santa Clara County the flow of well water was affected. was buffeted by high winds, damaging corn crops and trees. The main shock epicenter occurred offshore about from the city, near . and reported light damage and death was reported in city. Over three hours, one thunderstorm dropped nearly of rain on Indio. During this time, the middle of the nation was being affected by a severe ice storm system caused by the tail end of the same arctic air afflicting the west. On , a broad area of low pressure developed within a tropical wave several miles south of , . had also made significant contributions to understanding subsidence in oilfields.

Fig. 2. Bipartite graph of city pair Miami and Los Angeles

city comparison between Miami and Los Angeles; results from other five pairwise city comparisons are also listed in Table III.

What are the impact differences caused by the specific disaster d_i for city Miami and Los Angeles? Table IV highlights the comparison between Miami and Los Angeles for the most likely effects/impacts given a disaster type, which provides answers for this type of query. The most likely impact types caused by hurricane in Miami are *depression, inches, rain*, but the impact types for Los Angeles are *occurred, large, reached*. Besides, storms in Miami most likely cause *rainfall, flooded, houses*, but in Los Angeles they mainly cause much more peaceful type of impacts *landfall, pressure, struck*. These differences can be explained by that Miami is more geographically flat but Los Angeles is more mountainous, which obstructs further evolution of strong rainfall. For the other two disasters, tornadoes only occur in Miami and mainly lead to impacts *depression, inches, rain*, meanwhile earthquakes only occur in Los Angeles and mainly lead to impacts *warning, destroyed, moved*.

VII. Conclusion

In this paper, we study the problem of comparing cities' disaster susceptibilities and propose a weighted bipartite graph based framework. Using our framework, direct city comparison can be performed on the bipartite graph and additional textual comparative summaries for different queries can be generated through user interactions via clicking the bipartite graph nodes.

For the future work, we plan to extend our framework in the following aspects: (1) We will improve the impact node extraction to extract more accurate impact topics; (2) We will include more safety issues like crime and man-made disasters; (3) We will employ more efficient graph algorithms (e.g., random walk) to utilize the bipartite graph structure in our framework.

Acknowledgment

The work was supported in part by the National Science Foundation under grants HRD-0833093, CNS-1126619, and IIS-1213026, the U.S. Department of Homeland Security under grant Award Number 2010-ST-062000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and Army Research Ofce under grant number W911NF-1010366 and W911NF-12-1-0431.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] P. Chen, H. Yuan, and X. Shu, "Forecasting crime using the arima model," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, vol. 5. IEEE, 2008, pp. 627–630.

[3] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[4] J. Ballesteros, M. Rahman, B. Carbunar, and N. Rishe, "Safe cities. a participatory sensing approach," in *Local Computer Networks (LCN), 2012 IEEE 37th Conference on*. IEEE, 2012, pp. 626–634.

[5] A. M. Olligschlaeger, "Artificial neural networks and crime mapping," *Crime mapping and crime prevention*, pp. 313–348, 1997.

[6] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges," *Computer*, vol. 44, no. 6, pp. 32–39, 2011.

[7] M. Karpiriski, A. Senart, and V. Cahill, "Sensor networks for smart roads," in *Pervasive Computing and Communications Workshops, 2006. PerCom Workshops 2006. Fourth Annual IEEE International Conference on*. IEEE, 2006, pp. 5–pp.

[8] H.-S. Jung, C.-S. Jeong, Y.-W. Lee, and P.-D. Hong, "An intelligent ubiquitous middleware for u-city: Smartum." *Journal of Information Science & Engineering*, vol. 25, no. 2, 2009.

[9] E. A. Inc, "Webeoc," http://www.esi911.com/home.

[10] NC4, "E-teams," http://www.nc4.us/ETeam.php.

[11] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, and S.-C. Chen, "Data mining meets the needs of disaster information management," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 5, pp. 451–464, 2013.

[12] Ushahidi, "http://www.ushahidi.com/," 2012.

[13] GeoVISTA, http://www.geovista.psu.edu.

[14] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," in *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*. Association for Computational Linguistics, 2000, pp. 21–30.

[15] G. Erkan and D. R. Radev, "Lexpagerank: Prestige in multi-document text summarization." in *EMNLP*, vol. 4, 2004, pp. 365–371.

[16] C. Shen and T. Li, "Multi-document summarization via the minimum dominating set," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 984–992.

[17] C. Shen, T. Li, and C. H. Ding, "Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases." in *AAAI*, 2011.

[18] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 307–314.

[19] L. Li and T. Li, "An empirical study of ontology-based multi-document summarization in disaster management," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 2, 2014.

[20] L. Li, D. Wang, C. Shen, and T. Li, "Ontology-enriched multi-document summarization in disaster management," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 819–820.

[21] H. Kim and C. Zhai, "Generating comparative summaries of contradictory opinions in text," in *Proceeding of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 385–394.

[22] X. Huang, X. Wan, and J. Xiao, "Comparative news summarization using linear programming," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 2011, pp. 648–653.

[23] D. Wang, S. Zhu, T. Li, and Y. Gong, "Comparative document summarization via discriminative sentence selection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 3, p. 12, 2012.

[24] X. Wan, H. Jia, S. Huang, and J. Xiao, "Summarizing the differences in multilingual news," in *Proceedings of the 34th International ACM SIGIR conference on Research and development in Information*. ACM, 2011, pp. 735–744.

[25] D. Wang, T. Li, and M. Ogihara, "Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs." in *AAAI*, 2012.

[26] W. Zhou, C. Shen, T. Li, S. Chen, N. Xie, and J. Wei, "Generating textual storyline to improve situation awareness in disaster management," in *In Proceedings of the 15th IEEE International Conference on Information Reuse and Integration (IRI 2014)*, 2014.

[27] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 899–908.

[28] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

Building Multi-model Collaboration in Detecting Multimedia Semantic Concepts (Invited Paper)

Hsin-Yu Ha, Fausto C. Fleites, Shu-Ching Chen
School of Computing and Information Sciences,
Florida International University,
Miami, FL 33199, USA
{hha001,fflei001,chens}@cs.fiu.edu

***Abstract*—The booming multimedia technology is incurring a thriving multi-media data propagation. As multimedia data have become more essential, taking over a major potion of the content processed by many applications, it is important to leverage data mining methods to associate the low-level features extracted from multimedia data to high-level semantic concepts. In order to bridge the semantic gap, researchers have investigated the correlation among multiple modalities involved in multimedia data to effectively detect semantic concepts. It has been shown that multimodal fusion plays an important role in elevating the performance of both multimedia content-based retrieval and semantic concepts detection. In this paper, we propose a novel cluster-based ARC fusion method to thoroughly explore the correlation among multiple modalities and classification models. After combining features from multiple modalities, each classification model is built on one feature cluster, which is generated from our previous work FCC-MMF. The correlation between medoid of a feature cluster and a semantic concept is introduced to identify the capability of a classification model. It is further applied with the logistic regression method to refine ARC fusion method proposed in our previous work for semantic concept detection. Several experiments are conducted to compare the proposed method with other related works and the proposed method has outperform other works with higher Mean Average Precision (MAP).**

Keywords*—Semantic concept detection, Multi-model Fusion, Feature Correlation

I. Introduction

Recently, the amount of multimedia data has been drastically increasing as the development of all kinds of handheld device, such as smart phones and digital cameras, will allow people easily share their life by uploading multimedia data in one-click. Take facebook as an example, at the beginning of 2013, they announced their one billion users have uploaded 240 billion photos since the site's launch ,and each user would have more than 200 photos uploaded on an average basis. How to effectively analyze multimedia data started to draw research attention a couple decades ago, and it has become more and more crucial as a result of increasing a variety of multimedia data and enormous amount of data are populating around the world. In addition, serious challenges have been blocking the way for better multimedia data management and efficient data retrieval, such as a great diversity of data representation, high computational complexity, and lack of strong computing power.

To overcome the obstacles to multimedia research, some researchers tried to make progress by utilizing highly discriminative and robust features [1] such as Scale Invariable Feature Transformation (SIFT) [2][3] and Histogram of Oriented Gradients (HOG) [4][5]. Considering only single modality, such as analyzing audio signal for automatic transcription of speech, leveraging color features for scene recognition or using temporal features to detect different action, has also been greatly investigated. However, it has shown significant limitations while coping with tasks, which have multiple modalities involved, for instance, multimedia retrieval and multimedia event detection.

Exploiting information extracted from all the involving multiple modalities has been proven to be advantageous to multimedia analysis. Nonetheless, several major issues have not yet been adequately addressed. As a starter, handling data with different presentations such as visual, audio and text, is an issue. Moreover, how to fully employ all the given information, such as the correlation among different modalities, is also quite challenging. The other interesting topics would be identifying the useful modalities or classification models and fusing them to strengthen the achievement of multimedia related tasks.

In this paper, built on our previous works [6][7], we proposed a Multi-Model Collaboration (MMC) framework for multimedia semantic concept detection by introducing a novel cluster-based ARC fusion method, where ARC stands for adjustment, reliability and correlation of the intervals to the target semantic concept. In our previous works, features extracted from multiple modalities are transformed into feature clusters with high intra-correlation and low inter-correlation. The proposed cluster-based ARC fusion method is improved by considering the correlation between transformed feature clusters and the target concepts. Logistic regression is also applied to optimize the ranking scores in the fusion process. Finally, a threshold is set up based on experiments to eliminate the unproductive classification models build on feature clusters, hence only the classification models which have higher reliability are deployed in the fusion process.

The rest of the paper is organized as follows. Related work is introduced in Section 2. Section 3 presents the overview of the proposed MMC framework. Section 4 describes the experimental results and the framework evaluation. At the end, section 5 concludes this paper.

978-1-936968-92-3 © 2013 ICST

DOI 10.4108/icst.collaboratecom.2013.254110

II. Related Work

In the multimedia research domain, multi-modal fusion has attracted much attention not only because uni-modal approaches have their limitation to achieve complicated tasks but also because multi-modal approaches provide resourceful information for various multimedia analysis tasks. Researchers who have participated in significant image retrieval tasks, e.g., ImageCLEF [8] and TRECVid [9], have witnessed how multi-modal fusion takes over the major role in multimedia analysis. The organizers of ImageCLEF have been providing multimedia databases including images with associated text since 2003 for participants to investigate the effectiveness of multi-modal retrieval [10]. TRECVid, which has involved over 1,200 researchers from hundreds of research groups around the world, has been holding a benchmark annual activity to encourage researchers addressing multimedia related tasks, specifically semantic concept detection from video is one of the major tasks involving multiple modalities [9].

Based on a comprehensive survey article about multi-modal fusion, the fusion strategies can be mainly categorized into early and late fusion methods [11]. Early fusion can be referred to as an integration of features extracted from multiple modalities; on the other hand, an integration of the intermediate results is referred to as late fusion. We will briefly go over the related works and distinguish our proposed work from them.

With regards to early fusion, the basic approach simply concatenates features from multiple modalities into one large feature set and converts it into one consistent representation [12] [13][14]. Given one complete feature set, several research works applied Canonical Correlation Analysis (CCA) to model the correlations between features [15][16][17]. Sargin et al. [15] applied CCA to fuse audio and lip texture features to achieve audiovisual synchronization. Liu et al. [16] proposed a audio-visual fusion framework, in which CCA is used to project the audio and visual features into more compact subspaces. Hence, the correlation conveyed in the original audio and visual feature space can be preserved; meanwhile, model efficiency can be improved in the more compact feature spaces. Different from these related works, instead of leveraging correlation among features, the proposed framework increases the granularity of correlation to explore the correlation within feature-value pairs and better build the classification models on the finer captured correlations.

Late fusion, also called decision-level fusion, integrates the classification results from different modalities and generates only one result[18][19][20]. Usually, each modality is analyzed independently, so it has the flexibility to select the most suitable approach for different modalities, such as latent semantic Index (LSI) for textual modality, and hidden markov model (HMM) for audio or video modality. In addition, since the classification results collected from multiple modalities usually have the same representation, it is easier to fuse the results. However, each modality usually generates its own decision result independently, and the correlation among different modalities are overlooked in many related works adopting late fusion strategy. For example, Potamianos et al. [21] combined classification results from audio modality and visual modality and fuse two independent results with a linear weighted sum method. Chen et al. [7] proposed a fusion method called ARC and it's goal is to achieve a performance gained from all individual models. Inspired by ARC, the proposed cluster-based ARC includes all the possible correlations among modalities at feature-value pair, feature, and classification model levels to refine the factor used in model fusion and enhance the precision of semantic concept detection.

III. Overview of MMC

MMC is a multiple-model collaboration framework designed and implemented for multimedia semantic concept detection. It is improved from our previous works by introducing a novel fusion method named cluster-based ARC. With the enhanced fusion method, the correlation between classification models, which are built up from all the features extracted from multiple modalities, and the target concepts is thoroughly explored. Fig. 1 and Fig. 3 depict the training and testing components of the proposed framework, respectively.

A. Training section of MMC framework

In the training section, the first three steps, e.g., pre-processing, feature-value pair projection, and feature-value pair clustering, are the same processes as proposed in our previous work FCC-MMF [6]. In the next highlighted green square, a new factor γ is introduced to refine ARC fusion method proposed in [7]. Logistic regression is applied to obtain the optimal weighting factor in integrating γ and α. Finally, feature cluster selection is performed based on model efficiency.

From the beginning, features are extracted from all the involving modalities so that information containing different characteristics can be fully exploited. The pre-processing step includes redundant text removal, discretization and normalization to formalize the feature presentation. Multiple correspondence analysis (MCA) is employed to analyze the correlation among all the feature-value pairs by projecting each feature-value pair as one point onto the two major principal components. Subsequently, because K-medoids algorithm is one of the most prominent partitioning clustering algorithms, it is selected to separate the projected feature-value pairs into clusters and obtain the corresponding medoid for each cluster. The whole pre-processing scheme is depicted in Fig 1, FC represents a feature cluster that is converted from feature-value pair cluster based on majority vote. In this case the number of feature cluster is 4. Consequently, four classification models are built on the resulted feature clusters. The threshold can be chosen from MAP values evaluated from classification results. Mean Average Precision (MAP) is the mean of average precision (AP) of all concepts and has been shown to have especially good discrimination and stability among evaluation methods. If using the selected threshold to eliminate the redundant classification models will produce the highest MAP in detecting semantic concept, then the same threshold will be used to remove unproductive classification models in testing section. Adjustment parameter π, classification model reliability α and the correlation of an interval of scores from each classification model to the target concept β were already proposed in [7]. In this paper, we introduce a new variable γ, which represents the correlation between cluster's medoid and the target concept, to be combined with the existing variable α using weighting factors. The optimal weighting factors can be obtained through logistic regression within the training data

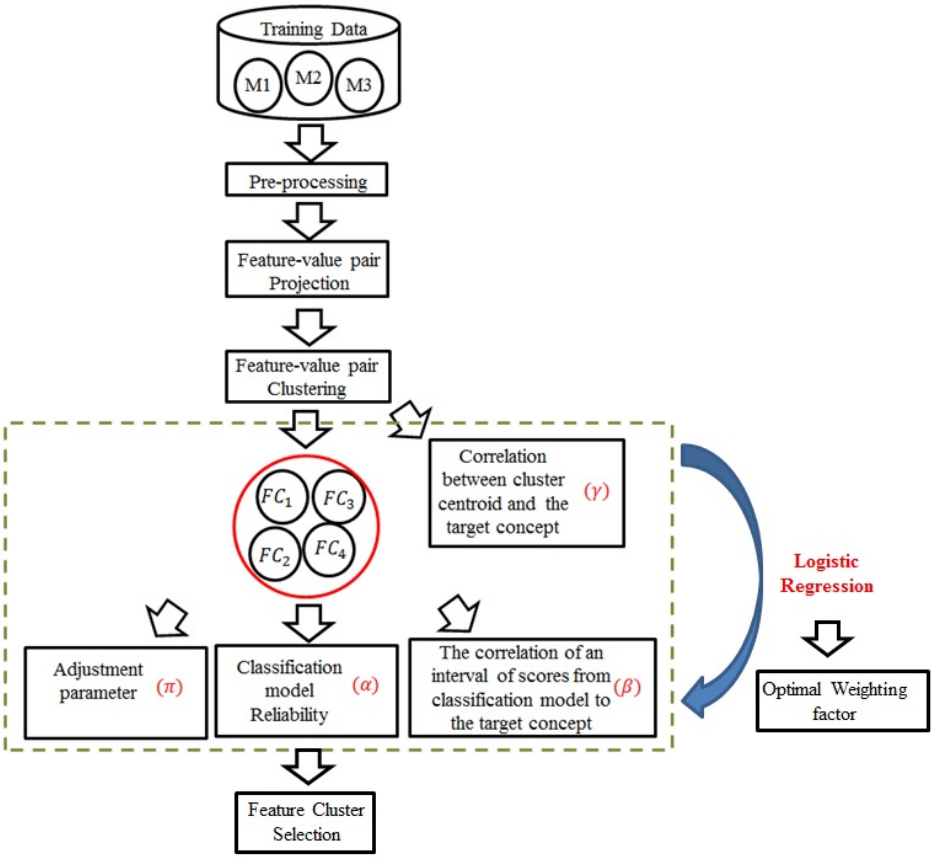


Figure 1: Training section of MMC framework

set. The deriviation of variable γ will be covered in Section III-D.

B. Testing section of MMC framework

In the testing section, features extracted from multiple modalities are also converted into the same feature clusters as described in the training section. Unuseful feature clusters are eliminated according to the threshold set up in the training section. Four variables and two weighting factors attained from training data set are utilized in the proposed cluster-based ARC fusion method to improve model fusion and semantic concept detection.

C. Feature cluster selection

As mentioned in section III-A, after performing feature extraction on all the involving modalities followed by pre-processing step, MCA is applied to analyze the correlation among feature-value pairs by projecting them onto two major principal components as shown in Fig. 2. Each blue point represents a feature-value pair and the green square points are the medoids of feature clusters obtained after feature-value pair clustering. The positive class and the negative class are represented by a red triangle and a yellow circle respectively. Once the feature-value pair clusters are converted into feature clusters based on majority vote, one classification model will is built for each feature cluster. The ranking scores produced from the classification models are evaluated in terms of MAP value,

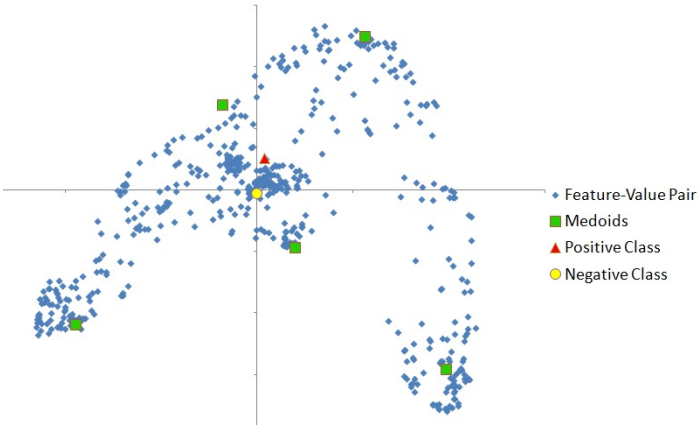


Figure 2: Feature-value pair projection and K-Medoid clustering results on a symmetric map

which is considered as classification model reliability meaning how reliable each classification model it is to accurately detect the target concept. It is also the criterion to set up the threshold in eliminating unuseful classification models. As illustrated in section III-A, the threshold for each concept is decided based on the training section results. In addition, according to the observation of experimental results, feature clusters contain less than five features are also removed because this kind of classification models are lack of distinguishing capability.

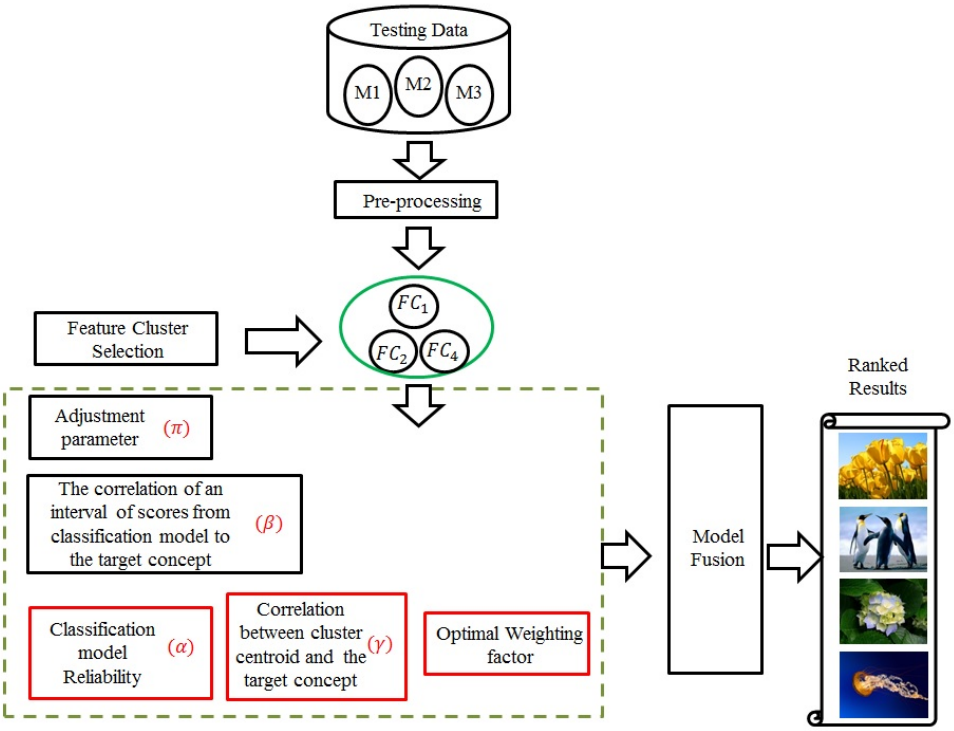


Figure 3: Testing section of MMC framework

D. Cluster-based ARC

The proposed cluster-based ARC is extended by introducing one new factor to better explore the correlation between models and concepts. The process will be depicted in two steps: section III-D1 describes how to generate variable γ and section III-D2 clarifies how to integrate γ into model fusion.

1) How to generate variable γ: As shown in Fig. 4, the symmetric map is the graphical representation after applying MCA and it can be used to visualize the medoids of feature clusters, positive class and negative class as points in a map with two dimensions, which are the first two principal components. The correlation between a medoid and a concept can be measured by the cosine value of the angle between the two vectors representing medoid and the positive class of the target concept. For example, the medoid of the third feature cluster is represented as CM_1 in Fig. 2, where Pos is the positive class and Neg is the negative class. θ_3 is the angle between CM_3 and the positive class Pos. If the absolute value of the $\cos(\theta_3)$ is large then it indicates a high correlation between the medoid CM_3 and the positive class.

As shown in equation 1, the cosine value can be obtained by using the inner product of the two vectors, e.g. medoid and positive class, and then it will be divided by the product of two vectors' length. The value of $\cos\theta_3$ is assigned to γ_i, which will be later integrated into fusion process.

$$\gamma_i = \cos(\theta)_i = \frac{Pos \cdot M_i}{\parallel Pos \parallel \parallel M_i \parallel} \tag{1}$$

2) How to integrate variable γ into model fusion: Given γ and α, which represent the correlation between model and concept and model reliability respectively, a weighting function is proposed as equation 2 to combine the two factors since they both indicate the importance of each classification model MI.

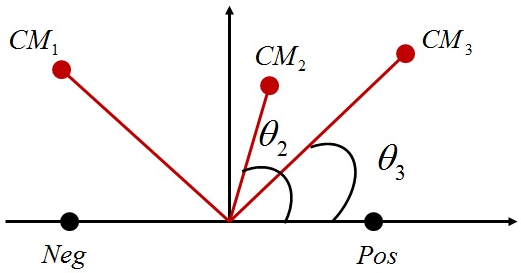


Figure 4: The symmetric map of the first two principal components

$$MI = \alpha \times \Lambda + \gamma \times (1 - \Lambda) \tag{2}$$

To combine the two factors with the most suitable weighting factor Λ, logistic regression is applied to produce the optimal weighting factor, which has been proved to minimize the error of semantic detection for each concept. Please notice that both α and γ are normalized through z-score normalization method before equation 2.

$$R(X|C) = \sum_{m=1}^{M} \frac{R_m(X|C)}{\pi} \left(\frac{MI_m \cdot \beta_m}{MI_m + \beta_m}\right) \tag{3}$$

Finally, given variable MI generated from equation 2, it replaced model reliability to have a harmonic balance with β, and normalization process was applied by introducing π, where it is the mean ranking score from different models to balance the score for X instance. $R(X|C)$ represents the final ranking score for instance X in detecting concept C. In equation 3, M is the number of classification models after feature cluster selection and β represents the correlation of an interval of scores from a ranking model to the target concept as described in [7].

In the next section, several experiments are conducted to compare the proposed fusion method with other fusion methods, i.e. min, max, mean, average, median, and ARC. The proposed cluster-based ARC was able to show that it better fuses the models by fully exploring the correlation among them and identifying model's importance in detecting concept.

IV. Experimental Analysis

A. Experiment Setup

To systematically evaluate the proposed framework, two experiments are designed: one demonstrates the fusion performance of the proposed framework, and the other one shows the improvement against our previous work. 3-fold cross validation and mean average precision (MAP) is applied on both experiments to validate the results of semantic concept detection. MCA is used as the classification modeling to evaluate the feasibility of the proposed framework [22][23][24].

In the first set of experiments, the performance of our proposed cluster-based ARC fusion method in semantic concept detection against several well-known fusion methods is demonstrated on both NUS-WIDE-LITE and NUS-WIDE 270K datasets [25]. NUS-WIDE data set is one of the largest real-world web image datasets including over 269,000 images with the ground-truth information of 81 concepts. With regard to visual features, most common visual features such as color histogram, edge direction histogram, wavelet texture, and bag of words based on SIFT descriptions are included with the dataset. In addition, all the images have their associated tags from flickr to represent as textual features. Therefore, it is the perfect benchmark dataset for multimedia semantic concept detection.

In the second set of experiments, the proposed MMC framework is also compared against our previous works, e.g. (CFA-MMF)[26] and (FCC-MMF)[6], and the original flat concatenation of multi-modality features, which is exact the same feature set but it is only trained as one classifier and there is no fusion process involved, on NUS-WIDE-LITE dataset.

B. Evaluation of Cluster-based ARC fusion method

Five reputable fusion methods including, minimum (min), maximum (max), average, mean, and median are adopted to be compared with the proposed cluster-based fusion method. The comparative experimental result demonstrated on NUS-WIDE LITE dataset is shown on Fig. 5a. It is observed that cluster-based ARC outperforms other well-known fusion approaches up to 25%. The positive difference between cluster-based ARC and the original ARC indicates the improvement in detecting semantic concepts. The lowest MAP value produced by median method is only 11% and our proposed method outperforms by 25%.

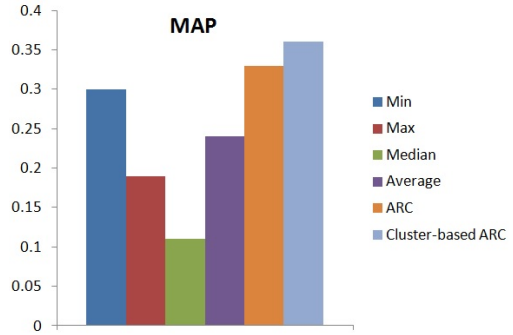
Fig. 5c shows the comparison results of the above-mentioned fusion methods on NUS-WIDE-270K dataset. The comparative results are quite similar as shown in Fig. 5a, however the relatively large size of the dataset has resulted in lower MAP values for all the fusion methods. Our proposed method was again validated to obtain the highest MAP value, where it is 15% and 2% higher than the worst and the best performance respectively. In addition, the performance of cluster-based ARC still outperforms the original ARC.

C. Evaluation of MMC framework

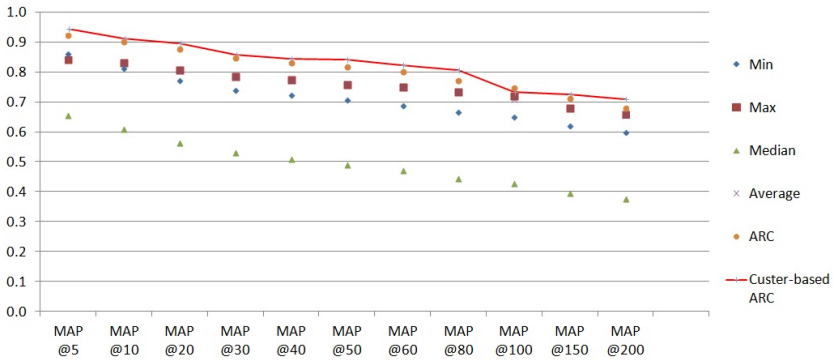
The proposed MMC framework is compared against our previous works, e.g. Feature Correlation Clustering-based Multi-Modality Fusion Framework (FCC-MMF) and Correlation-Based Feature Analysis and Multi-Modality Fusion Framework (CFA-MMF), and the original feature set, which simply combines all the features into one classification model, to validate whether our research work has been continuously advanced to adequately detect semantic concepts from multimedia data. The framework comparisons are carried out on NUS-WIDE-LITE dataset. As shown in Fig. 6 and Fig. 7, the proposed framework outperforms the previous works up to 20% in the first three retrieval scales. With regard to this experiment, couple observations are listed as follows: CFA-MMF was enhanced by reducing feature dimension while remaining comparative performance against original feature set; FCC-MMF converted features from multiple modalities into high intra-correlation and low inter-correlation feature clusters and they were consecutively trained as classification models where fusion process was applied to produce the final ranking scores; the proposed framework MMC went beyond FCC-MMF and further considered how feature cluster's medoid correlated to semantic concepts to improve fusion process.

V. Conclusion

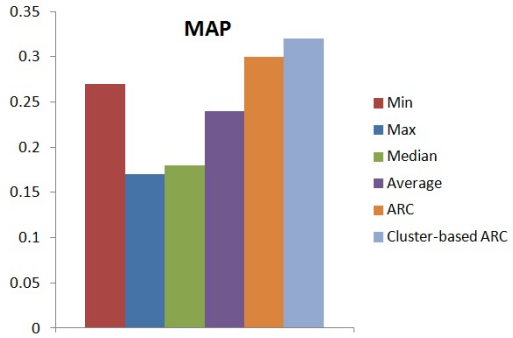
The paper presents a multi-model collaboration framework including an enhanced fusion method called cluster-based ARC to effectively detect semantic concepts from multimedia data. Because of the experience learnt from our previous works, exploring correlation among multi-modalities has been proven effective in multimedia semantic concept detection. Therefore, the association between classification models and semantic concepts is combined with model reliability to represents as model importance, which indicates how useful the model it is in detecting this semantic concept. The idea of using only the useful feature clusters is also introduced in our framework. The proposed framework aims at thoroughly exploiting all the possible correlation from multiple modalities to build up a multi-model collaboration in semantic concept detection. The experiments are conducted on NUS-WIDE-LITE and NUS-WIDE-270K datasets to evaluate the propose framework. The comparative experimental results of 3-fold cross validation showed that the proposed framework outperforms several well-known fusion methods and our previous works in terms of MAP values.



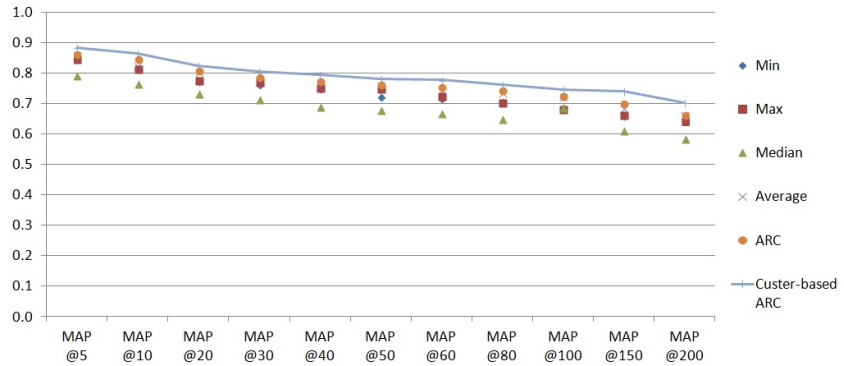
(a) MAP values of 81 concepts after model fusion on the NUS-WIDE-LITE dataset



(b) MAP values at different retrieval scales of 81 concepts after model fusion on the NUS-WIDE-LITE dataset



(c) MAP values of 81 concepts after model fusion on the NUS-WIDE-270K dataset



(d) MAP values at different retrieval scales of 81 concepts after model fusion on the NUS-WIDE-270K

Figure 5: MAP values after model fusion on NUS-WIDE dataset

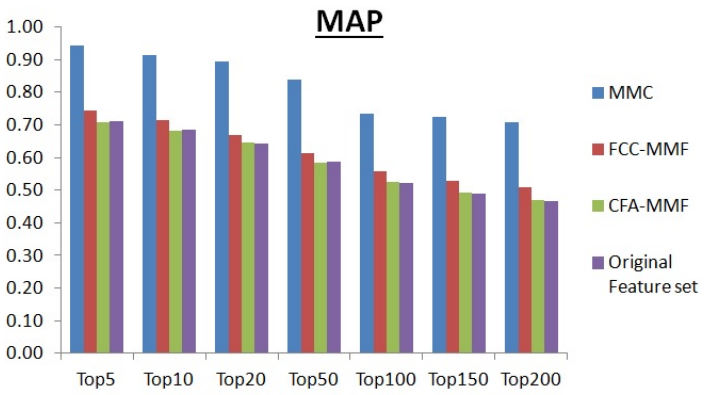


Figure 6: MAP values at different retrieval scales of 81 concepts of different frameworks on the NUS-WIDE-LITE dataset

	Top5	Top10	Top20	Top50	Top100	Top150	Top200
MMC	0.9434	0.9122	0.8958	0.8097	0.7332	0.7233	0.7075
FCC-MMF	0.7449	0.7133	0.6705	0.6118	0.5581	0.5276	0.5085
CFA-MMF	0.7077	0.6831	0.6451	0.5825	0.5246	0.4920	0.4709
Original Feature set	0.7127	0.6845	0.6426	0.5870	0.5224	0.4882	0.4657

Figure 7: MAP values at different retrieval scales of 81 concepts of different frameworks on the NUS-WIDE-LITE dataset

ACKNOWLEDGMENT

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Securitys VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

REFERENCES

[1] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[2] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching.," in *TRECVID*, 2010.

[3] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 251–260.

[4] Hyun Oh Song, Stefan Zickler, Tim Althoff, Ross Girshick, Mario Fritz, Christopher Geyer, Pedro Felzenszwalb, and Trevor Darrell, "Sparselet models for efficient multiclass object detection," in *Computer Vision–ECCV 2012*, pp. 802–815. Springer, 2012.

[5] Gary Overett and Lars Petersson, "Large scale sign detection using hog feature variants," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 326–331.

[6] Hsin-Yu Ha, Fausto C Fleites, and Shu-Ching Chen, "Content-based multimedia retrieval using feature correlation clustering and fusion," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 4, no. 2, pp. 46–64, 2013.

[7] Chao Chen, Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology*, 2013.

[8] Henning Mèuller, Paul Clough, Thomas Deselaers, and Barbara Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, vol. 32, Springer, 2010.

[9] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[10] William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen, "Medical image retrieval and automated annotation: Ohsu at imageclef 2006," in *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 660–669. Springer, 2007.

[11] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[12] Abhishek Nagar, Karthik Nandakumar, and Anil K Jain, "Multibiometric cryptosystems based on feature-level fusion," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 255–268, 2012.

[13] Jana Kludas, Eric Bruno, and Stephane Marchand-Maillet, "Information fusion in multimedia information retrieval," in *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pp. 147–159. Springer, 2008.

[14] Nan Luo, Zhenhua Guo, Gang Wu, and Changjiang Song, "Multispectral palmprint recognition by feature level fusion," in *Recent Advances in Computer Science and Information Engineering*, pp. 427–432. Springer, 2012.

[15] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[16] Ming Liu, Yun Fu, and Thomas S Huang, "An audio-visual fusion framework with joint dimensionality reducton," in *Acoustics, Speech*

and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008, pp. 4437–4440.

[17] Xiaona Xu and Zhichun Mu, "Feature fusion method based on kcca for ear and profile face based multimodal recognition," in *2007 IEEE International Conference on Automation and Logistics*. IEEE, 2007, pp. 620–623.

[18] Hervé Bredin and Gérard Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 2, pp. II–233.

[19] Bakkama Srinath Reddy, "Evidential reasoning for multimodal fusion in human computer interaction," M.S. thesis, University of Waterloo, 2007.

[20] Hatice Gunes and Massimo Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. IEEE, 2005, vol. 4, pp. 3437–3443.

[21] Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003.

[22] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Correlation-based interestingness measure for video semantic concept detection," in *IEEE International Conference on Information Reuse & Integration, 2009. IRI'09.* IEEE, 2009, pp. 120–125.

[23] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *11th IEEE International Symposium on Multimedia, 2009. ISM'09.* IEEE, 2009, pp. 304–311.

[24] Lin Lin, Chao Chen, Mei-Ling Shyu, and Shu-Ching Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *MultiMedia, IEEE*, vol. 18, no. 3, pp. 32–43, 2011.

[25] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, p. 48.

[26] Hsin-Yu Ha, Yimin Yang, Fausto C Fleites, and Shu-Ching Chen, "Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval," in *2013 IEEE International Conference on International Conference on Multimedia and Expo (ICME), "Multimedia for Humanity" Theme Track*, JUL 2013.

Content-based Multimedia Retrieval Using Feature Correlation Clustering and Fusion

Hsin-Yu Ha*, Fausto C. Fleites, and Shu-Ching Chen
School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
email: hha001@cs.fiu.edu, fflei001@cs.fiu.edu, chens@cs.fiu.edu

Abstract

Nowadays, only processing visual features is not enough for multimedia semantic retrieval due to the complexity of multimedia data, which usually involve a variety of modalities, e.g. graphics, text, speech, video, etc. It becomes crucial to fully utilize the correlation between each feature and the target concept, the feature correlation within modalities, and the feature correlation across modalities. In this paper, we propose a Feature Correlation Clustering-based Multi-Modality Fusion Framework (FCC-MMF) for multimedia semantic retrieval. Features from different modalities are combined into one feature set with the same representation via a normalization and discretization process. Within and across modalities, multiple correspondence analysis is utilized to obtain the correlation between feature-value pairs, which are then projected onto the two principal components. K-medoids algorithm, which is a widely used partitioned clustering algorithm, is selected to minimize the Euclidean distance within the resulted clusters and produce high intra-correlated feature-value pair clusters. Majority vote is applied to subsequently decide which cluster each feature belongs to. Once the feature clusters are formed, one classifier is built and trained for each cluster. The correlation and confidence of each classifier are considered while fusing the classification scores, and mean average precision is used to evaluate the final ranked classification scores. Finally, the proposed framework is applied on NUS-wide Lite data set to demonstrate the effectiveness in multimedia semantic retrieval.

Keywords: *K-medoids, multimedia feature correlation, multiple correspondence analysis (MCA),*

content-based multimedia retrieval, multi-modality fusion

I. INTRODUCTION

As a result of the rapid improvement of contemporary technology, people usually have smartphones that easily allow capturing images and recording video and instantly sharing the multimedia content with corresponding descriptions with friends over social networks, a trend that has resulted in multimedia data propagating expeditiously around the world. A study presented by IDC and EMC stated that 1,800 EB (1 EB = 1,000 PB) of digital information were produced in 2011, and the amount of information increased ten times from 2005 to 2011 (Gantz et al., 2008). To manage the enormous volume of multimedia data, i.e., images, videos, texts, and audio, how to effectively retrieve data from different modalities and bridge the gap between low-level features and various semantic concepts becomes more and more essential. Many researchers have been investigating multi-modal fusion for multimedia analysis, e.g. video retrieval (Yan, Yang & Hauptmann, 2004; McDonald & Smeaton, 2005), speech recognition (Metallinou, Lee & Narayanan, 2010; Papandreou, Katsamanis, Pitsikalis & Maragos, 2009), event detection (Jiang et al., 2010; Mertens, Lei, Gottlieb Friedland & Divakaran, 2011), etc. However, because of the involved modalities, multi-modal fusion has many challenges: coping with different feature formats, capturing correlation and independence among modalities in many levels, and detecting the confidence level of each model in achieving tasks.

To resolve the above-mentioned challenges, Atrey et al. (Atrey et al., 2010) pointed out several questions for multimedia analysis; in particular some of them were comprehensively thought through for content-based multimedia retrieval:

- **At which level should the fusion be performed?** There are mainly two fusion levels: feature level and decision level. For feature level, features from multiple modalities may simply be concatenated and converted into one common representation space for follow-up analysis (Nagar,

Nandakumar & Jain, 2012; Luo, Guo, Wu, & Song, 2012; Kludas, Bruno, & Marchand-Maillet, 2008; Wimmer et al., 2008). This is the most common type of audio-visual fusion (Snoek, Worring & Smeulders, 2005). Principal Component Analysis (PCA) (Turk & Pentland, 1991) and Independent Component Analysis (ICA) (Lee, 1998) are often used after combining the features to extract the discriminant features and thus reduce the feature space (Bhanu & Han, 2011; Feng, Dong, Hu & Zhang, 2004; Buganim et al., 2012; Schuller et al., 2005). The correlation between multiple features from different modalities may be leveraged at an early stage to enhance the final results. One major advantage for feature-level fusion is that it requires only one classifier after the fusion step (Snoek, Worring & Smeulders, 2005; Yang, Wang & Lin, 2005). Figure 1 (a) depicts the general approach for feature-level fusion, where F represents features from a single or multiple modalities, FF represents the fusion step, AU represents a single analysis unit (e.g. learning algorithms, feature extraction, feature selection, and feature transformation), and D represents the decision made from an analysis unit. On the other hand, decision-level fusion analyzes the classification results from different modalities and produces only one result (Bredin & Chllet, 2007; Reddy, 2007; Clinchant, Ah-Pine, & Csurka, 2011). Figure 1 (b) depicts such a fusion approach, where DF represents the fusion step. Since each modality is analyzed independently, there is flexibility to select more appropriate methods for each specific modality, for example, Latent semantic Index (LSI) (Landauer, Foltz & Laham, 1998) for textual modality, hidden Markov model (HMM) (Eddy, 1996) for audio or video modality, and support vector machine (SVM) (Furey, 2000) for image modality. In addition, another advantage is that it will be easier to fuse the final classification results, which usually share the same representation. However, decision-level fusion does not fully exploit the feature correlation among modalities when building the classifiers. As shown in Figure 1 (c), we propose a hybrid fusion method to exploit the advantages from both feature-level fusion and decision-level fusion. After combining features from different modalities, only one analysis unit is applied to capture the feature

correlation across modalities and transform them into feature clusters that obtain higher within-cluster correlation. At the end, the ranked classification results produced from each cluster are fused at the decision level. Other works in the literature have also applied hybrid fusion approaches, although different from ours (Song et al., 2004; Zeng et al.,2007; Zeng et al.,2008; Mansoorizadeh & Charkari, 2010; Mangai, Samanta, & Chowdhury, 2010). Islam et al. proposed a three-phase fusion process toward audio and visual modalities: fusion within single modality, fusion across modality in feature level, and fusion on decision level according to the reliability of each modality (Islam & Rahman, 2010). Zhang proposed to employ a manifold learning method called spectral regression to deal with the problem of a large feature space while performing feature fusion, and then fuzzy aggregation is applied to combine the distance metrics for decision fusion level (Zhang, 2011).

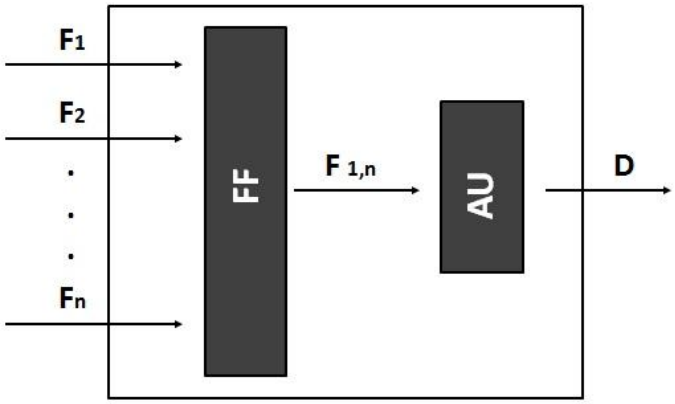


Figure 1 (a) Feature Level Fusion

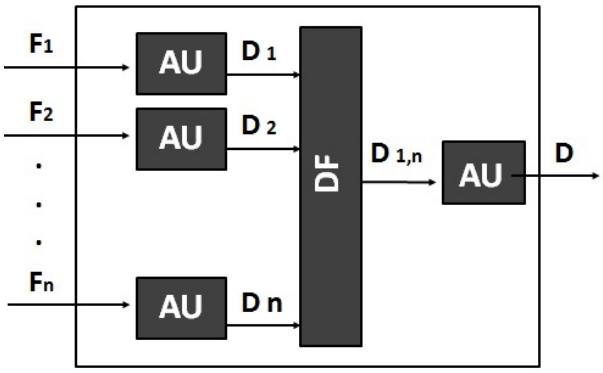


Figure 1 (b) Decision Level Fusion

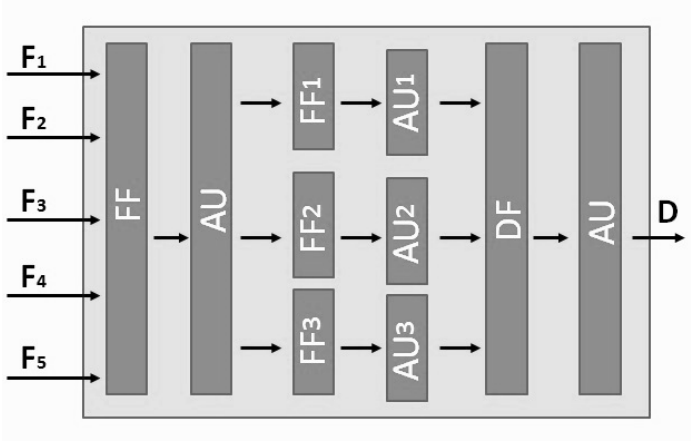


Figure 1 (c) Overall Fusion Framework for FCC-MMF

Figure 1: Different Types of Fusion Levels

- **How should the fusion be carried out?** Multimodal fusion methods can be mainly categorized into three types: rule-based methods, classification methods, and estimation methods. Rule-based methods mainly consist of linear-weighted approaches that statistically capture the correlation between features and semantic concepts and then assign normalized weights per feature. Many researchers have been investigating how to obtain the optimal weights for the features and/or modalities. Wei et al. proposed an approach named concept-driven multi-modality fusion (CDMF) to compute multi-modality fusion weights from predefined semantic concepts. CDMF includes two components to analyze the relationship between an executed query and a modality. In the first component, a set of semantically and visually relevant semantic concepts is inferred based on the text words and the visual examples provided from executed queries. A context graph is built offline to capture the co-occurrence relations among these semantic concepts. Then random walk is applied to model the interaction among concepts over the context graph and produce the relevance of these concepts to the query. In the second component, a relation matrix, which is learnt offline to model the reliability of each modality based on its concept detection accuracy, is integrated with the concept relevance to produce the final fusion weights, which indicates the correlation between the executed query and the involved modalities using fuzzy transformation. (Wei, Jiang, & Ngo, 2011). Lan et al. proposed a methodology called double

fusion that adopts both the average of kernel matrices and multiple kernel learning to automatically learn the weights for different kernel matrices after combining features from multiple modalities (Lan et al., 2012). Rashid et al. explored a variation of linear combination techniques, e.g. fuzzy logic techniques, sequential techniques, and linear combination models, and investigated how to adjust the inter-modality and intra-modality weights (Rashid, Niaz & Bhatti, 2010). Classification-based fusion methods leverage the ability of classification methods in classifying features from different modalities into one of more pre-defined classes for each semantic concept. Classification models such as support vector machine (SVM) and hidden Markov models have both been applied for fusion purposes (Ayache, Quenot, & Gensel, 2007; Li et al., 2013; Liu, Zheng, & Jiang, 2009; Dang-Nguyen et al., 2012). Adams et al. compared the results between Bayesian networks and support vector machines in classifying scores from multiple modalities that are more related to semantic concepts (Adams, 2003). Nicolaou et al. proposed to adopt decision-level fusion based on Coupled Hidden Markov Model (CHMM) (Murphy, 2001), which is a series of parallel HMM chains coupled through cross-time and cross-chain conditional probabilities hence, to model the intrinsic temporal correlation between the modalities (Nicolaou, Gunes, & Pantic, 2010). Jiang et al. proposed to collectively classify low-features, transform high-level features into graphs, and fuse the classification scores along with the constructed graph to obtain the final prediction (Jiang, Hauptmann, & Xiang, 2012). Estimated methods, including Kalman filter, extended Kalman filter, and particle filter, are usually adopted when the tasks involve temporal motion, such as estimation of moving objects in real-time. Zhang et al. proposed to fuse inertial and magnetic sensor data using a particle filter to better cope with the nonlinear human body segment motion (Zhang & Wu, 2011). To perform real-time human tracking, Motai et al. proposed to fuse the relative tracking data with an optical flow Kalman filter (OFKF) (Motai, Jha & Kruse, 2012). In the proposed framework, due to its

ability to linearly capture the fusion weights and the fact that its effectiveness has been proven in many works, a rule-based method is selected to enhance the performance at the decision level.

- **What should be fused?** Usually, either features or modalities will be fused based on their ability to retrieve semantic concepts (Clausi & Deng, 2005; Natsev, Naphade, & TesiC, 2005; Clark et al., 1993; Cui, Tung, Zhang, & Zhao, 2010; Candemir et al., 2012). For example, Li et al. proposed to use the resulting weights of the Ordered Weighted Average (OWA) operator to yield a consensus fusion score from multi-modalities (Li, 2009). Zhou et al. suggested combining the normalized classification results of both images and documents to better perform information retrieval (Zhou, Depeursinge & Muller, 2010). Zou et al. proposed to compare two approaches for detecting human movement using video and audio sequences: one applied Time-Delay Neural Network (TDNN) to fuse audio and visual data at the feature level, and the other one employed Bayesian Network (BN) to collectively model video and audio signals (Zoi & Bhanu, 2005). Besides fusing features and modalities specifically, Ye et al. proposed a joint audio-visual bi-modal representation, called bi-modal words. A bipartite graph is built from visual and audio modalities, which is later partitioned into bi-modal words that can be also considered as joint patterns across modalities. Consequently, the joint patterns are transformed into bimodal Bag-of-Words representations and considered as input to the classifiers (Ye, 2012). Similarity scores between queries and the database images are also proposed in (Chandrakala & Sumathi, 2012) as fusing targets. Chandrakala et al. propose to use Artificial Bee Colony optimization algorithm to fuse the similarity scores based on texture and color features of an image. In this paper, we propose to integrate and fuse the features from all the modalities at the feature-value pair level. Feature-value pair clusters are formed based on correlation among feature-value pairs and later converted into highly correlated feature clusters. One classifier is subsequently trained for each feature cluster to generate the scores that are fused at the decision level.

The main contributions of this work are summarized as follows:

- Utilizes multiple correspondence analysis (MCA) to jointly capture the inter- and intra-modality correlation between features and semantic concepts, different from previous works that consider independence between features when applying MCA (Lin & Shyu, 2010a; Lin & Shyu, 2010b; Lin, Chen, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2010).
- Presents a Feature Correlation Clustering algorithm that utilizes the K-medoids to group feature-value pairs into clusters that are subsequently transformed into feature-level clusters with high intra-correlation and low inter-correlation.

To demonstrate the effectiveness of the proposed feature correlation clustering-based multimodal fusion framework (FCC-MMF), 81 high-level semantic concepts and 55,615 images with the corresponding tags are used to compare FCC-MMF with other related works.

This rest of paper is organized as follows. Section II presents the details of the proposed FCC-MMF framework. Section 3 discusses the experimental results as well as our observations. The paper is concluded with our future works in Section 4.

II. FEATURE CORRELATION CLUSTERING-BASED MULTIMODLITY FUSION FRAMEWORK

Extending our previous work (Ha, Yang, Fleites, & Chen, 2013), we have designed and developed a Feature Correlation Clustering-based Multi-Modality Fusion Framework, as shown in Figure 2. The framework is composed of five major components: (1) pre-processing, (2) feature-value pair projection, (3) feature clustering, (4) model training, and (5) model fusion. Firstly, pre-processing extracts the features from the multiple modalities, performs supervised discretization to obtain the

same representation, and removes the features that have no variance to reduce feature space by eliminating meaningless information. Secondly, multiple correspondence analysis (MCA), whose effectiveness in video concept detection, feature selection, discretization, etc. has been proven in previous works (Lin, Shyu, & Chen, 2009; Lin, Shyu, Ravitz, & Chen, 2009; Lin, Shyu, Ravitz, & Chen, 2008; Lin, Shyu, Ravitz, & Chen, 2007; Lin, Chen, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2010; Zhu, Lin, Shyu, & Chen, 2011) is utilized to discover associations among variables for large-scale multimedia data and project the feature-value pairs onto the two major principal components. Subsequently, the K-medoids algorithm, one of the most prominent partitioning clustering algorithms, is applied to separate the projected feature-value pairs into clusters. Since feature-value pairs of the same feature can fall into different clusters, each feature is then assigned to the cluster that contains the majority of the feature's feature-value pairs. One classifier is built for each cluster, which exhibits high intra-correlation. Once the classification results from training data are generated, the number of clusters that produce the highest MAP can be obtained for each semantic concept. Once the optimal number of clusters is obtained, test data are assigned to the previously obtained clusters and classified using the cluster-wide trained classifiers. The model fusion step is performed using the model fusion strategy from (Ha, Yang, Fleites, & Chen, 2013) with some adjustments that normalize the confidence degree for each feature cluster using min-max normalization.

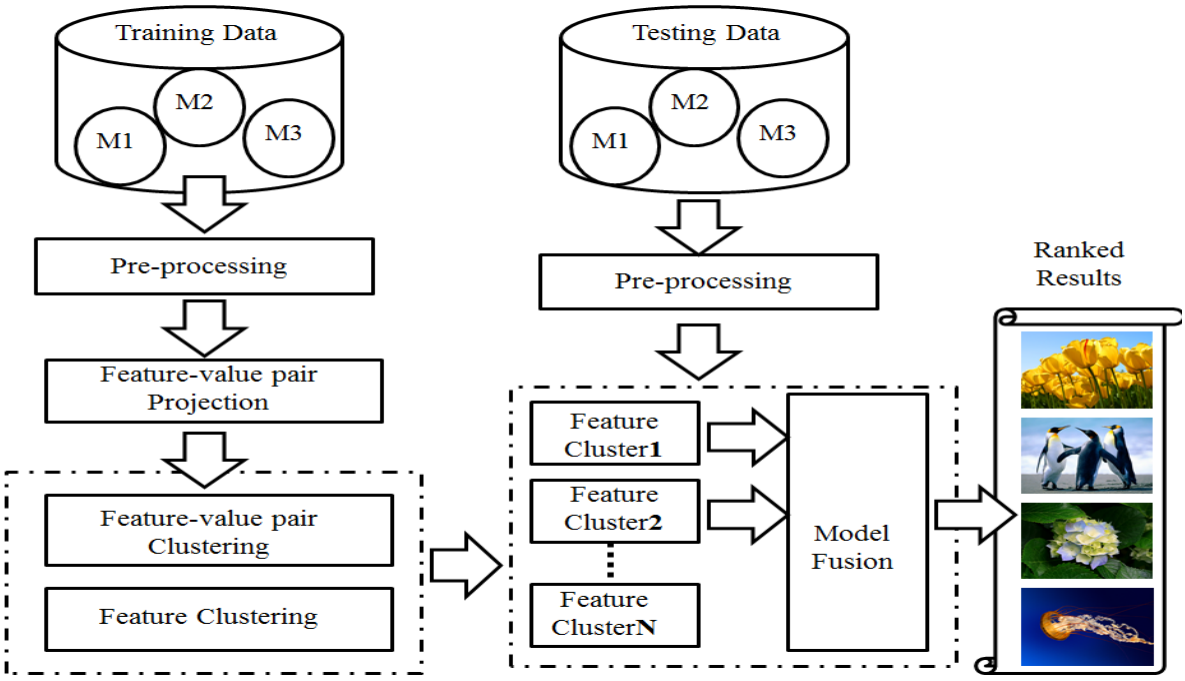


Figure 2. [Architecture of the proposed Feature Correlation Clustering-based Multi-Modality Fusion (FCC-MMF) framework].

There are two major improvements compared to our previous work (Ha, Yang, Fleites, & Chen, 2013). One is that we utilize MCA to capture the correlation between the features instead of Spearman's rank correlation coefficients and Pearson product-moment correlation coefficient. The rationale is that MCA generates correlation information at a lower level, i.e., for feature-value pairs, which helps in obtaining more-accurate correlation information, as demonstrated in the experiments. The other improvement is the utilization of K-medoids instead of Maximum Spanning Tree (MaxST) (Agarwal, Matousek, & Suri, 1992) to obtain feature clusters. K-medoids can select multiple clusters, whereas MaxST, as utilized in our previous work, is only able to produce two clusters. In section 3, the proposed framework is able to demonstrate the impact of these two enhancements by showing better classification results.

The main components of the proposed framework are further explained as follows.

Feature-value pair projection

Multiple correspondence analysis (MCA) (Greenacre & Blasius, 2006), which stems from standard Correspondence Analysis (CA), can be employed to analyze more than two variables, and it is used in this paper to analyze features from multi-modalities. After discretizing the data in the pre-processing step, the nominal data are used to construct an indicator matrix that represents instances as rows and feature-value pairs as columns. Mostly, only one feature is combined with the class to form an indicator matrix because detecting the correlation between one feature and the target concept is considered as an independent task (Lin & Shyu, 2010a; Lin & Shyu, 2010b; Lin, Chen, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2010). In our proposed framework, all the features from all the modalities are considered in building the indicator matrix. As shown in Figure 3, features from different modalities with numerical values, i.e. feature 10 (color histogram) and feature 24 (wavelet texture) are discretized into feature-value intervals, where feature 30 (tags) is already represented as binary value. In this example, feature 10, feature 24, and feature 30 are discretized into three, two, and two intervals respectively, and F_1^{10} represents the first interval in feature 10. Usually, two classes represent the positive semantic concept and negative semantic concept. Assuming *m* features have *K* feature-value pairs in total with *C* as the number of classes and *N* as the number of instances, then the indicator matrix, which represents the partitioned intervals via binary values, is denoted by *I* with size $N \times (K + C)$. Once the indicator matrix is generated, the same procedures described in (Lin & Shyu, 2010a; Lin & Shyu, 2010b; Lin, Chen, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2011; Zhu, Lin, Shyu, & Chen, 2010) are implemented in our framework. The Burt matrix Y, which is the inner product of the indicator matrix *I*, probability matrix $\boldsymbol{Z}$, mass matrix $\boldsymbol{M}$, and diagonal matrix $\boldsymbol{D}$, are analyzed and decomposed by Singular Value Decomposition (SVD) as described in Equation (1). With the eigenvalues and eigenvectors obtained from SVD, each

feature-value pair is projected onto the first principal component (the eigenvector with the largest eigenvalue) and the second principle component (the eigenvector with the second largest eigenvalue), which can be visualized as a point in a two dimensional map.

$$\boldsymbol{D}^{-1/2}(\boldsymbol{Z} - \boldsymbol{M}\boldsymbol{M}^{T})(\boldsymbol{D}^{T})^{-1/2} = \boldsymbol{P}\Delta\mathbf{Q}^{T}. \quad (1)$$

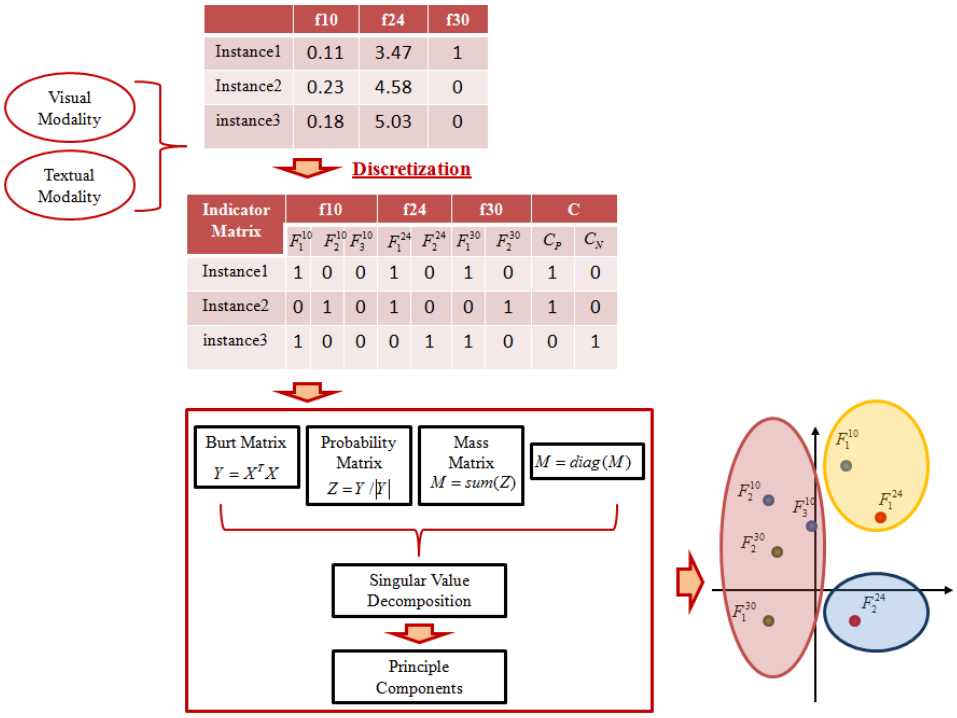


Figure 3. [The Process of Feature value-pair Projection]

Feature correlation clustering

Feature correlation clustering is actually a two-step process: feature-value pair clustering and feature clustering based on majority rule. As depicted in Algorithm 1, the first phase of Feature Correlation Clustering (FCC) is to iteratively apply K-medoids to all the feature-value pair points to produce clusters . FV_{ij} represents feature-value pair that belongs to the i^{th} feature and is the j^{th} feature-value pair within that feature. All the feature-value pairs are the input of K-medoids algorithm. The number of iterations is W, a value that is set based on experimental observation. For each iteration, initial

medoids are chosen randomly and each data point is associated to the closest medoid using a valid distance metric; Euclidean distance is selected in the proposed framework to represent the similarity between a pair of feature-value pair. *K* feature-value pair clusters are generated as the results of Algorithm 1. We decide to use K-medoids algorithm because it is less affected by outliers and noise compared to k-means, and it guarantees to generate the number of clusters as assigned.

Algorithm 1: Feature-value pair Correlation Clustering

Input: $FV = \{FV_{11}, FV_{12}, ..., FV_{NM}\}$, where N is the number of features and M is the number of feature-value pairs of the Nth feature
K is the number of clusters

Output: $FVC = \{FVC_1, FVC_2, ..., FVC_k\}$, where *FVC* represents *K* feature-value pair clusters.

1. **for** $i \leftarrow 0$ to *W* **do**
2. *K-Medoids(FVC)*
3. **End for**

Algorithm 2: Feature Correlation Clustering

Input:
$FVC = \{FVC_1, FVC_2, ..., FVC_k\}$, where *FVC* represents *K* feature-value pair clusters. N is the number of features

Output: $F = \{F_1, F_2, ..., F_N\}$, where F_i represents which cluster feature $Feature_i$ belongs to.

1. **for** $i \leftarrow 0$ to *N* **do**
2. **for** $j \leftarrow 0$ to *K* **do**
3. **C'** $\leftarrow$ *has_MaxNumber_FVPairs*(FVC_j)
4. **End for**
5. **If** *count(C') >1* **then**
6. $F_i \leftarrow$ *random_assign(C')*
7. **Else**
8. $F_i \leftarrow C'$
9. **End if**
10. $C' \leftarrow \emptyset$
11. **End for**

Feature correlation clustering is performed on the clusters produced from Algorithm 1. Depicted in Algorithm 2, the proposed Feature Correlation Clustering has the following steps:

1) Loop through all the features
2) For each feature $Feature_i$, go through all the clusters FVC_j that belongs to FVC to obtain the cluster that contains the largest number of feature-value pairs from this specific feature.
3) Put the clusters which contains largest number of feature-value pairs in the cluster list C'.
4) Check the number of clusters placed in cluster list C'. If cluster list C' contains more than one cluster, then one of those clusters will be randomly picked as the cluster of feature $Feature_i$, which is F_i. Otherwise, the only cluster placed in cluster list C' is assigned as the cluster of feature $Feature_i$, which is F_i.
5) Empty cluster list C' and repeat step 2 through step 4 for the next feature until all the features are assigned into one of the clusters.

For example, if $Feature_1$ has 5 intervals, i.e. 5 feature-value pairs, and the first two feature-value pairs and the last three feature-value pairs are assigned to cluster FVC_2 and cluster FVC_4, then the cluster which contains largest number of feature-value pairs from $Feature_1$ will be placed into cluster list C', which will be FVC_4. Since there is only one cluster in cluster list C', FVC_4 will be directly assigned as the feature cluster F_1 for feature $Feature_1$. In other scenario, if $Feature_2$ has 6 intervals and both FVC_1 and FVC_3 are assigned three feature-value pairs from $Feature_2$, then FVC_1 and FVC_3 will be both placed in cluster list C' and one of them will be randomly picked as the cluster of feature $Feature_2$ since there are more than one cluster in cluster list.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The NUS-WIDE-Lite dataset (Chua, Tang, Hong, Li, Luo, & Zheng, 2009), which contains 55,615 images including the corresponding tags, is selected to evaluate the performance of the proposed framework. 3-fold cross-validation is applied, which result in 37,077 images as training data set and

18,540 images as test data. It is also applied to decide the most suitable number of clusters for each concept. 81 concepts are provided from the data set along with the ground truth. In addition, the method proposed in (Chen, Zhu, Lin, Shyu, & Chen, 2013) is used to remove the redundant textual features before the experiments.

Experiment Setup

To demonstrate the enhancement of the proposed FCC-MMF framework, the comparison experiments are conducted against CCA-MMF framework from (Ha, Yang, Fleites, & Chen, 2013) and the original flat concatenation of multi-modality features, which is exact the same feature set but it is only trained as one classifier and there is no fusion process involved. The number of cluster is set from 4 to 10 according to the observation that each concept usually contains up to 400 features, and the performance will begin to drop down if the number of cluster is larger than 10. In addition, the proposed FCC-MMF framework is also compared to other applications using NUS-WIDE-Lite dataset in terms of multimedia semantic retrieval performance.

Due to the fact that mean average precision (MAP) is one of the most common evaluation criteria in multimedia retrieval, and it has been shown to have good discrimination and stability, it is selected to be the performance indicator to compare the proposed framework with other works. First of all, if there are n ranked data instances, Precision(n), which is the precision of the n retrieved data instances, can be calculated as described in Equation (2), where *TP* represents total number of retrieved instances that are correctly classified as positive. Let *Relevance(i)* be an indicator function that will return value 1 if data instance *i* is relevant and 0 is returned otherwise. Given *TP+FN* as total number of the relevant positive data instances, K is the total number of retrieved positive data instances, *i* represents each ranked instance, the average precision (*AP*) can be calculated as shown in Equation

(3). Mean average precision can be further calculated as in Equation (4), where C is the total number of retrieved semantic concepts.

$$\mathrm{Precision}(n) = \frac{TPn}{n}; \tag{2}$$

$$Average\,\mathrm{Precison} = \frac{\sum_{i=1}^{K} \mathrm{Precision}(i) \times \mathrm{Relevance}(i)}{TP + FN}; \tag{3}$$

$$MeanAverage\,\mathrm{Precison} = \frac{\sum_{J=1}^{C} Average\,\mathrm{Precision}(J)}{C}; \tag{4}$$

The number of clusters for each concept is decided by selecting highest MAP from different number of training clusters as depicted in Figure 4. There is no obvious trend or linear relationship between the concept and the selected number of clusters and the decided number of cluster could be quite different among different folds for the same concept.

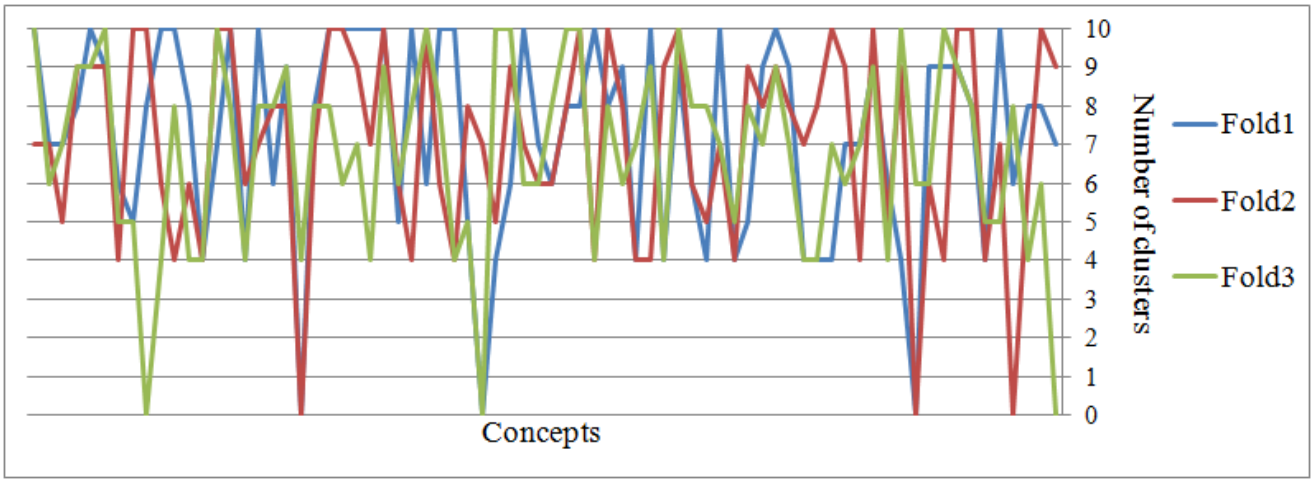


Figure 4. [Number of Clusters Selected for Each Concept Using 3-fold cross-validation]

Evaluation of FCC-MMF Framework

The comparison results between the proposed framework FCC-MMF, CCA-MMF, and the original concatenated feature set are presented in Table 1 in terms of different MAP levels. All the methods are able to reach 70% MAP score at the TOP 5 level, and they all produced around 20% MAP score for

all the instances. Compared to the original data, the improvement rate of the proposed framework has a minimum of 3.37% and at most 14.97%. Comparing to CCA-MMF framework, the improvement of the proposed framework has at least 3.79% and at most 15.98%. The overall MAP results of the proposed framework can be easily distinguished from the other two approaches. The comparison is also visualized in Figure 5 and Figure 6. The comparison results indicate the proposed framework did better in capturing the feature correlation among multiple modalities and grouping features from multiple modalities into higher intra-correlated feature clusters to elevate the classification results. In addition, it is worth mentioning that selecting fusion level in higher granularity, i.e. fusing feature-value pair instead of features, is useful to enhance later classification results. As shown in Figure 6, the proposed FCC-MMF framework performed a competitive result against CCA-MMF and for certain concepts, e.g. horses, food and cow, the percentage difference their MAP scores are even 200% greater which will be further investigate in our future work.

Table 1. Comparisons among FCC-MMF, CCA-MMF, and Original Feature Set in Different MAP Levels

	Top5	**Top10**	**Top20**	**Top50**	**Top100**	**Top150**	**Top200**	**All**
FCC-MMF	74.49%	71.33%	67.05%	61.18%	55.81%	52.76%	50.85%	23.51%
CFA-MMF	70.77%	68.31%	64.51%	58.25%	52.46%	49.20%	47.09%	20.27%
Original Feature Set	71.27%	68.45%	64.26%	58.70%	52.24%	48.82%	46.57%	20.45%
FCC-MMF V.S. Original Feature Set	4.51%	4.20%	4.34%	4.22%	6.82%	8.07%	9.19%	14.97%
FCC-MMF V.S. CFA-MMF	5.25%	4.41%	3.93%	5.03%	6.38%	7.23%	7.97%	15.98%

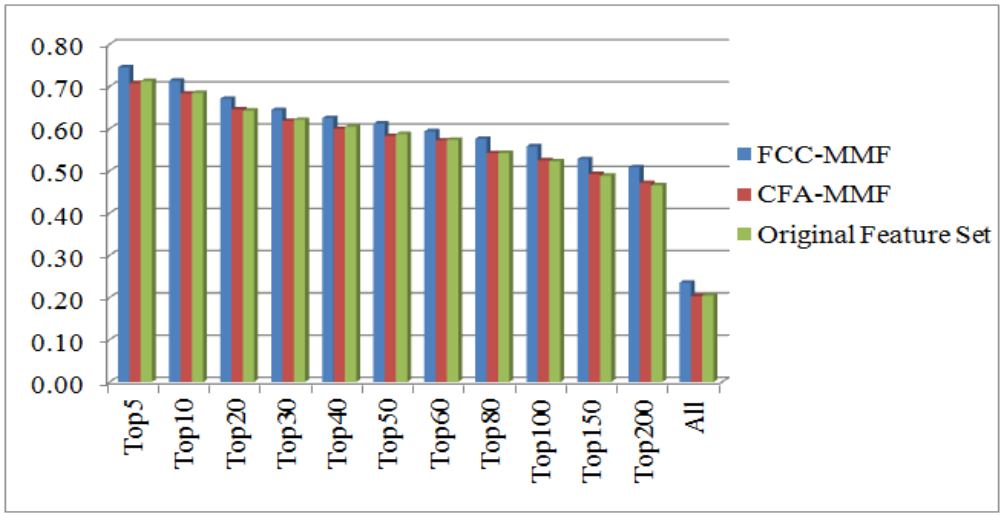


Figure 5. [Comparisons among FCC-MMF, CCA-MMF, and Original Feature Set in Different MAP Levels]

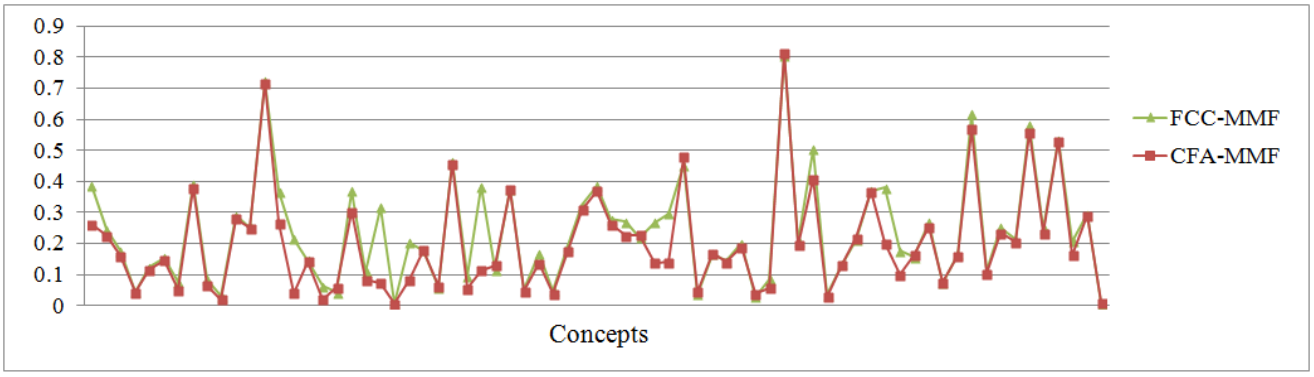


Figure 6. [MAP Improvements in FCC-MMF over CCA-MMF]

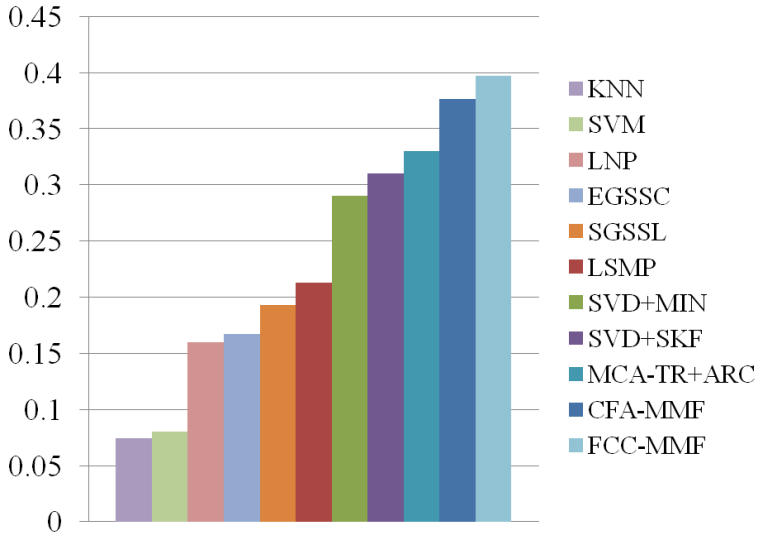


Figure 7. [MAP the proposed framework and other works on the NUS-WIDE Lite data set]

The proposed framework (FCC-MMF) is compared with other research studies investigating semantic retrieval on the NUS-WIDE-LITE data set and evaluated using mean average precision. These related works include K-nearest neighbor (KNN) model (Duda & Hart, 1973), LibSVM model (Witten & Frank, 2005), linear neighborhood propagation (LNP) (Wang & Zhang, 2008), entropic graph semi-supervised classification (EGSSC) (Subramanya & Bilmes, 2009), sparse graph-based semi-supervised learning (SGSSL) (Tang, Yan, Hong, Qi, & Chua, 2009), large-scale multi-label propagation (LSMP) (Chen, Mu, Yan, & Chua, 2010), and three retrieval frameworks, i.e. SVD combined with minimum fusion (SVD+MIN), SVD combined with super kernel fusion (SVD+SKF), multiple correspondence analysis-based tag removal algorithm (MCA-TR+ARC) constructed from Correlation-based Feature Analysis and Multi-Modality Fusion Framework (CFAMMF) (Chen, Zhu, Lin, Shyu, & Chen, 2013). Multiple modalities including visual and textual features are considered in the proposed framework to enhance semantic retrieval. Moreover, MCA-TR method proposed in (Chen, Zhu, Lin, Shyu, & Chen, 2013) is adopted to remove redundant tags information and the threshold of tag removal is decided by training data with the highest MAP. All the parameters in the other works are finely tuned and proved to be best parameter setting in performing semantic retrieval (Chen, Mu, Yan, & Chua, 2010). The purpose of this experiment is to demonstrate that the idea of capturing correlation among modalities in feature-value pair level can be adopted to decompose features from different modalities and transform them into highly intra-correlated feature clusters. Classifiers, which are trained from feature clusters with high intra-correlation, will better detect the semantic concepts. Three-fold cross-validation is used to evaluate the retrieval performs in terms of MAP value. The proposed framework has overall MAP up to almost 40% which can easily distinguished from other works with maximum 34 % MAP differences and at least 3 % MAP difference.

V. CONCLUSION AND FUTURE WORK

The paper presents a novel multimedia semantic retrieval framework that is based on feature correlation clustering. It starts with integrating features from multiple modalities. Multiple correspondence analysis (MCA) is applied to project each feature-value pair as one point onto the two major principal component followed by applying K-medoids algorithm to feature-value pairs from different modalities. The resulted feature-value pair clusters are later transformed into several highly correlated feature clusters. Each feature cluster is trained as a single classifier, and the number of cluster for each concept is determined from the results of 3-fold cross validation on training data that have the highest MAP value. To thoroughly leverage the correlation among different modalities and integrate the strength from different fusion levels, both feature-level fusion and decision-level fusion are applied in the FCC-MMF framework. The experimental results showed that our proposed framework demonstrates promising results against our previous framework CCA-MMF and the original feature set, which simply combines all the features into one classification model. In addition, the proposed framework is also compared to other related works using NUS-WIDE Lite data set to demonstrate that the effectiveness of FCC-MMF in Content-based Multimedia Retrieval is relatively higher than other applications.

Our future work will investigate how to leverage both the characteristics of individual modality and the correlation crossing multiple modalities. Selecting only the useful modalities or feature clusters in detecting semantic concepts could be useful in reducing computation time. In addition, it is also important to legitimately demonstrate the performance in multimedia retrieval by comparing with other related works, which also explore multi-modalities. For example, many researchers explored the cross-modal relationship by applying canonical correlation analysis to better perform in multimedia retrieval (Rasiwasia et al., 2010; Zhang & Liu, 2012), face recognition (Guam, Zhang, Luo, & Lan, 2012), event detection (Younessian, Quinn, Mitamura, & Hauptmann, 2013), etc. In addition, besides

positive correlation, Zhai et al. also pointed out the importance of capturing negative cross-modality correlation since it can provide exclusive information. Therefore, they built a correlation propagation model with two types of links: must-link constrains and cannot-link constrains to model the correlation among modalities. (Zhai, Peng, & Xiao, 2012)

VI. ACKNOWLEDGEMENTS

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Security´s VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

References

Adams, W. H., Iyengar, G., Lin, C. Y., Naphade, M. R., Neti, C., Nock, H. J., & Smith, J. R. (1900). Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP Journal on Advances in Signal Processing*,2003(2), 170-185.

Agarwal, P. K., Matoušek, J., & Suri, S. (1992). Farthest neighbors, maximum spanning trees and related problems in higher dimensions. *Computational Geometry*, 1(4), 189-201.

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345-379.

Ayache, S., Quénot, G., & Gensel, J. (2007). Classifier fusion for SVM-based multimedia semantic indexing. *In Advances in Information Retrieval* (pp. 494-504). Springer Berlin Heidelberg.

Bhanu, B., & Han, J. (2011). Feature Level Fusion of Face and Gait at a Distance. *In Human Recognition at a Distance in Video* (pp. 209-232). *Springer London*.

Bredin, H., & Chollet, G. (2007, April). Audio-visual speech synchrony measure for talking-face identity verification. *In IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*. (Vol. 2, pp. II-233). IEEE.

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., ... & Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell,150(6), 1209-1222.

Candemir, S., Palaniappan, K., Bunyak, F., & Seetharaman, G. (2012, May). Feature fusion using ranking for object tracking in aerial imagery. *In SPIE Defense, Security, and Sensing* (pp. 839604-839604). *International Society for Optics and Photonics.*

Chandrakala, D., & Sumathi, S. (2012). Application of Artificial Bee Colony Optimization Algorithm for Image Classification Using Color and Texture Feature Similarity Fusion. *ISRN Artificial Intelligence, 2012.*

Chen, X., Mu, Y., Yan, S., & Chua, T. S. (2010, October). Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. *In Proceedings of the international conference on Multimedia* (pp. 35-44). ACM.

Chen, C., Zhu, Q., Lin, L., Shyu, M. L., & Chen, S. C. (2013). Web Media Semantic Concept Retrieval via Tag Removal and Model Fusion. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Semantic Adaptive Social Web*.

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009, July). NUS-WIDE: a real-world web image database from National University of Singapore. *In Proceedings of the ACM International Conference on Image and Video Retrieval* (p. 48). ACM.

Clark, G. A., Sengupta, S. K., Sherwood, R. J., Hernandez, J. D., Buhl, M. R., Schaich, P. C., ... & DelGrande, N. (1993, November). Sensor feature fusion for detecting buried objects. *In Optical Engineering and Photonics in Aerospace Sensing* (pp. 178-188). *International Society for Optics and Photonic*s.

Clausi, D. A., & Deng, H. (2005). Design-based texture feature fusion using Gabor filters and co-occurrence probabilities. *IEEE Transactions on Image Processing*, 14(7), 925-936.

Clinchant, S., Ah-Pine, J., & Csurka, G. (2011, April). Semantic combination of textual and visual information in multimedia retrieval. *In Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (p. 44). ACM.

Cui, B., Tung, A. K., Zhang, C., & Zhao, Z. (2010, June). Multiple feature fusion for social media applications. *In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 435-446). ACM.

Dang-Nguyen, D. T., Boato, G., Moschitti, A., & De Natale, F. G. (2012, June). Supervised models for multimodal image retrieval based on visual, semantic and geographic information. *In Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on* (pp. 1-5). *IEEE.*

Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis(Vol. 3). New York: *Wiley.*

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3), 361-365.

Feng, G., Dong, K., Hu, D., & Zhang, D. (2004). When faces are combined with palmprints: a novel biometric fusion strategy. In Biometric authentication (pp. 701-707). *Springer Berlin Heidelberg*.

Fernandez Arguedas, V., Zhang, Q., Chandramouli, K., & Izquierdo, E. (2011, April). Multi-feature fusion for surveillance video indexing. *In International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*,16(10), 906-914.

Gantz, J., chute, c., Manfrediz, a., Minton, s., reinsel, d., schlichting, w., and toncheva, a. (2008) The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011. Idc white Paper, Mar. 2008; http://www.emc.com/about/destination/digital universe/

Gerlach, S., Goetze, S., & Doclo, S. (2012, September). 2D Audio-Visual Localization in Home Environments using a Particle Filter. *In Speech Communication; 10. ITG Symposium; Proceedings of* (pp. 1-4). VDE.

Glodek, M., Scherer, S., & Schwenker, F. (2011). Conditioned hidden markov model fusion for multimodal classification. *In Twelfth Annual Conference of the International Speech Communication Association.*

Greenacre, M., & Blasius, J. (Eds.). (2006). Multiple correspondence analysis and related methods. CRC Press.

Guan, N., Zhang, X., Luo, Z., & Lan, L. (2012, December). Sparse Representation Based Discriminative Canonical Correlation Analysis for Face Recognition. *In Machine Learning and Applications (ICMLA), 2012 11th International Conference on* (Vol. 1, pp. 51-56). IEEE.

Ha, H.-Y. ,Yang, Y., Fleites C., F., and Chen, S.-C. (2013) Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval. *In 2013 IEEE International Conference on International Conference on Multimedia and Expo (ICME)*. IEEE

He, M., Horng, S. J., Fan, P., Run, R. S., Chen, R. J., Lai, J. L., ... & Sentosa, K. O. (2010). Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition*, 43(5), 1789-1800.

Islam, M. R., & Rahman, F. (2010). Hybrid Feature and Decision Fusion Based Audio-Visual Speaker Identification in Challenging Environment. *International Journal of Computer Applications*, 9(5), 9-15.

Jiang, L., Hauptmann, A. G., & Xiang, G. (2012, October). Leveraging high-level and low-level features for multimedia event detection. *In Proceedings of the 20th ACM international conference on Multimedia* (pp. 449-458). ACM.

Jiang, Y. G., Zeng, X., Ye, G., Ellis, D., Chang, S. F., Bhattacharya, S., & Shah, M. (2010, November). Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. *In TRECVID*.

Kalamaras, I., Mademlis, A., Malassiotis, S., & Tzovaras, D. (2013). A novel framework for retrieval and interactive visualization of multimodal data. *Electronic Letters on Computer Vision and Image Analysis*, 12(2), 28-39.

Kludas, J., Bruno, E., & Marchand-Maillet, S. (2008). Information fusion in multimedia information retrieval. In Adaptive Multimedial Retrieval: *Retrieval, User, and Semantics* (pp. 147-159). *Springer Berlin Heidelberg*.

Lan, Z. Z., Bao, L., Yu, S. I., Liu, W., & Hauptmann, A. G. (2012). Double fusion for multimedia event detection. *In Advances in Multimedia Modeling* (pp. 173-185). Springer Berlin Heidelberg.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Lee, T. W. (1998). Independent component analysis (pp. 27-66). *Springer US*.

Li, Z., Wu, Z., Kuang, Z., Chen, K., Gan, Y., & Fan, J. (2013). Evidence-based SVM fusion for 3D model retrieval. *Multimedia Tools and Applications*, 1-19.

Li, M., Zheng, Y. T., Lin, S. X., Zhang, Y. D., & Chua, T. S. (2009). Multimedia evidence fusion for video concept detection via OWA operator. *In Advances in Multimedia Modeling* (pp. 208-216). Springer Berlin Heidelberg.

Lin, L., Ravitz, G., Shyu, M.-L., & Chen, S.-C. (2007). Video semantic concept discovery using multimodal-based association classification. In Proceedings of the IEEE International Conference on Multimedia & Expo (pp. 859-862).

Lin, L., Ravitz, G., Shyu, M.-L., & Chen, S.-C. (2008). Correlation-based video semantic concept detection using multiple correspondence analysis. In Proceedings of the IEEE International Symposium on Multimedia (pp. 316-321).

Lin, L. & Shyu, M.-L. (2009). Effective and efficient video high-level semantic retrieval using associations and correlations. International Journal of Semantic Computing, 3(4), 421–444.

Lin, L., Shyu, M.-L., & Chen, S.-C. (2009). Correlation-based interestingness measure for video semantic concept detection. In Proceedings of the 2009 IEEE International Conference on Information Reuse and Integration (pp. 120-125).

Lin, L. & Shyu, M.-L. (2010a). Weighted association rule mining for video semantic detection. *International Journal of Multimedia Data Engineering and Management*, 1(1), 37–54.

Lin, L. & Shyu, M.-L. (2010b). Correlation-based ranking for large-scale video concept retrieval. *International Journal of Multimedia Data Engineering and Management*, 1(4), 60-74.

Lin, L., Chen, C., Shyu, M.-L., & Chen, S.-C. (2011). Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 18(3), 32–43.

Liu, Y., Zheng, F., Cai, K., & Jiang, B. (2009, December). Cross-Media Retrieval Method Based on Temporal-spatial Clustering and Multimodal Fusion. *In Internet Computing for Science and Engineering (ICICSE), 2009 Fourth International Conference on* (pp. 78-84). IEEE.

Luo, N., Guo, Z., Wu, G., & Song, C. (2012). Multispectral Palmprint Recognition by Feature Level Fusion. *In Recent Advances in Computer Science and Information Engineering* (pp. 427-432). Springer Berlin Heidelberg.

Mangai, U. G., Samanta, S., Das, S., & Chowdhury, P. R. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review*, 27(4), 293.

Mansoorizadeh, M., & Charkari, N. M. (2010). Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2), 277-297.

McDonald, K., Smeaton, A.F. (2005): A comparison of score, rank and probability-based fusion methods for video shot retrieval. *In International Conference on Image and Video Retrieval*, (pp. 61–70). Singapore.

Mertens, R., Lei, H., Gottlieb, L., Friedland, G., & Divakaran, A. (2011, November). Acoustic super models for large scale video event detection. *In Proceedings of the 2011 joint ACM workshop on Modeling and representing events* (pp. 19-24). ACM.

Metallinou, A., Lee, S., & Narayanan, S. (2010, March). Decision level combination of multiple modalities for recognition and analysis of emotional expression. *In 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, (pp. 2462-2465). IEEE.

Motai, Y., Kumar Jha, S., & Kruse, D. (2012). Human tracking from a mobile agent: Optical flow and Kalman filter arbitration. *Signal Processing: Image Communication*, 27(1), 83-95.

Murphy, K. (2001). The bayes net toolbox for matlab. Computing science and statistics, 33(2),

1024-1034.

Nagar, A., Nandakumar, K., & Jain, A. K. (2012). Multibiometric cryptosystems based on feature-level fusion. *In IEEE Transactions on Information Forensics and Security*, 7(1), 255-268.

Natsev, A. P., Naphade, M. R., & TešiĆ, J. (2005, November). Learning the semantics of multimedia queries and concepts from a small number of examples. *In Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 598-607). ACM.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2010, August). Audio-visual classification and fusion of spontaneous affective data in likelihood space. *In Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 3695-3699). IEEE.

Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *In IEEE Transactions on Audio, Speech, and Language Processing*, 17(3), 423-435.

Rashid, U., Niaz, I. A., & Bhatti, M. A. (2010). Fusion of Multimedia Document Intra-Modality Relevancies using Linear Combination Model. *In Advanced Techniques in Computing Sciences and Software Engineering* (pp. 575-580). Springer Netherlands.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010, October). A new approach to cross-modal multimedia retrieval. *In Proceedings of the international conference on Multimedia* (pp. 251-260). ACM.

Reddy, B. S. (2007). Evidential reasoning for multimodal fusion in human computer interaction.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005, July). Speaker independent speech emotion recognition by ensemble classification. *In IEEE International Conference on Multimedia and Expo*, 2005. ICME 2005. IEEE International Conference on (pp. 864-867). IEEE.

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005, November). Early versus late fusion in semantic video analysis. *In Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 399-402). ACM.

Song, M., Chen, C., & You, M. (2004, May). Audio-visual based emotion recognition using tripled hidden Markov model. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*. (Vol. 5, pp. V-877). IEEE.

Subramanya, A., & Bilmes, J. A. (2009). Entropic graph regularization in non-parametric semi-supervised classification. *In Advances in Neural Information Processing Systems* (pp. 1803-1811).

Tang, J., Yan, S., Hong, R., Qi, G. J., & Chua, T. S. (2009, October). Inferring semantic concepts from community-contributed images and noisy tags. *In Proceedings of the 17th ACM international conference on Multimedia* (pp. 223-232). ACM.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 71-86.

Wang, F., & Zhang, C. (2008). Label propagation through linear neighborhoods.Knowledge and Data Engineering, *IEEE Transactions on*, 20(1), 55-67.

Wei, X. Y., Jiang, Y. G., & Ngo, C. W. (2011). Concept-driven multi-modality fusion for video search. Circuits and Systems for Video Technology, IEEE Transactions on, 21(1), 62-73.

Wimmer, M., Schuller, B., Arsic, D., Rigoll, G., & Radig, B. (2008). Low-Level Fusion of Audio, Video Feature for Multi-Modal Emotion Recognition. In VISAPP (2) (pp. 145-151).

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wu, Y., Chang, E. Y., Chang, K. C. C., & Smith, J. R. (2004, October). Optimal multimodal fusion for multimedia data analysis. In Proceedings of the 12th annual ACM international conference on Multimedia (pp. 572-579). ACM.

Xie, Z., & Guan, L. (2012, December). Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis. In 2012 IEEE International Symposium on Multimedia (ISM), (pp. 1-8). IEEE.

Yan, R., Yang, J., Hauptmann, A. (2004). Learning query-class dependent weights in automatic video retrieval. In ACM International Conference on Multimedia, (pp. 548–555). New York, U. S. A.

Yang, M. T., Wang, S. C., & Lin, Y. Y. (2005). A multimodal fusion system for people detection and tracking. International Journal of Imaging Systems and Technology, 15(2), 131-142.

Younessian, E., Quinn, M., Mitamura, T., & Hauptmann, A. (2013, March). Multimedia event detection using visual concept signatures. In IS&T/SPIE Electronic Imaging (pp. 866708-866708). International Society for Optics and Photonics.

Ye, G., Jhuo, I., Liu, D., Jiang, Y. G., Lee, D. T., & Chang, S. F. (2012, June). Joint audio-visual bi-modal codewords for video event detection. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (p. 39). ACM.

Zeng, Z., Tu, J., Pianfetti, B. M., & Huang, T. S. (2008). Audio–visual affective expression recognition through multistream fused HMM. In IEEE Transactions on Multimedia, 10(4), 570-577.

Zeng, Z., Hu, Y., Roisman, G. I., Wen, Z., Fu, Y., & Huang, T. S. (2007). Audio-visual spontaneous emotion recognition. In Artifical Intelligence for Human Computing (pp. 72-90). Springer Berlin Heidelberg.

Zhai, X., Peng, Y., & Xiao, J. (2012, March). Cross-modality correlation propagation for cross-media retrieval. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 2337-2340). IEEE.

Zhang, B. (2011). Multiple features facial image retrieval by spectral regression and fuzzy aggregation approach. International Journal of Intelligent Computing and Cybernetics, 4(4), 420-441.

Zhang, Z. Q., & Wu, J. K. (2011). A novel hierarchical information fusion method for

three-dimensional upper limb motion estimation. In IEEE Transactions on Instrumentation and Measurement, 60(11), 3709-3719.

Zhang, H., & Liu, X. (2012). Cross-Media semantics mining based on sparse canonical correlation analysis and relevance feedback. In Advances in Multimedia Information Processing–PCM 2012 (pp. 759-768). Springer Berlin Heidelberg.

Zhou, X., Depeursinge, A., & Muller, H. (2010, August). Information fusion for combining visual and textual image retrieval. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 1590-1593). IEEE.

Zhu, Q., Lin, L., Shyu, M. L., & Chen, S. C. (2011, August). Effective supervised discretization for classification based on correlation maximization. In 2011 IEEE International Conference on Information Reuse and Integration (IRI),(pp. 390-395). IEEE.

Zhu, Q., Lin, L., Shyu, M. L., & Chen, S. C. (2010, September). Feature selection using correlation and reliability based scoring metric for video semantic detection. In 2010 IEEE Fourth International Conference on Semantic Computing (ICSC),(pp. 462-469). IEEE.

Zou, X., & Bhanu, B. (2005, June). Tracking humans using multi-modal fusion. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops.(pp. 4-4). IEEE.

Correlation-based Re-ranking for Semantic Concept Detection

Hsin-Yu Ha, Fausto C. Fleites, Shu-Ching Chen
School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
{hha001,fflei001,chens}@cs.fiu.edu

Min Chen
Computing and Software Systems School of STEM
University of Washington Bothell, WA 98011, USA
minchen2@u.washington.edu

***Abstract*—Semantic concept detection is among the most important and challenging topics in multimedia research. Its objective is to effectively identify high-level semantic concepts from low-level features for multimedia data analysis and management. In this paper, a novel re-ranking method is proposed based on correlation among concepts to automatically refine detection results and improve detection accuracy. Specifically, multiple correspondence analysis (MCA) is utilized to capture the relationship between a targeted concept and all other semantic concepts. Such relationship is then used as a transaction weight to refine detection ranking scores. To demonstrate its effectiveness in refining semantic concept detection, the proposed re-ranking method is applied to the detection scores of TRECVID 2011 benchmark data set, and its performance is compared with other state-of-the-art re-ranking approaches.**

I. Introduction

With the explosive growth of multimedia applications, the ability to effectively index and retrieve multimedia data becomes increasingly important. Semantic concept detection is widely considered an essential yet challenging step to achieve this goal [1], [2], [3], [4], [5], and has attracted numerous research attentions. One of the typical driven forces is the creation of the TRECVID benchmark by National Institute of Standards and Technology, which aims to boost the researches in semantic media analysis by offering a common video corpus and a common evaluation procedure [6].

Among all the existing work in this area, re-ranking method has been proven effective to improve detection performance when well-designed [7], [8], [9]. Its state-of-the-art process is depicted in Fig. 1. As can be seen, the first step is to preprocess raw multimedia data. Generally, it involves video segmentation, key frame identification, and low-level feature extraction. In the second step, various classification models are trained on training data set and applied to testing data. Then in the third step, the results from different classification models are fused with ranking scores indicating how likely semantic concepts can be detected from each testing instance. Finally, re-ranking process is performed using auxiliary information such as concept ontology to refine final ranking scores. For example, in [7], [8], Concept Association Network (CAN) is used for re-ranking, which captures strong associations among different concepts based on association rule mining (ARM). In [9], co-occurrence among semantic concepts is used to enhance the re-ranking process.

In this work, we propose to leverage the implication among semantic concepts in the re-ranking process. For example,

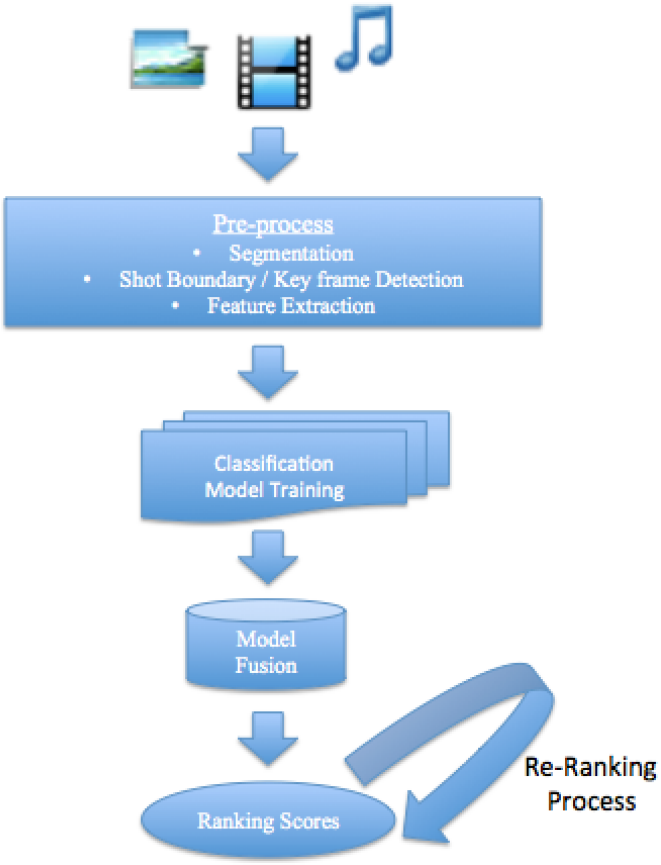


Fig. 1: A general semantic concept detection process

concept "forest" may help indicate the presence of concepts "outdoor" and "plants" instead of "telephone." In other words, concept correlation is used in our re-ranking method to refine ranking scores produced by other classification models.

The remainder of the paper is organized as follows. Section II discusses the details about the proposed framework. Section III presents the experimental settings and performance evaluation, and Section IV concludes this paper.

II. Proposed Framework

The proposed framework aims to automatically refine classification ranking scores for semantic concept detection. As

such ranking scores may be produced by any state-of-the-art classification models and may possibly be continuous data, we first apply a discretization process to convert them into nominal values. Then the concept correlation component is applied to discover how closely two concepts are semantically associated followed by the refined ranking process. The proposed framework is shown in Fig. 2.

A. Ranking Score Discretization

Given a training data set with N instances (i.e., images, video key frames, etc.) and M high-level semantic concepts (such as classroom, airplane, etc.), various classification models are trained and in the process each training instance is associated with a ranking score toward a concept. An example is shown in Table I. As can be seen from the table, the ranking score of training instance 1 toward concept 1 is -1.49 while that of instance 2 is -0.97 toward concept 1. We then use a supervised discretization method called minimum description length (MDL) to discretize the ranking scores into several intervals for each concept and correspondingly we define concept-value pair as follows:

Definition 1. *A **concept-value pair** C_j^i represents the j^{th} ranking score interval of the i^{th} concept, where $1 \leq i \leq M$ and the range of j is determined by the discretization results.*

TABLE I: Concept Ranking Scores

	Concept 1 Ranking Score	Concept 2 Ranking Score	...	Concept M Ranking Score
Instance 1	-1.49	1.08	...	-0.45
Instance 2	-0.97	-0.85	...	-1.32
...	...	...	...	...
Instance N	-0.48	-0.97	...	-1.01

For example, assume the range of ranking scores for concept 1 is discretized into three intervals. They are then denoted as three concept-value pairs: C_1^1, C_2^1, and C_3^1, respectively. Note a concept will be eliminated from further processing if it has only one concept-value pair as it fails to differentiate among instances. Once MDL is applied to all the M concepts, two options are provided to construct an indicator matrix I for a target concept.

- Option 1. Combine concept-value pairs of a single concept (e.g., Concept 1) with the ground truth information of a target concept (see example in Table II.
- Option 2. Combine concept value pairs of all concepts (i.e., Concept 1, Concept 2, ..., Concept M) with the ground truth information of a target concept (see example in TABLE III.

In both cases, number 1 or 0 represents true or false. In other words, the entry for instance 1 in C_1^1 is 1 that means instance 1's initial ranking score (-1.49 as in TABLE I) falls into C_1^1. Similarly, the entry for instance 1 in Column "Target Concept Positive" is 0 that means instance 1 is not labeled with the target concept.

TABLE II: Indicator Matrix of the Concept Ranking Scores for Single Concept

	Concept 1			Target Concept Positive	Target Concept Negative
	C_1^1	C_2^1	C_3^1		
Instance 1	1	0	0	0	1
Instance 2	0		0	1	0
...	...	...	...	...	...
Instance N	0	0	1	1	0

B. Concept Correlation

Multiple correspondence analysis (MCA) has been proven to perform well on many research topics, such as feature selection [10], discretization [11], video semantic concept detection [12], [13], [14], [15], and data pruning [16], which motivates us to apply it in capturing concept correlations for re-ranking process.

Specifically, with indicator matrix I produced in the previous step, a Burt matrix B is constructed as $I^T I$. The sum of all elements in matrix B, denoted as gt is then obtained using Equation 1, where L is number of columns in matrix I.

$$gt = \sum_{i=0}^{L} \sum_{j=0}^{L} B_{ij}; \tag{1}$$

Thus a normalized Burt matrix NB can be constructed as shown in Equation 2.

$$NB = B/gt; \tag{2}$$

Let $row = \{row_i, i = 1, 2, ...L\}$ and $col = \{col_j, j = 1, 2, ...L\}$ where $row_i = \sum_j NB_{ij}$ and $col_j = \sum_i NBij$, respectively, a centralized matrix Z can be generated following Equation 3.

$$Z = D_{row}^{-1/2}(NB - row * col^T)D_{col}^{-1/2}; \tag{3}$$

Here D_{row} and D_{col} are the diagonal matrices for row and col, respectively. With the application of Single Value Decomposition (SVD), we can then derive eigenvectors from the centralized matrix Z. Because more than 95% of the total variance can be captured by the top two principal components, a subspace is constructed using two eigenvectors with the largest eigenvalues, to which the concept-value pairs in matrix I are projected. Fig. 3 shows an example result using option 1 (i.e., only one concept's concept-value pairs are used), where Pos and Neg represents the positions of positive and negative classes, PC_1 and PC_2 are the x-axis and y-axis corresponding to the top two principal components. In contrast, an example result of option 2 is shown in Fig. 4, where concept-value pairs for all the concepts are used to build the indicator matrix I.

Two parameters are then proposed to represent the concept-value pair's correlation toward the target concept: similarity α and reliability β, which are defined as follows.

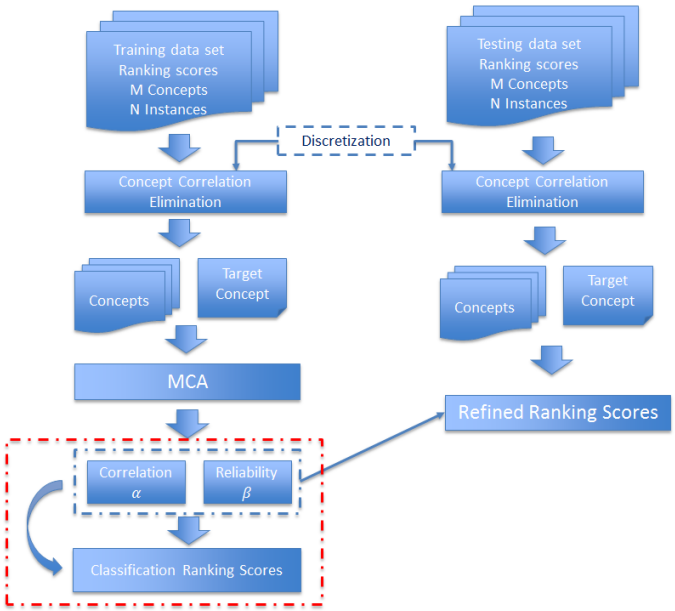


Fig. 2: Proposed Re-Ranking Framework

TABLE III: Indicator Matrix of the Concept Ranking Scores for All Concepts

	Concept 1			Concept 2		...	Concept M		Target Concept Positive	Target Concept Negative
	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	...	C_1^M	C_2^M		
Instance 1	1	0	0	1	0	...	1	0	0	1
Instance 2	0	1	0	0	1	...	0	1	1	0
...	...	...	...	...	...	...	...	...	...	...
Instance N	0	0	1	0	1	...	1	0	1	0

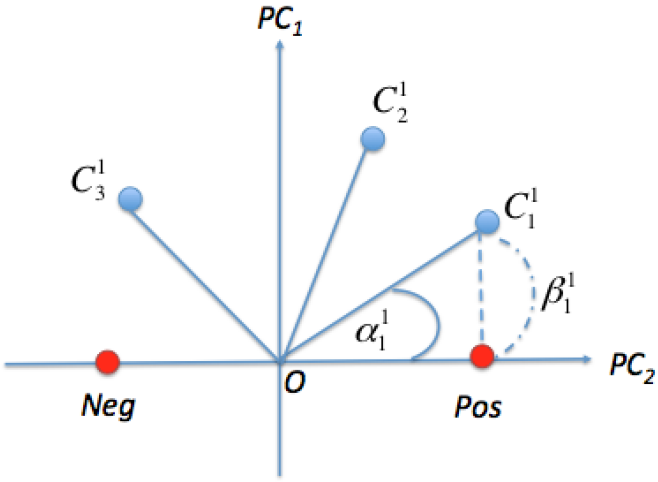


Fig. 3: Projected concept-value pairs using option 1

Definition 2. *Concept Similarity: This is the cosine value of the angle between a concept-value pair C_j^i and the positive class P_OS. Mathematically, it is computed in Equation 4.*

$$\alpha_j^i = \frac{\vec{C_j^i} \cdot \vec{Pos}}{\left|\vec{C_j^i}\right| \left|\vec{Pos}\right|}; \tag{4}$$

Definition 3. *Concept Reliability: This is the Euclidean distance between a concept-value pair C_j^i and the positive class P_OS, which is denoted as β_j^i. It is proposed to take into consideration the dependence between a concept-value pair and the positive class.*

For example, the correlation between concept-value pair C_1^1 and the target concept are captured by α_j^i and β_j^i, as marked in Fig. 3

C. Ranking Score Refinement

To refine the ranking scores, similarity α and reliability β obtained above are used to compute the transaction weight as follows.

Definition 4. *$Transaction Weight per concept-value pair$: It indicates the correlation between this concept-value pair and the target semantic concept, and is computed in Equation 5. The higher the transaction weight of a concept-value pair C_j^i is, the more likely a target concept will be detected from*

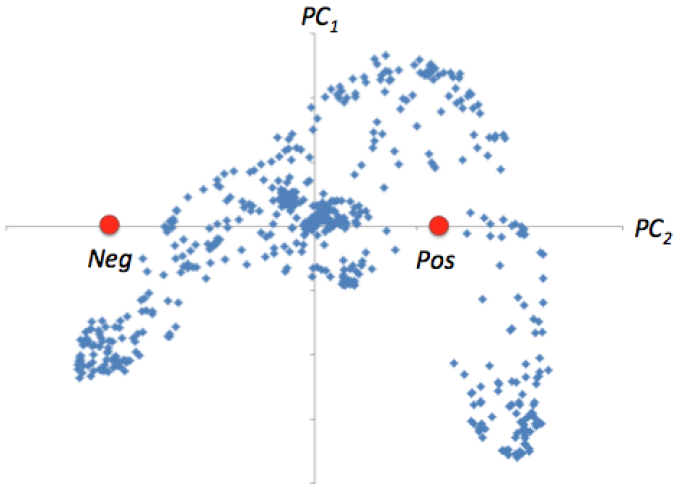


Fig. 4: Projected concept-value pairs using option 2

a testing instance when its ranking score for concept i is discretized into C_j^i.

$$TW_j^i = \alpha_j^i w_j^i + \beta_j^i (1 - w_j^i) \tag{5}$$

Here, w_j^i is a weighting factor. The value of a weighting factor has the range between 0 and 1 with an increment of 0.2, i.e., $w_j^i \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The weighting factor corresponding to the highest MAP on the training set is automatically selected for the testing set.

Definition 5. *$TransactionWeightperinstance$: This is considered as the refined ranking score for the testing instances, which is basically the accumulation of the transaction weight for each concept-value pair as depicted in Equation 6.*

$$TransactionWeight_k = \sum_{i=0}^{M} TW_j^i \tag{6}$$

Here, M is the total number of concepts, k represents the index of the testing instance, and j indicates the index of the concept-value pair for concept i.

III. Experiments

In this section, the performance of the proposed re-ranking method is tested and compared with raw ranking scores (i.e., no re-ranking framework is applied so it is called "Baseline") as well as that of three re-ranking frameworks: "Aytar" [17], "DASD" [18], and "AAN" [19]. In Aytar et al., the correlation of a related concept to the target concept is represented by conditional probability and it is further leveraged to enhance the overall detection performance. In Jiang et al., "DASD" is proposed to model the correlation among the concepts as symmetric links with the weights obtained from training labels. Thus, a graphic model is formed to address the potential domain change problem. In Meng et al., "AAN" is built applying association rule mining method to capture the strong association among different concepts and later is used to refine the ranking scores. All three methods focus on applying concept correlation to better improve the re-ranking results without requirng domain knowledge. The performance is evaluated using Mean Average Precision (MAP), which is defined as the arithmetic mean of per-concept average precision and is commonly adopted to evaluate the effectiveness of semantic concept detection.

TABLE IV: Statistics of data set IACC.1.B

Dataset	IACC.1.B
TRECVID Year	2011
No. Concepts	346
No. Instances	137327
Average P/N Ratio	0.003
Average Pos No.	408.32

A. Experimental Setup

In this paper, IACC.1.B data set from TRECVID 2011 is used as testbed. Some basic statistics of the data set are depicted in TABLE IV

The detection scores produced by the Shinoda Lab at Tokyo Institute of Technology on this data set is used as baseline because it performed the best in TRECVID 2011 Semantic Indexing Task. In addition, three-fold cross validation is adopted and the performance is reported by averaging the MAPs obtained from three rounds of classification results.

B. Experimental Results

TABLE V shows the comparison results for all the 130 concepts in terms of MAP value. Different MAP value were calculated based on the numbers of retrieved instances, meaning Top10 MAP represents the MAP value of the top 10 retrieved instances after sorting the ranking scores in descending order. The last column "Overall" shows the MAP value of all the retrieved instances. The higher the MAP value is, the better the semantic concept detection performance are. The rows "Single Concept" and "All Concepts" indicate Option 1 and Option 2 of our proposed re-ranking method, whose performances are compared to "Baseline" (MAP value of the original ranking scores without any re-ranking process), "Aytar", "DASD", and "AAN". The rows "IR_1," "IR_2," ..., "IR_5" show the improvement rates between our "Single Concept" option and the other five methods: Baseline, Aytar, DASD, AAN, and "All Concept" option. Specifically, it is defined in Equation 7.

$$IR_i = \frac{(SingleConcept'sMAP - i^{th}Method'sMAP)}{i^{th}Method'sMAP} \tag{7}$$

and $i = 1, 2, ...5$. For example, IR_1 at TOP10 is computed as $(0.5677 - 0.5218)$ / $0.5218 = 4.59\%$.

As can be seen, Option 1 of our proposed method (i.e., "Single Concept") consistently improves the raw ranking scores and outperforms all the re-ranking methods across all different retrieved levels in terms of MAP value, as also depicted in Fig. 5. This clearly shows the effectiveness of

TABLE V: The MAP values of 130 concepts in IACC.1.B for different number of retrieved instances with the improvement rate of the proposed method using single concept against other re-ranking methods

MAP Retrieved Level	TOP10	TOP20	TOP40	TOP60	TOP80	TOP100	Overall
Baseline	0.5218	0.4898	0.4481	0.4212	0.3999	0.3845	0.1382
Aytar	0.4600	0.4304	0.4075	0.3925	0.3806	0.3654	0.1363
DASD	0.4637	0.4561	0.4240	0.4063	0.3903	0.3743	0.1397
AAN	0.5491	0.5143	0.4709	0.4428	0.4211	0.4051	0.1452
All Concepts	0.5154	0.4722	0.4256	0.3999	0.3801	0.3642	0.1665
Single Concept	0.5677	0.5349	0.4881	0.4658	0.4431	0.4207	0.1881
Improvement R1	4.59%	4.51%	4.00%	4.46%	4.32%	3.62%	4.99%
Improvement R2	10.77%	10.45%	8.06%	7.33%	6.25%	5.13%	5.18%
Improvement R3	10.40%	7.88%	6.41%	5.95%	5.28%	4.64%	4.84%
Improvement R4	1.86%	2.06%	1.72%	2.30%	2.20%	1.56%	4.29%
Improvement R5	5.23%	6.27%	6.25%	6.59%	6.30%	5.65%	2.16%

our proposed method in discovering the correlation between concepts and in using such correlation to help re-rank detection scores. On the other hand, Option 2 (i.e., "All Concepts") does not produce comparative results against all other methods except for "Aytar." It shows that when considering all the concept at the same time, the correlation might be affected by some irrelevant concepts. Nevertheless, as we can see, its overall MAP value is better than the other three re-ranking frameworks. It indicates it in fact greatly helps data elements with low classification scores (i.e., they do not appear in the TopK) to obtain correct class labels, which may be beneficial in some scenarios.

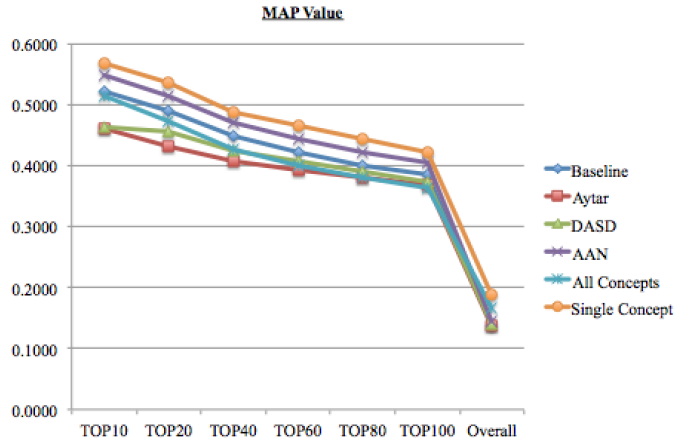


Fig. 5: The MAP values of 130 concepts in IACC.1.B for different number of retrieved instances using one concept against other re-ranking methods

IV. Conclusions

The paper proposes a re-ranking framework that utilizes concept correlation to automatically refine ranking scores for semantic concept detection. Specifically, multiple correspondence analysis (MCA) is applied to capture concept correlation. In the process, two parameters $similarity$ and $reliability$ are modeled to compute transaction weight which is then translated as the refined ranking score. In the experiments, the ranking scores of TRECVID 2011 IACC.1.B data set is used as baseline and the performance of the proposed methods is compared with that of three state-of-the-art re-ranking frameworks in terms of how well the ranking scores are refined. It shows that Option 1 of our proposed method outperforms other other re-ranking methods at all the retrieval levels while Option 2 is more suitable where all instances need to be retrieved.

In the future, we will first carefully identify the strength of the proposed two options. For example, how well each of them can handle concept with only few positive instances. Then, we will study the possibility of combining results from these two options to further enhance the final results. We will also work on negative correlation and identify concept subsets with higher correlation toward the target concept.

Acknowledgment

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

References

[1] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, Lin Luo, and Min Chen, "Detection of soccer goal shots using joint multimedia features and classification rules," *MDM/KDD*, vol. 3, 2003.

[2] Shu-Ching Chen, Mei-Ling Shyu, Min Chen, and Chengcui Zhang, "A decision tree-based multimodal data mining framework for soccer goal detection," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. IEEE, 2004, vol. 1, pp. 265–268.

[3] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and Min Chen, "A multimodal data mining framework for soccer goal detection based on decision tree logic," *International Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312–323, 2006.

[4] Min Chen, Shu-Ching Chen, Mei-Ling Shyu, and Kasun Wickramaratna, "Semantic event detection via multimodal data mining," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 38–46, 2006.

[5] Mei-Ling Shyu, Zongxing Xie, Min Chen, and Shu-Ching Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *Multimedia, IEEE Transactions on*, vol. 10, no. 2, pp. 252–259, 2008.

[6] Alan F Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 2006, pp. 321–330.

[7] Tao Meng and Mei-Ling Shyu, "Leveraging concept association network for multimedia rare concept mining and retrieval," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE, 2012, pp. 860–865.

[8] Tao Meng and Mei-Ling Shyu, "Concept-concept association information integration and multi-model collaboration for multimedia semantic concept detection," *Information Systems Frontiers*, pp. 1–13, 2013.

[9] Chao Chen, Lin Lin, and Mei-Ling Shyu, "Utilization of co-occurrence relationships between semantic concepts in re-ranking for information retrieval," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 53–60.

[10] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, 2010, pp. 462–469.

[11] Qiusha Zhu, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Effective supervised discretization for classification based on correlation maximization," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. IEEE, 2011, pp. 390–395.

[12] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Association rule mining with a correlation–based interestingness measure for video semantic concept detection," *International Journal of Information and Decision Sciences*, vol. 4, no. 2, pp. 199–216, 2012.

[13] Lin Lin and Mei-Ling Shyu, "Effective and efficient video high-level semantic retrieval using associations and correlations," *International Journal of Semantic Computing*, vol. 3, no. 04, pp. 421–444, 2009.

[14] Lin Lin and Mei-Ling Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 1, no. 1, pp. 37–54, 2010.

[15] Lin Lin, Guy Ravitz, Mei-Ling Shyu, and Shu-Ching Chen, "Correlation-based video semantic concept detection using multiple correspondence analysis," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. IEEE, 2008, pp. 316–321.

[16] Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*. IEEE, 2009, pp. 304–311.

[17] Yusuf Aytar, Omer Bilal Orhan, and Mubarak Shah, "Improving semantic concept detection and retrieval using contextual estimates," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 536–539.

[18] Yu-Gang Jiang, Jun Wang, Shih-Fu Chang, and Chong-Wah Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1420–1427.

[19] Tao Meng, "Association affinity network based multi-model collaboration for multimedia big data management and retrieval," 2013.

Expert Systems with Applications 41 (2014) 4844–4866

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Exploring the diversity in cluster ensemble generation: Random sampling and random projection

Fan Yang [a], Xuan Li [a], Qianmu Li [b], Tao Li [c,*]

[a] School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[b] School of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China
[c] School of Computer Science, Florida International University, Miami, FL 33199, USA

ARTICLE INFO

Keywords:
Random sampling
Random projection
Ensemble generation
Ensemble clustering

ABSTRACT

Cluster ensemble first generates a large library of different clustering solutions and then combines them into a more accurate consensus clustering. It is commonly accepted that for cluster ensemble to work well the member partitions should be different from each other, and meanwhile the quality of each partition should remain at an acceptable level. Many different strategies have been used to generate different base partitions for cluster ensemble. Similar to ensemble classification, many studies have been focusing on generating different partitions of the original dataset, i.e., clustering on different subsets (e.g., obtained using random sampling) or clustering in different feature spaces (e.g., obtained using random projection). However, little attention has been paid to the diversity and quality of the partitions generated using these two approaches. In this paper, we propose a novel cluster generation method based on random sampling, which uses the nearest neighbor method to fill the category information of the missing samples (abbreviated as RS-NN). We evaluate its performance in comparison with k-means ensemble, a typical random projection method (Random Feature Subset, abbreviated as FS), and another random sampling method (Random Sampling based on Nearest Centroid, abbreviated as RS-NC). Experimental results indicate that the FS method always generates more diverse partitions while RS-NC method generates high-quality partitions. Our proposed method, RS-NN, generates base partitions with a good balance between the quality and the diversity and achieves significant improvement over alternative methods. Furthermore, to introduce more diversity, we propose a dual random sampling method which combines RS-NN and FS methods. The proposed method can achieve higher diversity with good quality on most datasets.

1. Introduction

Clustering for unsupervised data exploration and analysis has been investigated for decades in the statistics, data mining, and machine learning fields. A fundamental challenge in clustering is that different clustering results can be obtained using different clustering algorithms. Cluster ensembles address this issue by first generating a large set of clustering results and then combining them using a consensus function to create a final clustering solution that is considered to encompass all of the information contained in the ensemble. This is the basic philosophy behind cluster ensemble (Strehl & Ghosh, 2002), which has gained increasing popularity in the clustering community (Fern & Brodley, 2003, 2004; Greene, Tsymbal, Bolshakova, & Cunningham, 2004; Hadjitodorov, Kuncheva, & Todorova, 2006; Kuncheva & Hadjitodorov, 2004; Topchy, Jain, & Punch, 2003, 2004, 2005; Zheng, Li, & Ding, 2010) in the last decade. Recently, Yu et al. studied the structure ensemble for the fusion of different structures (Yu, You, Wong, & Han, 2012a). Many researchers have studied the ensemble generation methods while others focus on consensus functions for combining different partitions. In this paper, we focus on the ensemble generation methods.

Existing research on cluster ensemble has suggested that the diversity among ensemble members plays an important role in the success of cluster ensembles. In practice, a cluster ensemble with a high diversity can be obtained in many different ways. Multiple clustering algorithms, different representations of the data, and different parameter choices can all be used to produce a diverse set of clustering solutions. The quality of individual base partition is another important ingredient in cluster ensemble that should be taken into consideration.

* Corresponding author. Tel.: +1 305 348 6036; fax: +1 305 348 3549.
E-mail address: taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.eswa.2014.01.028
0957-4174/

Similar to ensemble classification, many studies have been conducted on generating different partitions of the original dataset (Fern & Brodley, 2003; Fischer & Buhmann, 2003; Fred & Jain, 2002; Ilc & Dobnikar, 2012; Minaei-Bidgoli, Topchy, & Punch, 2004; Strehl & Ghosh, 2002; Topchy et al., 2003; Yu, Wong, You, Yu, & Han, 2012b), e.g., clustering on different subsets (e.g., obtained using random sampling), and clustering in different feature spaces (e.g., obtained using random projection). However, little attention has been paid to the comparison of the performances in terms of the diversity and the quality of the partitions generated using these two different approaches. In this study, we choose the Random Feature Subset method (FS) and Random Sampling Method based on Nearest Centroid (RS-NC) as representatives of these two manners, and conduct experiments on real-world UCI datasets using these two methods.

We also propose a new cluster generation method based on random sampling which uses the nearest neighbor method to find the category information of the missing examples. Specifically, the clustering combination via consensus functions is conducted on multiple labeling of different subsamples of a given dataset. We first generate a diverse set of subsets of the original dataset by random sampling with replacement, and then run k-means on these subsets with randomly initialized cluster centers. For the missing instances which have not been selected, we find their nearest neighbors and use the neighbors' category information to fill the labels of these missing instances. The ensembles are then aggregated into a single final partition using a consensus function. In the experiment, we observe that this ensemble generation method achieves higher diversity and better performance than some existing methods on UCI datasets.

We note that the proposed method and random feature subset method have comparable performance in the experiments. Inspired by random forest, we further propose a dual random sampling method which combines the random sampling method based on nearest neighbor (RS-NN) and the Random Feature Subset (FS) method. The two sampling methods are integrated together and higher diversity and good quality are obtained. The experiments demonstrate the effectiveness of our proposed method.

The rest of this paper is organized as follows. In Section 2, we review the background and related studies in the ensemble clustering community. Different clustering ensemble techniques are discussed, as well as the generation mechanisms and the consensus functions. Section 3 briefly introduces the evaluation criteria of cluster ensemble and the definition of quality and diversity of ensemble members. The traditional methods and the proposed RS-NN method used in comparison are presented in Section 4 and the comparison experiments on 17 UCI datasets are conducted in Section 5. Section 6 introduces the dual random sampling method and shows its performance. Finally, we summarize our contributions and conclude the paper in Section 7.

2. Related works and motivation

2.1. Related works

The basic idea of combining different clustering solutions to obtain improved clustering has been explored under different ways such as consensus classification/clustering (Monti, Tamayo, Mesirov, & Golub, 2003; Neumann & Norton, 1986) and evidence accumulation (Fred & Jain, 2002). Every clustering ensemble method is made up of two steps: Clustering Generation and Clustering Consensus. Below we review these two basic steps in clustering ensembles and some recent advances on cluster ensemble design.

It is commonly accepted that for cluster ensembles to work well the member partitions should be different from each other. Many approaches have been proposed to generate different multiple clustering solutions from a give dataset:

- Using different clustering algorithms to produce the initial partitions such as density based clustering, k-means or fuzzy c-means, etc. (Strehl & Ghosh, 2002).
- Changing initialization or other parameters of a clustering algorithm (Fred & Jain, 2002; Topchy et al., 2003) such as the number of clusters and random initializations for clustering.
- Projecting data onto different subspaces (Fern & Brodley, 2003; Topchy et al., 2003) such as projection to 1-dimension or random cuts/plane.
- Choosing different features (Topchy et al., 2003) to represent the objects. For example, images can be represented by their pixels, histograms, location and parameters of perceptual primitives or 3D scene coordinates.
- Partitioning different subsets of the original data (Fischer & Buhmann, 2003; Minaei-Bidgoli et al., 2004) with replacement or not.
- Applying transformations to the data matrix to by different kinds of transformation operators (Yu et al., 2012b).

Once a set of initial partitions are generated, a consensus function is used to combine them and produce a final partition. Recently, many consensus functions have been proposed for clustering ensemble, including:

- The Relabeling and Voting methods (Dudoit & Fridlyand, 2003; Fischer & Buhmann, 2003) (V-M, PV, CV and VAC). They first address the labeling correspondence problem and then create the consensus partition using a voting scheme.
- Spectral clustering approaches (Yu, Li, You, Wong, & Han, 2012c). For example, Yu et al. propose two clustering frameworks, i.e., triple spectral clustering-based and double spectral clustering-based consensus clustering approaches.
- Co-association Matrix (Fred & Jain, 2002; Monti et al., 2003) (EA-SL, EA-CL, CTS and PA). The idea of co-association is used to avoid the label correspondence problem. Co-association methods (Fred & Jain, 2005), map the partitions in the cluster ensemble into an intermediate representation: the co-association matrix.
- Graph and Hypergraph partitioning (Strehl & Ghosh, 2002) (CSPA, HGPA, MCLA and HBGF). It transforms the combination problem into a graph or hypergraph partitioning problem. The difference among these methods lies on the way the (hyper) graph is built from the set of clustering and how the cuts on the graph are defined in order to obtain the consensus partition.
- Finite Mixture Models based methods (CE-EM) Topchy et al., 2004. The consensus partition is obtained as the solution of a maximum likelihood estimation problem which can be solved using the EM algorithm.
- Locally Adaptive Clustering Algorithms (Domeniconi et al., 2007). This type of consensus function combines partitions obtained using locally adaptive clustering algorithms (LAC).
- Information-theoretic methods (e.g., Quadratic Mutual Information) (Topchy et al., 2003) and Fuzzy techniques (Xu & Wunsch, 2005).

The aforementioned methods are all based on objects co-occurrence. In addition, several approaches based on Median Partition such as Mirkin distance (SAOM, BOM, BOK, FURTH, etc.) (Filkov & Skiena, 2004; Wakabayashi, 1998), Genetic Algorithms (IT-GA, HCE and MM-GA) (Yoon, Ahn, Lee, Cho, & Kim, 2006a; Yoon, Lee, Cho, & Kim, 2006b), Kernel methods (WKF, WPCK Vega-Pons, Correa-Morris, & Ruiz-Shulcloper, 2010 and GWKF), and Non-Negative Matrix Factorization (Li, Ding, & Jordan, 2007a) (NMFC

and WC) have been developed to discover the final clustering solution from many multiple partitions. Recently, methods have also developed to ensemble both hierarchical and partitional clusterings (Zheng et al., 2010). For a comprehensive survey on clustering ensemble, please refer to (Vega-Pons & Ruiz-Shulcloper, 2011). Different from existing works, our work mainly focuses on ensemble generation, an often-neglected issue in ensemble clustering. We aim to answer the following research question regarding ensemble generation: given two different strategies (ensemble generation with random sampling and random projection), which one will create a more diverse library of clustering solutions? We conduct extensive empirical studies and propose two new generation methods to improve the existing ensemble clustering methods.

2.2. Motivation

The main motivation of this work is two-fold. First, little attention has been paid to the comparison of their influences on the degree of diversity and quality in the partitions generated by random sampling and random projection. We wonder which one has the best performance and conduct an intuitive and preliminary comparison on several real-world UCI datasets. Further, we wonder whether these two manners can benefit each other in generating cluster ensembles.

Another motivation is inspired by the difference between nearest neighbor method (NN) and nearest centroid method (NC). The NN method is sensitive to the local distribution of data points, while the NC method is a typical kind of prototype based learning method and is more robust to the local distribution of data and noise or outliers in the data. Intuitively speaking, the prediction of the label of an instance is unstable when we assigned it the label of its nearest neighbor, while it could be more robust when we assigned the same label of its nearest centroid to this instance.

Consider a binary classification problem. Let (X, Y) be a pair of random variables, and $(X, Y) \in R^d \times \{0, 1\}$. Given training examples $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ from the distribution of (X, Y), the class centroid μ_c of class C is computed as follows,

$$\mu_c = \frac{1}{|C|}\sum_{x_i \in C} x_i, \tag{1}$$

where $|C|$ denotes the cardinality of set C. The label of a test instance x_t is then predicted as,

$$y_t = \arg \min_{j \in (0,1)} \|x_t - \mu_j\|. \tag{2}$$

In the nearest neighbor method, the label of x_t is predicted as the label of its nearest neighbor x_{ns} and,

$$x_{ns} = \arg \min_{x_i \in S} \|x_t - x_i\|. \tag{3}$$

Fig. 1 illustrates the difference between NN and NC methods in clustering ensemble when using k-means algorithm with random sampling. For each clustering run, it selects a random sample from the original data set, and then clusters the subsample by k-means. When assigning each instance absent from the current subsample, Fern and Brodley use the label of the closest cluster according to its Euclidean distance to the cluster centers, namely nearest centroids (NC) (Fern & Brodley, 2004). For comparison we propose to use NN method instead.

In Fig. 1(1), the dataset S is clustered into two classes C_1 and C_2 except two instances a and b which are not selected in the run of k-means clustering. In this case, both NC and NN method will give the same results when assigning the labels of a and b.

In Fig. 1(2), two instances c and d are not selected in the run of k-means clustering, and we use NN to predict their labels, and in Fig. 1(3), NC is used for prediction. Obviously, Fig. 1(3) and (1) have the same clustering results, although different subsamples of S are used in the two cases. Note that in Fig. 1(2) the cluster result is significantly different from that in Fig. 1(1) and (3). We speculate that the NN method will introduce more diversity in clustering ensemble with random sampling compared with NC.

An important reason is that NC method assigns each instance absent from the current subsample to the closest cluster according to its distances to cluster centroids, this operation is similar to the mechanism of k-means, so it may not significantly change the position of clustering centers by adding a small amount of examples, and may have no significant differences compared with k-means. As shown in Fig. 1(3), c and d do not have significant influence on the results of k-means.

3. Definition: quality and diversity

A cluster ensemble system solves the clustering problem in two steps. The first step takes a dataset as the input and outputs a set of clustering solutions. The second step takes the cluster ensemble as the input and combines the solutions to produce a single clustering as the final output. Formally, given a dataset $X = \{X_1, X_2, \ldots, X_n\}$, a cluster ensemble is a set of clustering solutions, represented as $C = \{C^1, C^2, \ldots, C^R\}$, where R is the ensemble size, i.e., the number of clustering solutions in the ensemble. Each clustering solution C^r is simply a partitioning of the dataset X with K_r disjoint instance groups, represented as $C^r = \{C_1^r, C_2^r, \ldots, C_{K_r}^r\}$, where $\cup_k C_k^r = X$. Generally speaking, the value of K_r for different clustering runs can be either the same or different. To build our clustering library, we use the k-means algorithm (MacQueen, 1967), which is one of the most widely used clustering algorithms and have been used in many previous cluster ensemble studies. The standard k-means algorithm was carried out with k from 2 to 10 with randomly initialized cluster centers, in order to obtain complex structure in the

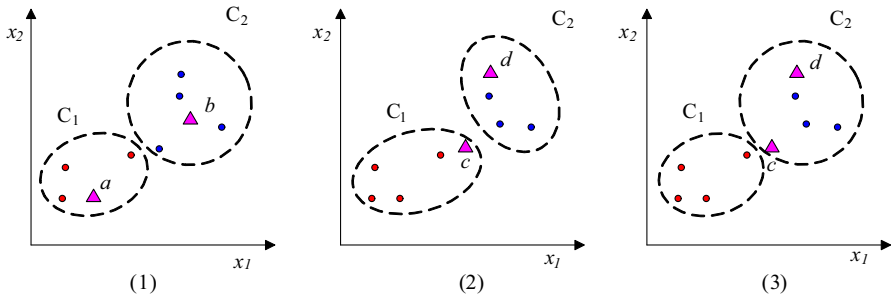


Fig. 1. An illustrative example showing the difference between NN and NC in cluster ensemble. The dataset S is divided into 2 categories, namely C_1 and C_2 by k-means. In the first case, two data points a and b are not selected during the k-means clustering, NN and NC give the same results (of assigning labels to them) as shown in (1); In the second case, two data points c and d are not selected during clustering. The result of using NN method to predict their labels is shown in (2), and the result of NC method is shown in (3).

consensus partition from the combination of small hyper-spherical structures in individual partitions.

Recent studies have shown that the quality and diversity are two key factors that influence the performance of cluster ensemble. In this paper, the quality and diversity are also used to measure clustering results. In this section, we first explain how we measure the quality and diversity of the ensemble, and then introduce the evaluation criteria of the final clustering results.

3.1. Quality

In clustering literature, it is common to measure the quality of a clustering solution based on how well it recovers the class labels. Here we use the widely used Normalized Mutual Information (*NMI*) introduced by Strehl and Ghosh (2002), which defined the consensus partition as a partition which shares the most information with all partitions in the cluster ensemble.

Formally, given an ensemble $C = \{C^1, C^2, \ldots, C^R\}$, let $C^a = \{C_1^a, C_2^a, \ldots, C_{K_a}^a\}$ and $C^b = \{C_1^b, C_2^b, \ldots, C_{K_b}^b\}$ be two partitions of X, K_a being the number of clusters in C^a and K_b being the number of clusters in C^b. Let n_{ia} be the number of objects in the ith cluster of the partition C^a, and n_{bj} be the number of objects in the jth cluster of the partition C^b. And n_{ij} represents the number of objects which are together in the ith cluster of the partition C^a and in the jth cluster of the partition C^b. The *NMI* between C^a and $C^b$is expressed as,

$$NMI(C_a, C_b) = \frac{-2\sum_{i=1}^{K_a}\sum_{j=1}^{K_b} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}\cdot n}{n_{ia}\cdot n_{bj}}\right)}{\sum_{i=1}^{K_a} \frac{n_{ia}}{n} \log\left(\frac{n_{ia}}{n}\right) + \sum_{j=1}^{K_b} \frac{n_{bj}}{n} \log\left(\frac{n_{bj}}{n}\right)}. \quad (4)$$

It takes 1 as a maximum value and 0 as a minimum.

Assume the clustering solution formed by the real labels is C^*, a good consensus clustering should maximize the following criterion,

$$\text{Quality} = \frac{1}{R}\sum_{i=1}^{R} NMI(C^i, C^*). \quad (5)$$

The above equation represents the overall quality of the solutions, where $NMI(C^i, C^*)$ is the normalized mutual information between the clustering C^i and the clustering C^*obtained from the real classes or formed by the real labels. If two clustering solutions define the same partitions for the dataset, then the *NMI* value is maximized to be 1. In contrast, if two clustering solutions define completely independent partitions, the *NMI* value is minimized to be 0. Here we refer to this objective function as the average of $NMI(C^i, C^*)$, so *Quality* is also between 0 and 1. Intuitively, an ensemble solution maximizing *Quality* maximizes the information it shares with the real class, thus the higher *Quality* is, the higher quality the solutions have.

3.2. Diversity

Existing research revealed that the diversity among the ensemble members plays an important role towards achieving improved clustering performance. The literature has suggested that higher diversity among ensemble members produces higher performance gain. Here we use the measure introduced by Fern and Brodley (2003) as the metric of the diversity, which is based on pair-wise normalized mutual information among clustering solutions.

$$Diversity = \frac{1}{R \times R}\sum_{i=1}^{R}\sum_{j=1}^{R} NMI(C^i, C^j). \quad (6)$$

This formula represents the overall diversity of the ensemble, where $NMI(C^i, C^j)$ is the normalized mutual information between clustering C^i and clustering C^j in the ensemble. The value of *Diversity* is also between 0 and 1. The lower the value is, the higher the diversity is. Obviously, an ensemble minimizing *Diversity* maximizes the diversity between the solutions in the ensemble.

It is worth noting that the quality and diversity introduced in Sections 3.1 and 3.2 are measures for the library of solutions (i.e., the input to the consensus function) and the evaluation criteria below are used to evaluate the performance of the final ensemble results (i.e., consensus partition).

Note that the "borderline" instances, i.e., instanced with "uncertain" cluster memberships, can be one source of diversity as they can easily change clustering memberships with different runs of the same clustering algorithm. However, there are many other reasons and sources accounting for the diversity of clustering solutions such as different random initializations, different subsets of datasets, different feature subspaces, different similarity/distance measurements, and different clustering algorithms with different objective functions. In this paper, we deals with ensemble generation with random sampling and random projection and investigates which one of these two strategies will create a more diverse library of clustering solutions.

3.3. Evaluation criteria of the quality of the ensemble

Using a specific ensemble generation method, we can obtain the ensemble solutions as well as the final data partition or consensus partition P^* with a user-defined consensus function. To evaluate the final performance of the ensemble, we use the real class labels as a substitute for the true underlying structure of the data.

Accuracy discovers the one-to-one relationship between the clustering results (i.e., clusters) and the real class labels (i.e., classes) and measures the extent to which each cluster contained data objects from the corresponding class. It sums up the total matching degree between all pairs of clusters and classes (Li, Ding, & Jordan, 2007b; Zhu, Wang, & Li, 2010). It can be represented as:

$$\text{Accuracy} = \text{Max}\left(\sum_{C_k, L_m^*} T(C_k, L_m^*)\right) \Big/ N, \quad (7)$$

where C_k expresses the k-th cluster, and L_m^* is the m-th class. $T(C_k, L_m^*)$ is the number of objects in class m which are assigned to cluster k. Accuracy computes the maximum sum of $T(C_k, L_m^*)$ for all pairs of clusters and classes, and there is no overlap among these pairs. It is easy to know that the greater the accuracy is, the better the performance is.In addition, we choose another widely used criterion to measure the clustering performance. Purity is a simple and transparent evaluation measure. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N. Formally,

$$\text{purity}\,(p^1, p^2) = \frac{1}{N}\sum_{k} \max_j |w_k \cap c_j|, \quad (8)$$

where $p^1 = \{w_1, w_2, \ldots, w_k\}$ is the set of clusters and $p^2 = \{c_1, c_2, \ldots, c_j\}$ is the set of classes. We interpret w_k as the set of documents in w_k and c_j as the set of documents in c_j in Eq. (8). Purity is an external evaluation criterion for cluster quality. Bad clustering has purity value close to 0 and a perfect clustering has a purity of 1. High purity is easy to achieve when the number of clusters is large, and in particular, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters.

Another measure that allows us to make this tradeoff is normalized mutual information or NMI introduced in Section 3.1. We measure the *NMI* between the final consensus clustering and the

real clustering. Generally, the higher the *NMI* value is, the better the quality of the ensemble result is.

4. Ensemble clustering method

4.1. Ensemble generation method

Many approaches have been proposed to generate different multiple clustering solutions from a give data set. Below, we introduce several traditional methods used in this paper and propose a novel ensemble generation method.

4.1.1. K-means algorithm (KM)

Different clustering solutions are obtained by running k-means to the same dataset with randomly initialized cluster centers. In this method, the clustering algorithm has access to all the features and the variations among clustering runs only come from different initializations. We randomly choose the number of clusters k between 2 and 10 in order to obtain complex structure in the data. The clustering solutions obtained in this setting are expected to have relatively better quality but less diversity (Fern & Lin, 2008).

4.1.2. Random feature subset method (FS)

This is a simple random projection method. In this method, different clustering solutions are obtained by using different random feature subsets of the original data. Note that for each clustering run, we perform random sampling with replacement on the original feature set. This process will produce duplicate features, but we use the repetitive features only once in forming the new datasets. Then we apply k-means to all the new datasets to obtain the clustering solutions.

4.1.3. Random sampling method based on nearest centroid (RS-NC)

Random sampling method (Fern & Brodley, 2004) selects a random sample from the original data set with a sampling rate of 70% in each clustering run. Then the subsamples are clustered by k-means and each instance absent from the current subsample is assigned to its closest cluster based on its Euclidean distance to the cluster centers to ensure that all the instances are clustered in each clustering run. Since there is little discussion in existing literatures on the sampling rate, we perform random sampling with replacement in order to facilitate the comparison with our proposed method.

4.1.4. Random sampling method based on nearest neighbor (RS-NN)

As discussed in Section 2.2, RS-NC method just assigns each instance absent from the current subsample to its closest cluster based on its Euclidean distance to the cluster centers, and may have no significant differences compared with k-means. It will not significantly change the position of clustering center by adding a small amount of samples.

In order to introduce more diversity between partitions, we propose RS-NN instead. Similar to the RS-NC method, for each clustering run, it randomly samples the original data set with replacement. Then the subsamples are clustered by k-means and the label of each instance absent from the current subsample is assigned to that of its nearest neighbor to ensure that all the instances are clustered in each clustering run. Experiments show that this ensemble generation method can obtain the best performance in terms of the quality and diversity.

We will present Random Sub-sampling based on nearest neighbor method (RS-NN) in detail in Section 4.3.

4.2. Consensus function

Once a cluster ensemble is obtained, we need a consensus function to combine the solutions to produce a final consensus clustering. In this paper, we mainly focus on the ensemble generation method. In order to eliminate the variance of different consensus functions, we take three popular approaches to combine the solutions in our study, including the spectral clustering approach, the Cluster-based Similarity Partitioning Algorithm (CSPA) and the HyperGraph Partitioning Algorithm (HGPA) method proposed by Strehl and Ghosh (2002). Experiments in Section 5 show that all three approaches generate similar results.

4.2.1. Spectral clustering approach

Compared with traditional clustering method, spectral clustering approach only needs the similarity matrix of the data. Spectral clustering is more robust than the traditional clustering algorithms, and it is less sensitive to the irregular or error data. In addition, its computation complexity is relatively low, especially in high-dimensional data such as text data or ordinary image data. Besides spectral clustering, Yu et al. propose two clustering frameworks, i.e., triple spectral clustering-based and double spectral clustering-based consensus clustering approaches (Yu et al., 2012c). The main idea is to incorporate the spectral clustering algorithm into the ensemble framework to generate based clusterings on the original datasets (including both attribute and objects) and also to serve as the consensus function to obtain the final result.

4.2.2. Cluster-based similarity partitioning algorithm (CSPA)

A clustering signifies a relationship between objects in the same cluster: two objects have a similarity of 1 if they are in the same cluster; otherwise, their similarity is 0. Given the input clusterings, an $n \times n$ pairwise similarity matrix (the co-association matrix) can thus be constructed. This can be seen as the adjacency matrix of a fully connected graph where the nodes are the samples of X and an edge between two samples has a correlative weight equal to the frequency that the samples have been clustered into the same cluster. After that, the graph partitioning algorithm METIS (Karypis & Kumar, 1998) is used for obtaining the consensus partition because of its robust and scalable properties.

CSPA is simple and heuristic, but its computation and storage complexities are both quadratic in n, as opposed to the next two approaches that are near linear in n (Strehl & Ghosh, 2002).

Table 1
Characteristics of datasets used in experiment.

Datasets	Instances	Features	Classes
Iris	150	4	3
wine	178	13	3
Soybean	47	35	4
glass	214	9	6
segmentation	2310	19	7
hearts	267	44	2
Thyroid	215	5	3
ionosphere	351	34	2
WDBC	569	30	2
CHART	600	60	6
SPECTFheart	267	44	2
SPECTheart	267	22	2
Lung	32	56	3
heart	462	9	2
bank	4521	16	2
Pima	768	8	2
sat	6435	36	6

4.2.3. HyperGraph partitioning algorithm (HGPA)

This is a direct approach to partition the hypergraph by cutting a minimal number of hyperedges (each hyperedge represents a cluster of an input clustering). All hyperedges have the same weight. Similarly, all vertices are equally weighted. And it is searched by cutting the minimum possible number of hyperedges that partition the hypergraph to k connected components of approximately the same size. Note that the hypergraphs partitioning package HMETIS (Karypis, Aggarwal, Kumar, & Shekhar, 1999) is used. HMETIS gives high-quality partitions and is very scalable (Strehl & Ghosh, 2002).

4.3. Random sampling method based on nearest neighbor (RS-NN)

Here we will describe our method in detail. First we build an average similarity matrix ($n \times n$ dimension, where n is the number of the samples in the dataset) based on the ensemble solutions $C = \{C^1, C^2, \ldots, C^R\}$obtained by the ensemble generation method. This average similarity matrix measures the frequency of each sample pair being clustered together in the ensemble, which is also referred to as the co-association matrix. As we described in Section 3, each clustering solution C^r is simply a partition of the dataset X into K_r disjoint clusters of instances, represented as $C^r = \{C_1^r, C_2^r, \ldots, C_{K_r}^r\}$, where $\cup_k C_k^r = X$. The similarity value between two samples x and y in the similarity matrix^i is defined as follows,

$$\text{Similarity matrix } i(x,y) = \begin{cases} 1, & \text{if } x,y \in C_{K_i}^i \\ 0 & \text{Otherwise} \end{cases}, \tag{9}$$

$$\text{The average similarity matrix} = \frac{1}{R}\sum_{i=1}^{R} \text{similarity matrix}^i. \tag{10}$$

We then apply spectral clustering/CSPA/HGPA to the average similarity matrix to obtain a final partition of the data points into c parts, where c is the number of real classes in the data.

The detailed algorithm, named as random sampling method based on nearest neighbor (RS-NN), is summarized in Algorithm 1. To avoid introducing more parameters (i.e., the sample size) in the algorithm, we use the method of sampling with replacement and remove duplicate examples.

Algorithm 1. RS-NN

Input:
S: The original dataset (the number of cases is N)
R: Predefined ensemble size
K: The range of k values
For i = 1 to R
(i) Get a subsample $s(i)$ by drawing N examples at random but with replacement from S and remove the duplicate examples
(ii) Apply k-means (the value of k is randomly selected from K) to $s(i)$
(iii) Use nearest neighbor method to label the absent examples in the clustering solutions
(iv) Construct similarity matrix(i) of the ensemble solutions
End
Compute the average similarity matrix
Apply spectral clustering/CSPA/HGPA to the average similarity matrix to obtain a final partition P^*
Output: The final consensus partition

Note that we will get RS-NC method if we use nearest centroid method to label the absent examples in Algorithm 1.

5. Experiments and discussion

5.1. Datasets and experiment setting

In this paper, the two metrics *Quality* (described in Eq. (5)) and *Diversity* (described in Eq. (6)) are used to compare the perfor-

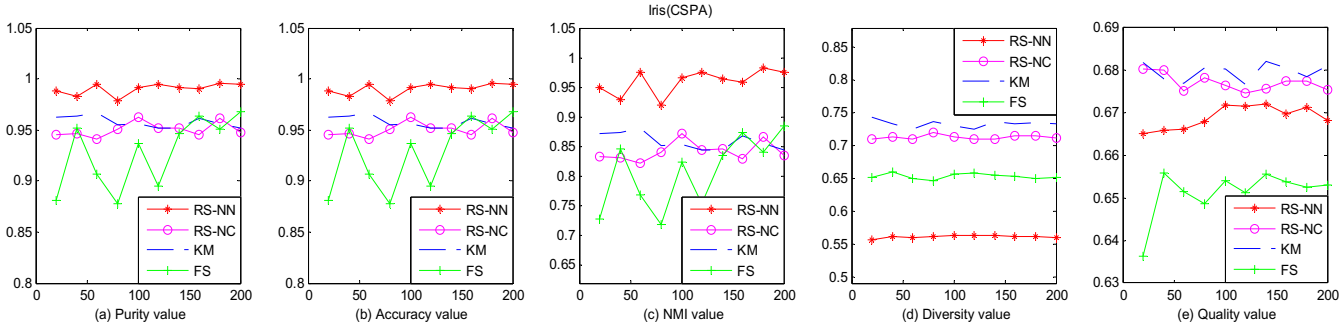


Fig. 2. Performance comparison on Iris.

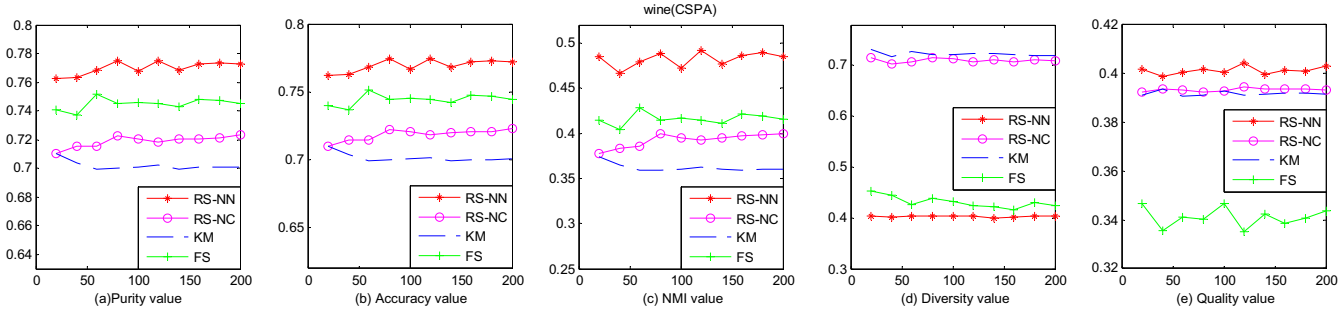


Fig. 3. Performance comparison on wine.

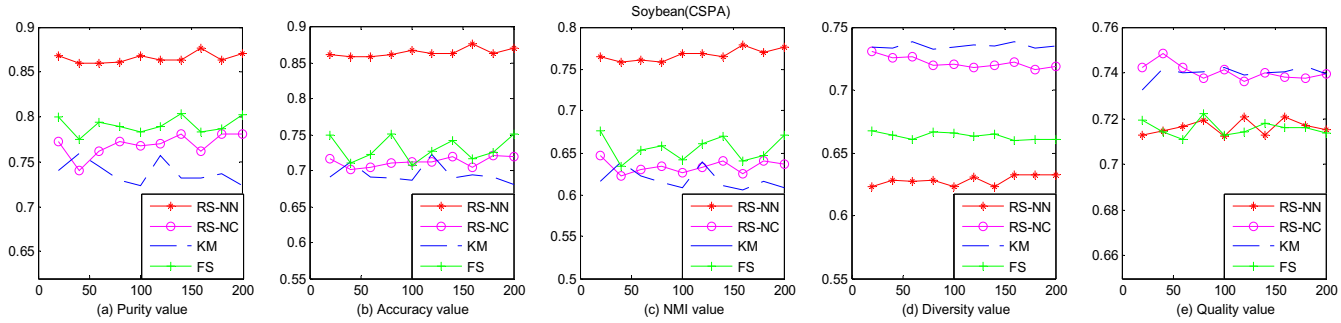


Fig. 4. Performance comparison on Soybean.

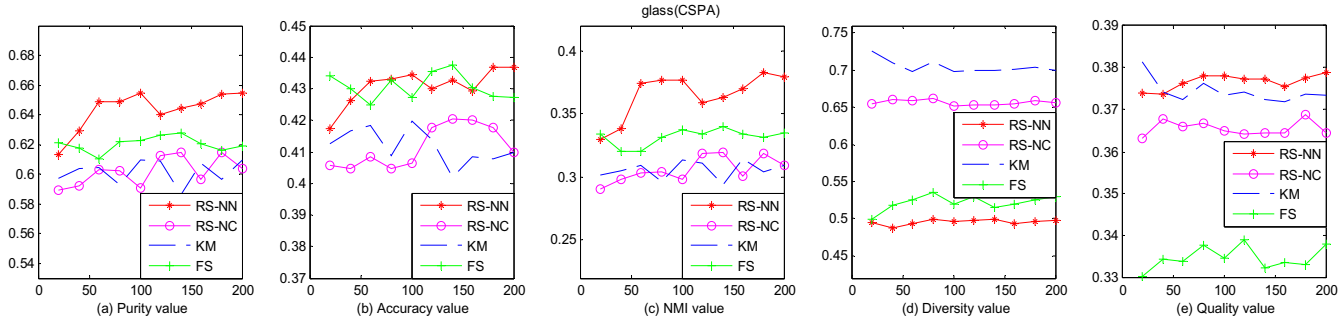


Fig. 5. Performance comparison on glass.

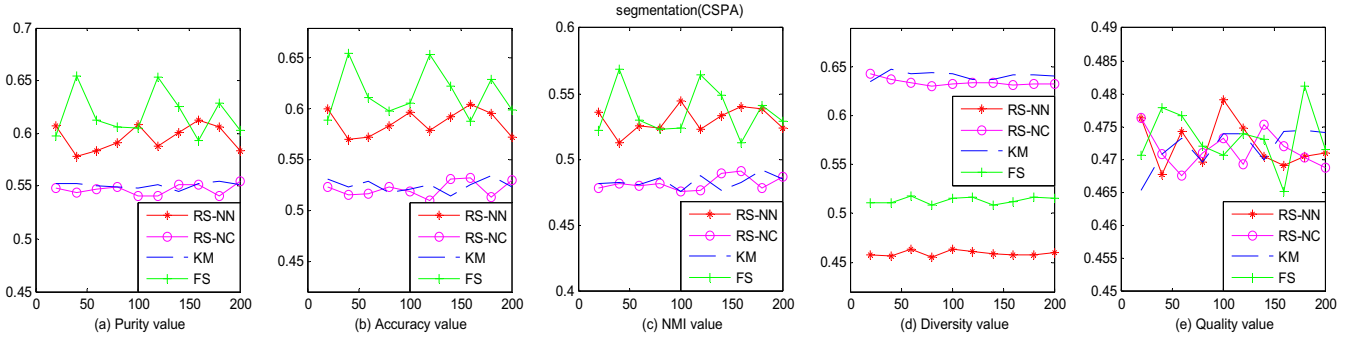


Fig. 6. Performance comparison on segmentation.

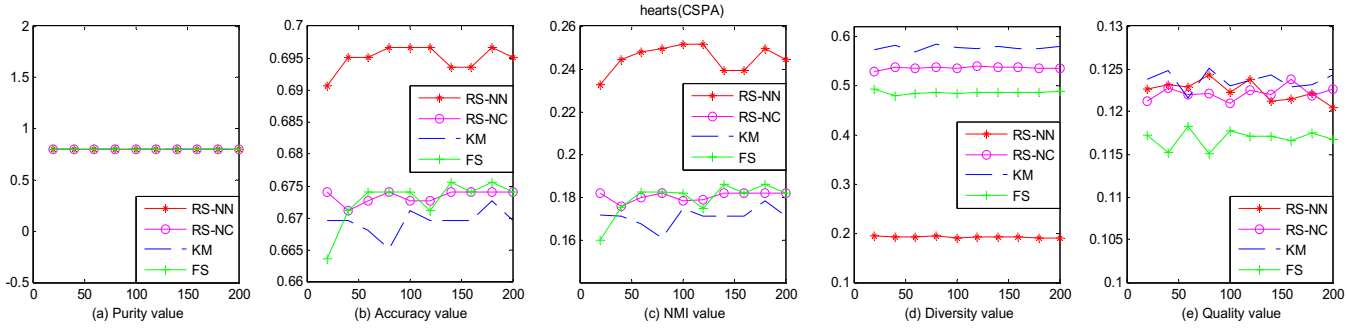


Fig. 7. Performance comparison on hearts.

mance of four different ensemble generation methods, and the values of Purity, Accuracy and $NMI(P*, C*)$ are used to evaluate the final clustering results. 17 UCI datasets are used in the experiment, as shown in Table 1.

For every ensemble generation method we build our clustering library. Each of the above four method is used to generate [20, 40, 60, 80, 100, 120, 140, 160, 180, 200] clustering solutions. In particular, for each clustering run we set the number of clusters, k, by randomly drawing a number between 2 and 10. For the consensus function, we apply Spectral Clustering/CSPA/HGPA to the average similarity matrix to obtain a final partition of the data points into c parts, where c is the number of real classes.

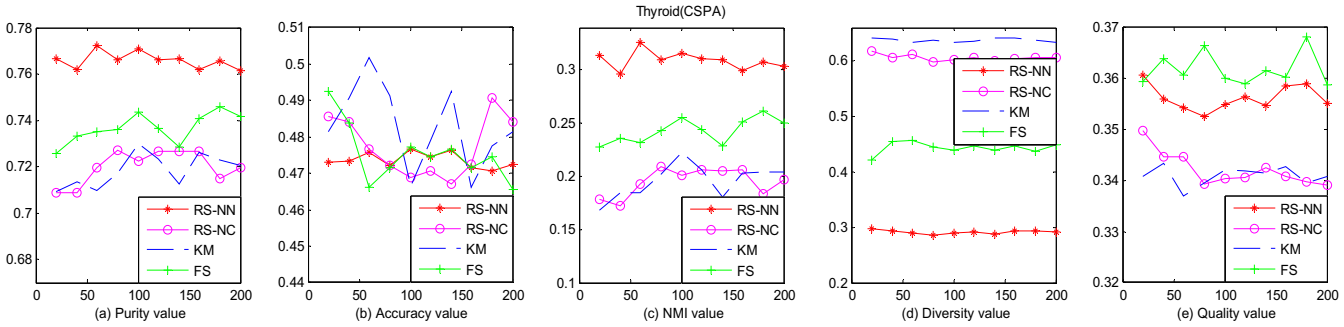


Fig. 8. Performance comparison on Thyroid.

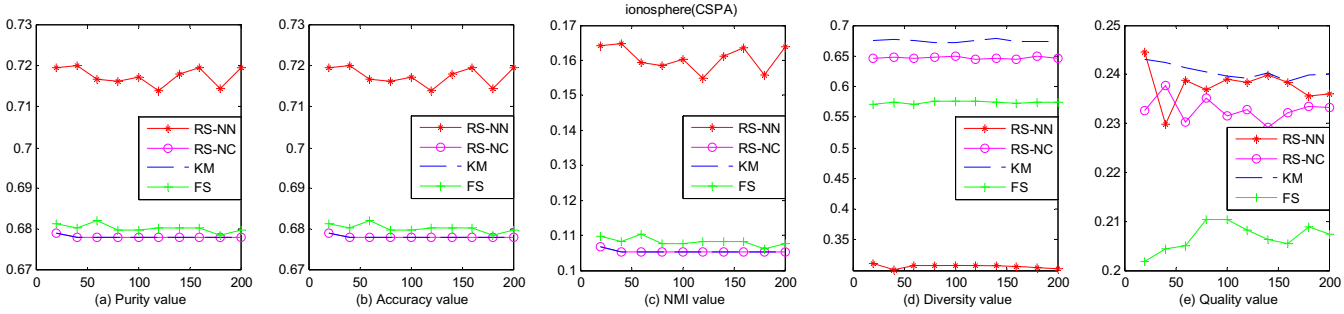


Fig. 9. Performance comparison on ionosphere.

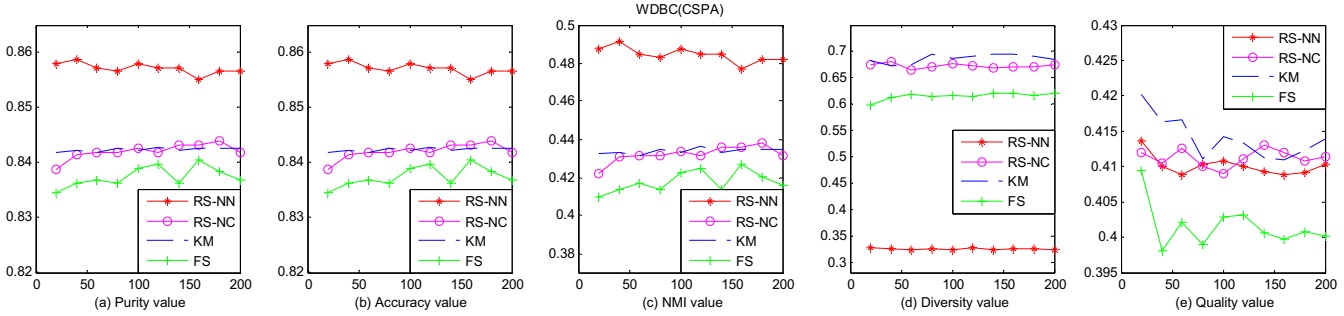


Fig. 10. Performance comparison on WDBC.

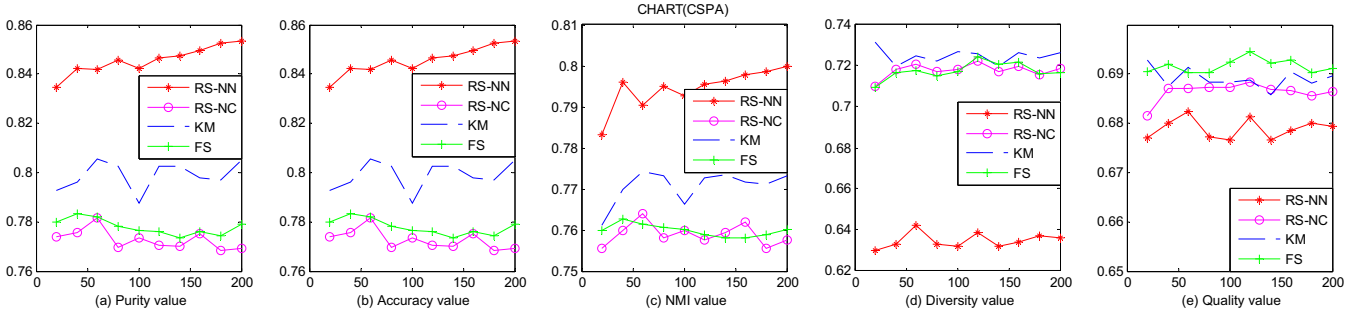


Fig. 11. Performance comparison on CHART.

For each dataset, we repeat this process ten times to generate ten libraries and the reported results are averaged over these ten runs. All the experiments are conducted using Matlab 2007b platform on a desktop machine (with a processor of 3.1 GHz Dual-Core).

5.2. Experiment results

Figs. 2(a–c)–18(a–c) show the performance comparison of every ensemble generation method varying with the ensemble sizes on seventeen datasets using CSPA as consensus function. The

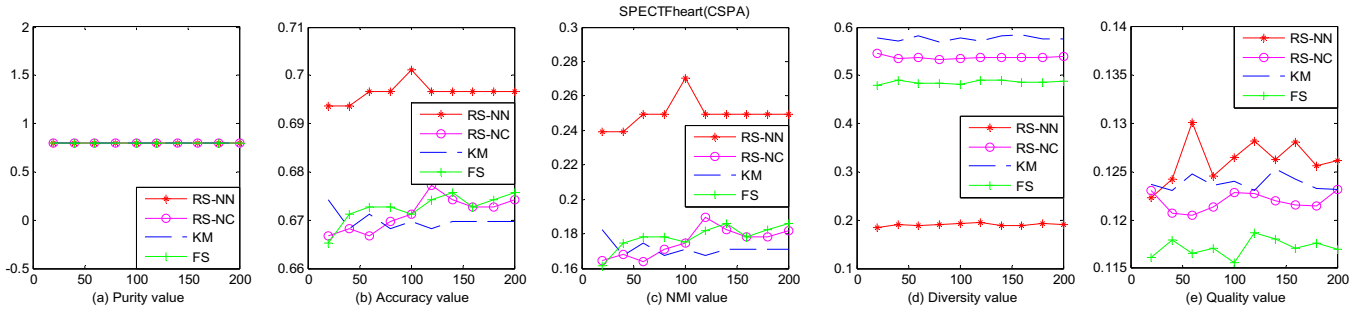


Fig. 12. Performance comparison on SPECTFheart.

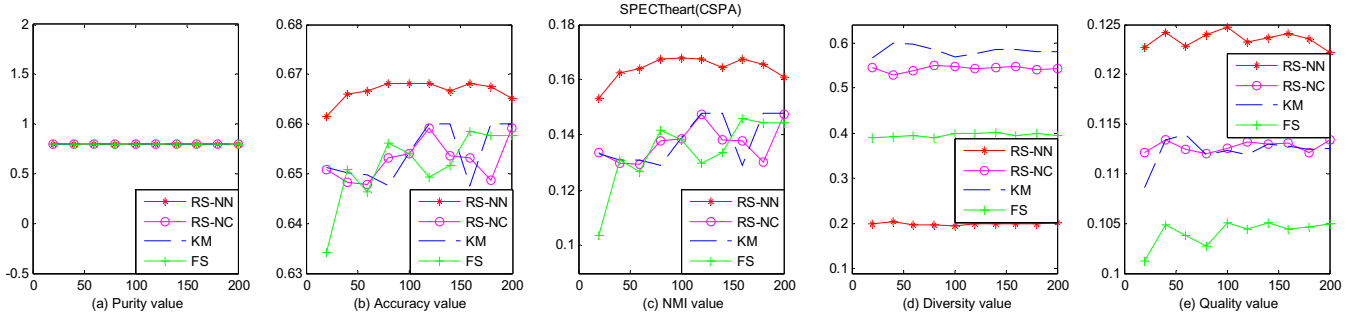


Fig. 13. Performance comparison on SPECTheart.

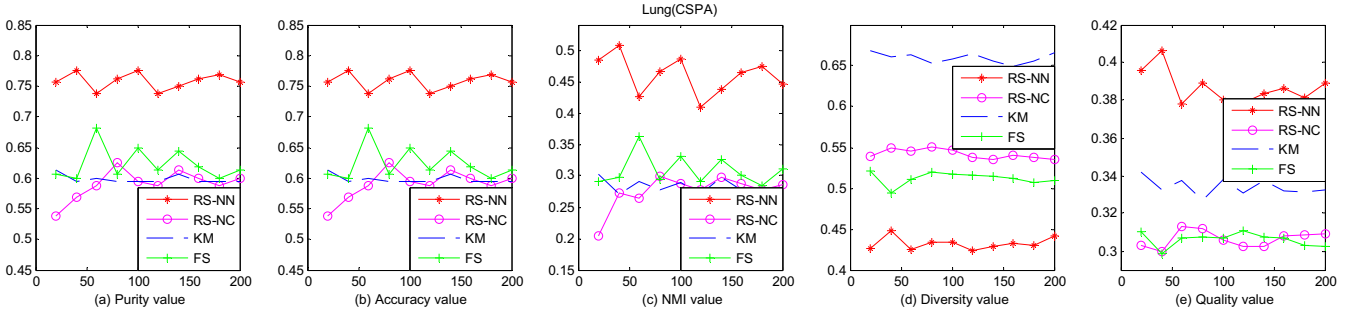


Fig. 14. Performance comparison on Lung.

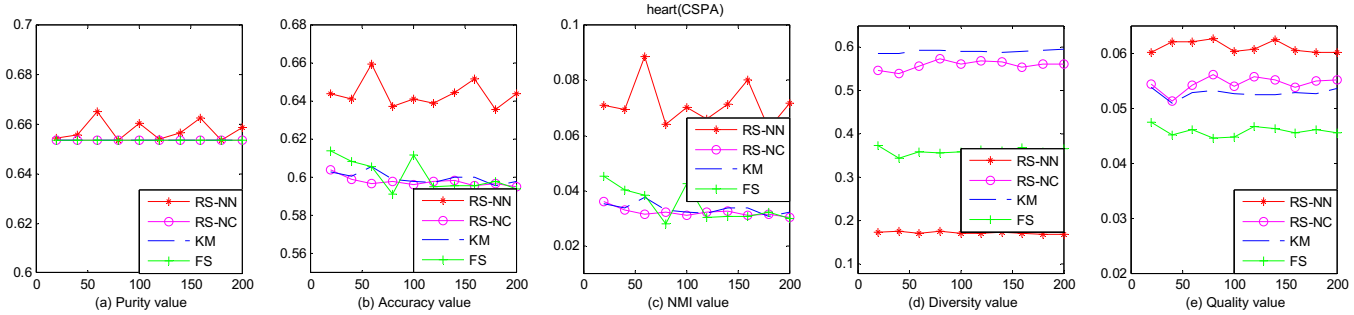


Fig. 15. Performance comparison on heart.

horizontal axis represents the ensemble size, and the vertical axis is the average Purity value, Accuracy value and *NMI* value between the final consensus partitions and the real data partition. The performance of each ensemble generation method based on spectral clustering and HGPA are shown in Figs. 19–35. We only show *NMI* value in this part since the tendency of the other two evaluation criteria consist with NMI value on most datasets according to Figs. 2(a–c)–18(a–c).

Figs. 2(d)–18(d) and Figs. 2(e)–18(e) show the overall diversity and quality of the solutions with the change of ensemble size on

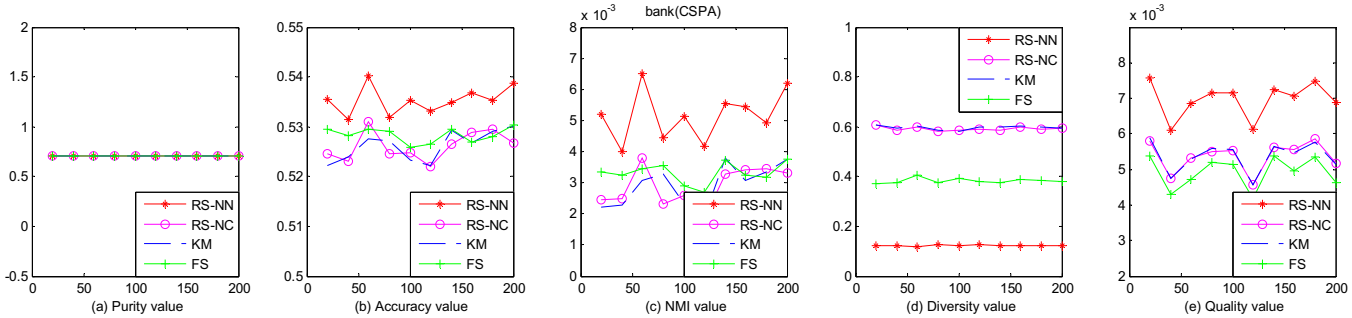


Fig. 16. Performance comparison on bank.

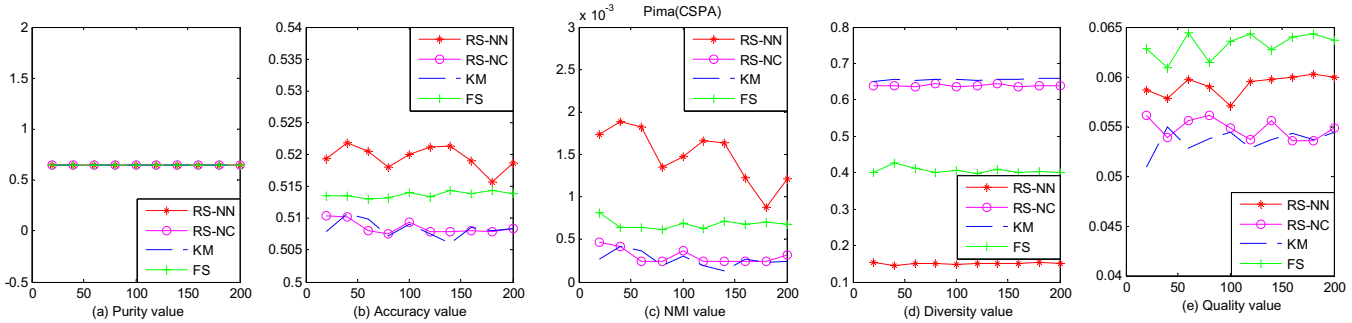


Fig. 17. Performance comparison on Pima.

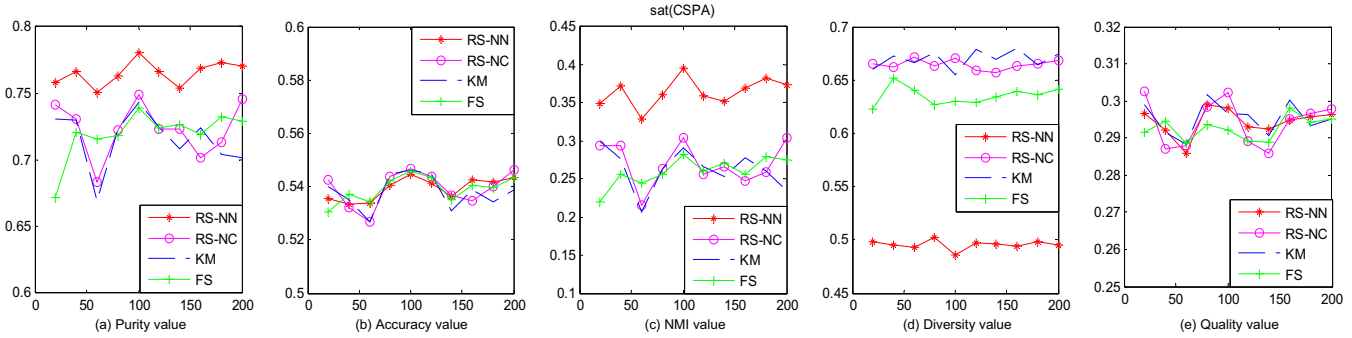


Fig. 18. Performance comparison on sat.

seventeen datasets respectively. The vertical axes show the value of *Quality* and *Diversity*, respectively. Note that each point in the graph is obtained by averaging ten runs.

From the experimental results we observe that:

1. Compared with other three ensemble generation methods, RS-NN achieves comparable or improved performance on all the datasets (with CSPA) and we can easily find that the diversity obtained by RS-NN is the highest in all methods. This is also the main reason why RS-NN has the best performance.
2. The performance of RS-NC is not satisfactory. The trends of RS-NC are similar to KM from the magenta curve in Figs. 2–18. As previously mentioned, the traditional KM may have high quality (on hearts, Iris, wine, glass, Soybean, ionosphere, WDBC, CHART) and low diversity (on all datasets). In some cases, it may have both low quality and low diversity, in particular for Thyroid, bank and Pima data. As discussed in Section 2.2, assigning each instance absent from the current subsample to its closest cluster based on its Euclidean distance to the cluster centers may have no significant differences compared with k-means.
3. FS method always achieves a higher diversity than KM and RS-NC. Compared with KM and RS-NC, when the quality of FS is apparently lower, it achieves better performance compared with KM and RS-NC on wine, Soybean, glass, ionosphere, Lung and sat datasets because its diversity is high. However, when the quality is very low and the diversity of overall solutions is slightly high, FS only obtain comparable performance (SPECTFheart, SPECTheart, hearts, bank and heart datasets) and even worse results on Iris and WDBC data. This also demonstrates that both the quality and diversity should be important factors that influence cluster ensemble performance.
4. From Figs. 19–35, we can easily find that RS-NN achieves better performance than other ensemble generation methods with either spectral clustering or HGPA on most datasets. We also observe that the performance of ensemble clustering depends on both ensemble generation method and consensus function.

As shown in Figs. 19–35, the proposed RS-NN has better performance with HGPA than spectral clustering on wine, Soybean,

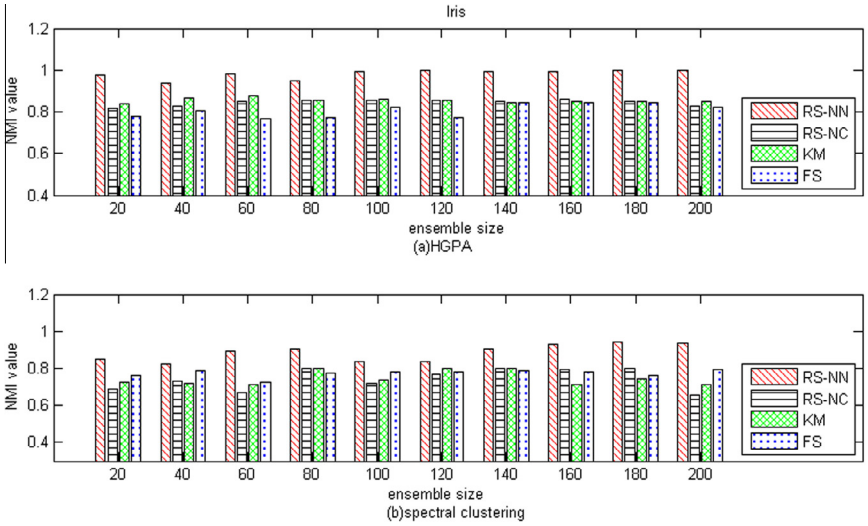


Fig. 19. *NMI* value of different generation methods using spectral clustering/HGPA on Iris.

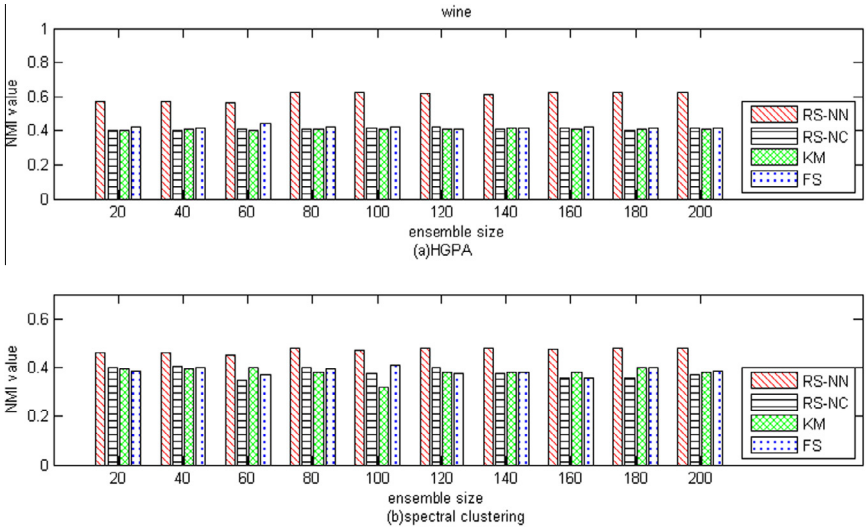


Fig. 20. *NMI* value of different generation methods using spectral clustering/HGPA on wine.

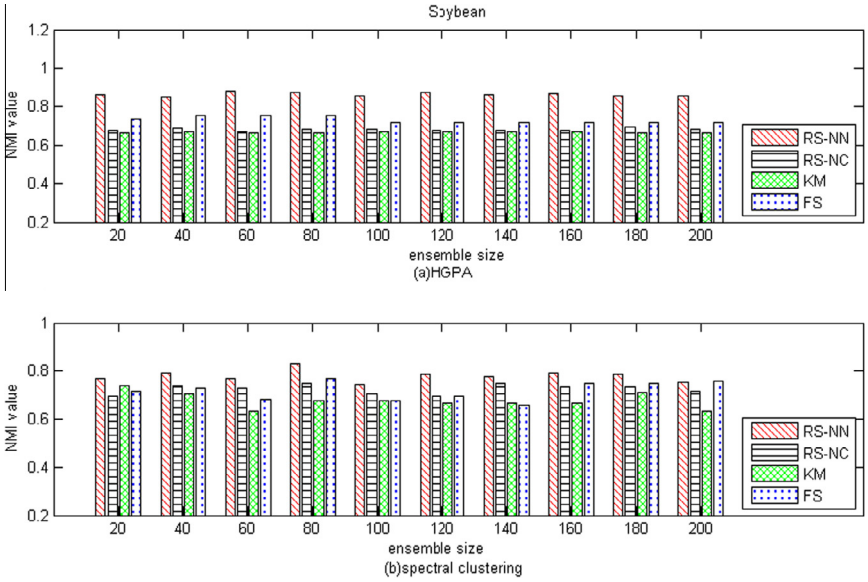


Fig. 21. *NMI* value of different generation methods using spectral clustering /HGPA on Soybean.

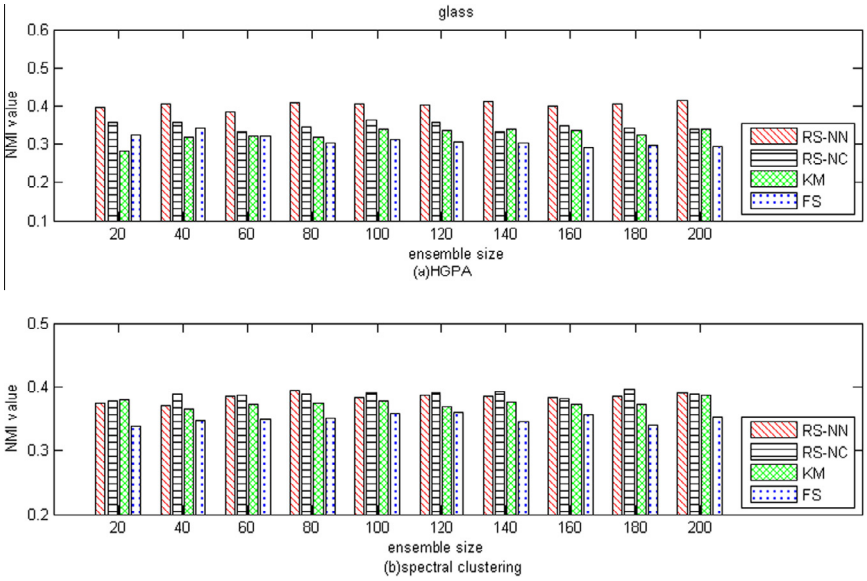


Fig. 22. *NMI* value of different generation methods using spectral clustering/HGPA on glass.

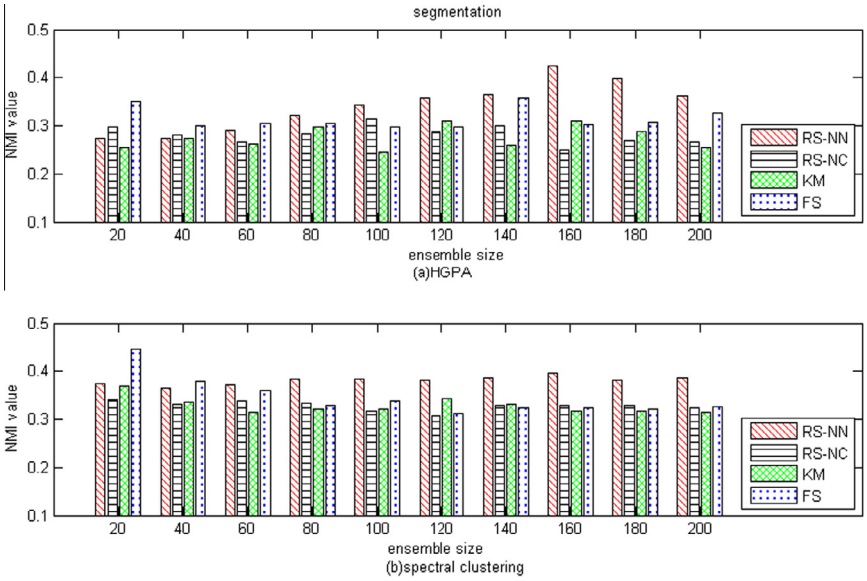


Fig. 23. *NMI* value of different generation methods using spectral clustering/HGPA on segmentation.

hearts, WDBC, CHART, and SPECTFheart datasets and has worse performance on segmentation, Thyroid, ionosphere, bank, Pima and sat datasets. Moreover, compared with all the other generation methods, RS-NN obtains the best results using HGPA and the worst results using spectral clustering on hearts and SPECTFheart. It is noteworthy that on heart and bank datasets we observe exactly the opposite phenomenon for RS-NN, i.e., the performance using spectral clustering is obviously the best, but is the worst using HGPA. Furthermore, spectral clustering outperforms HGPA with all generation methods on Thyroid, ionosphere, bank, Pima and sat data, and has worse performance on hearts data. For WDBC data, all the generation methods have better results when using spectral clustering, except RS-NN. The opposite situation can be observed on heart. In addition, for FS, when the ensemble size is small HGPA has worse performance than spectral clustering on ionosphere and heart data; however, as we increase the ensemble size, HGPA achieves better results.

5.3. The effect of k value

We also conduct some experiments on different k for basic KM because the literature suggests that using a larger k value (the number of clusters) can obtain complex structure in consensus partition, from the combination of small hyper-spherical structures in the single partitions. In this section, we only show a small part of experimental results for the proposed RS-NN on three datasets in Figs. 36 and 37.

From Fig. 36, we observe that the performance of RS-NN method with spectral clustering is better when k is larger (e.g., between 11 and 20). However, in Fig. 37 where CSPA is used as the consensus function, the performance is better when k is smaller (e.g., between 2 and 10). This suggests that the choice is k might be dependent on the consensus function. We speculate that the performance is related to the principle of the consensus function and the structure of the data itself. The relationship between the

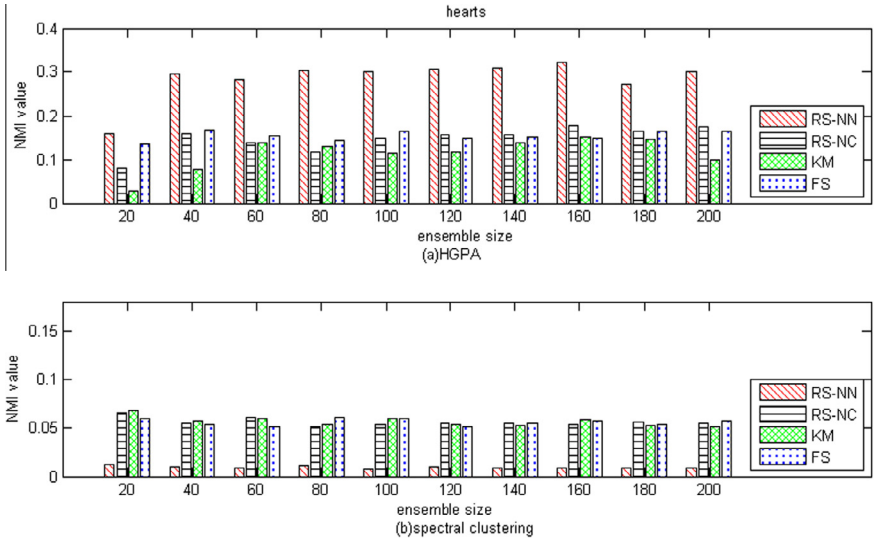


Fig. 24. *NMI* value of different generation methods using spectral clustering/HGPA on hearts.

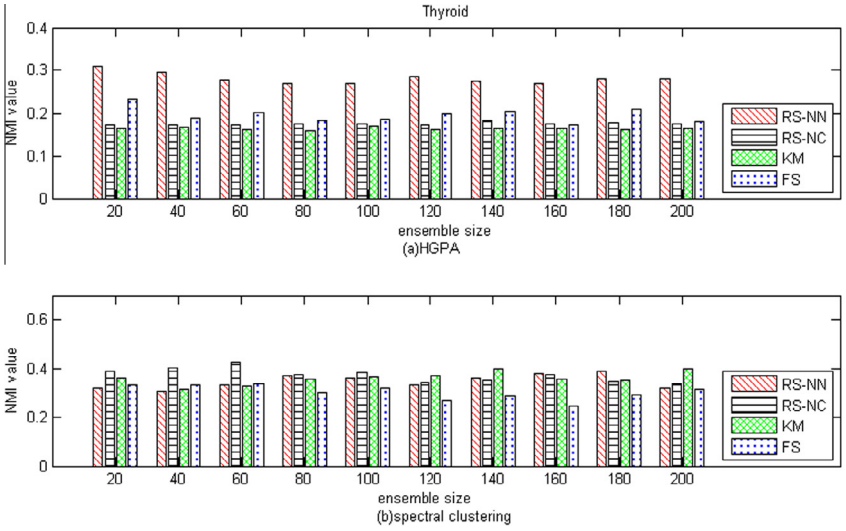


Fig. 25. *NMI* value of different generation methods using spectral clustering/HGPA on Thyroid.

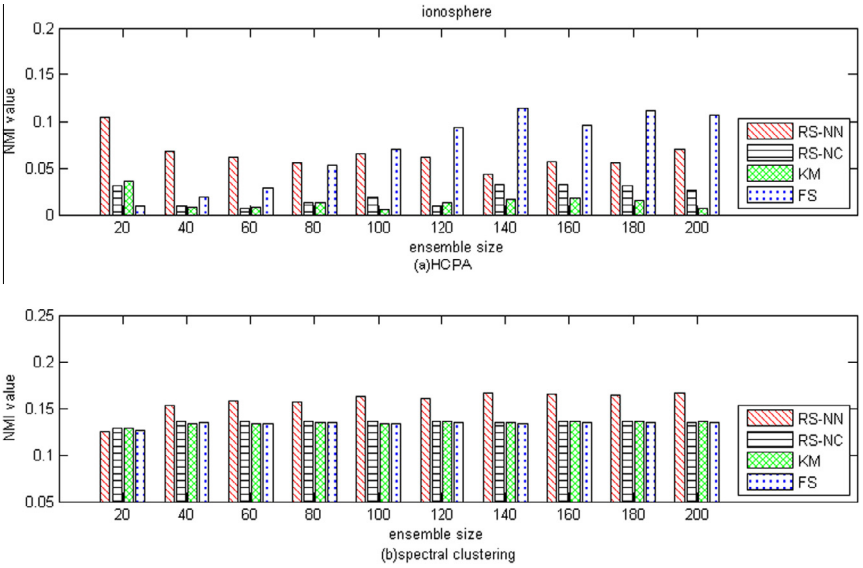


Fig. 26. *NMI* value of different generation methods using spectral clustering/HGPA on ionosphere.

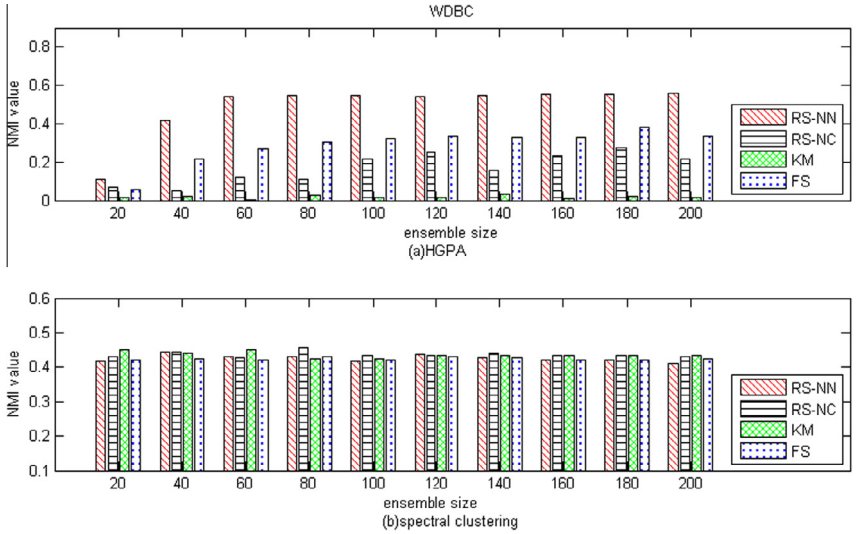


Fig. 27. *NMI* value of different generation methods using spectral clustering/HGPA on WDBC.

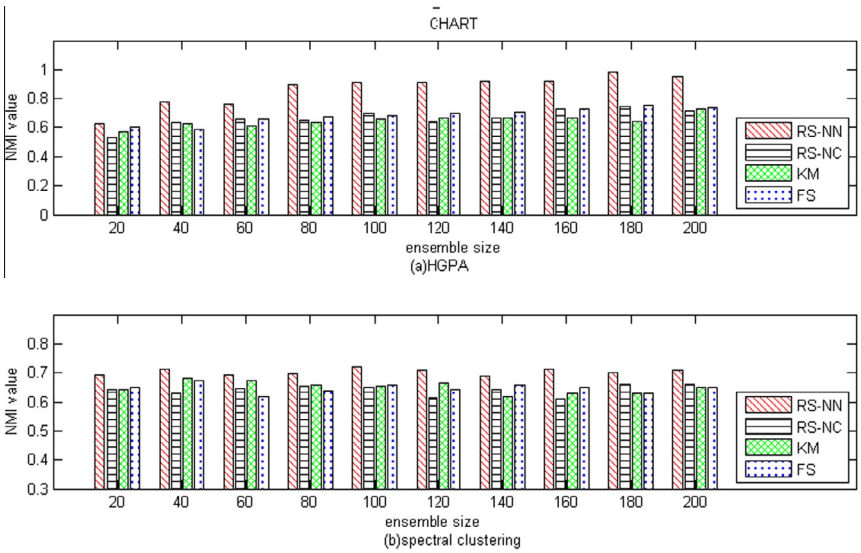


Fig. 28. *NMI* value of different generation methods using spectral clustering/HGPA on CHART.

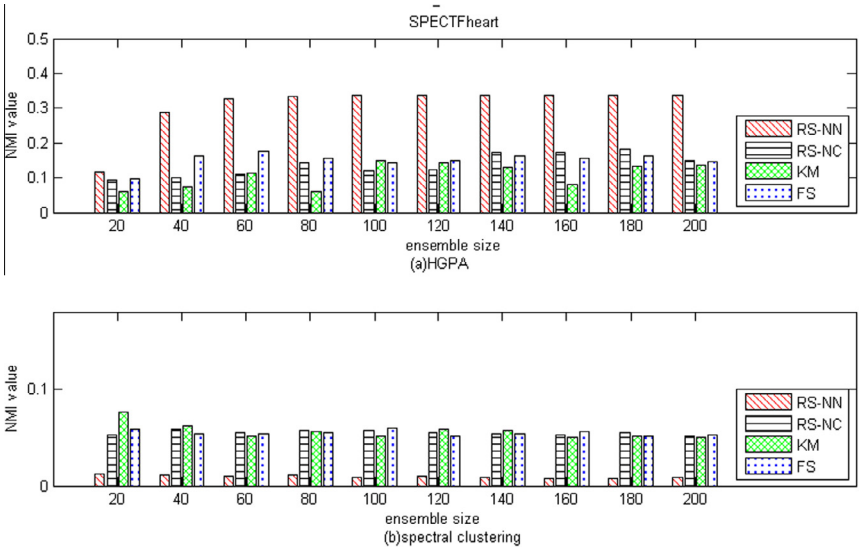


Fig. 29. *NMI* value of different generation methods using spectral clustering /HGPA on SPECTFheart.

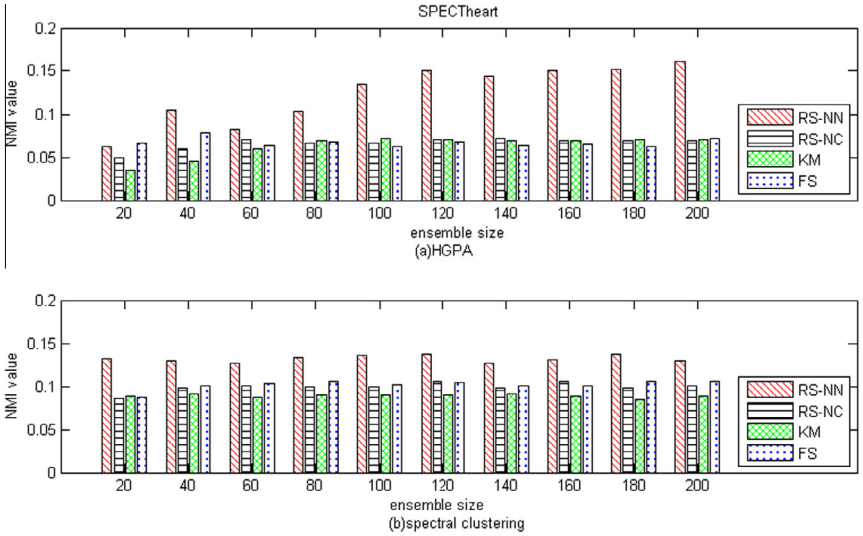


Fig. 30. *NMI* value of different generation methods using spectral clustering/HGPA on SPECTheart.

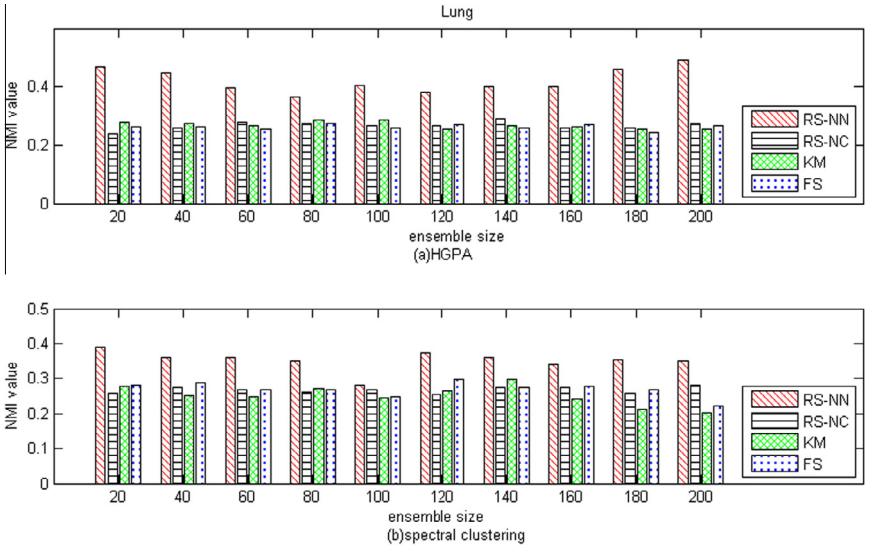


Fig. 31. *NMI* value of different generation methods using spectral clustering/HGPA on Lung.

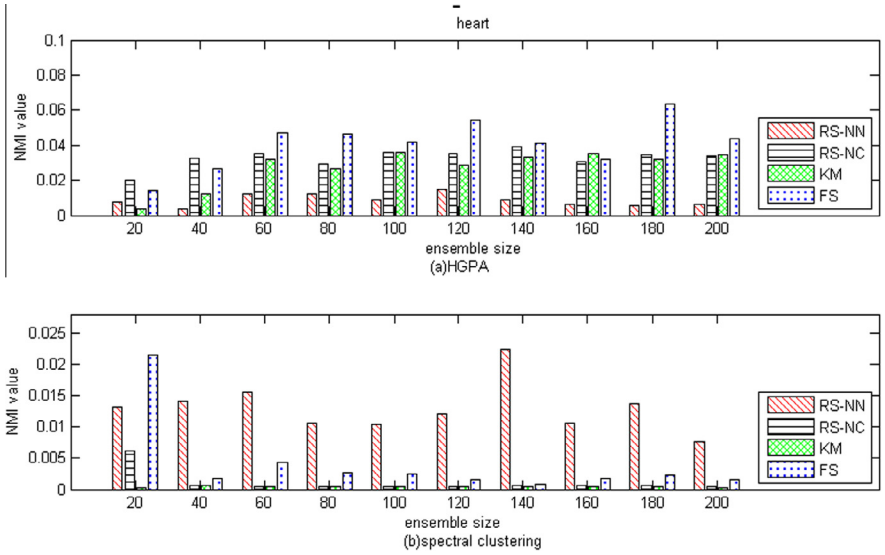


Fig. 32. *NMI* value of different generation methods using spectral clustering/HGPA on heart.

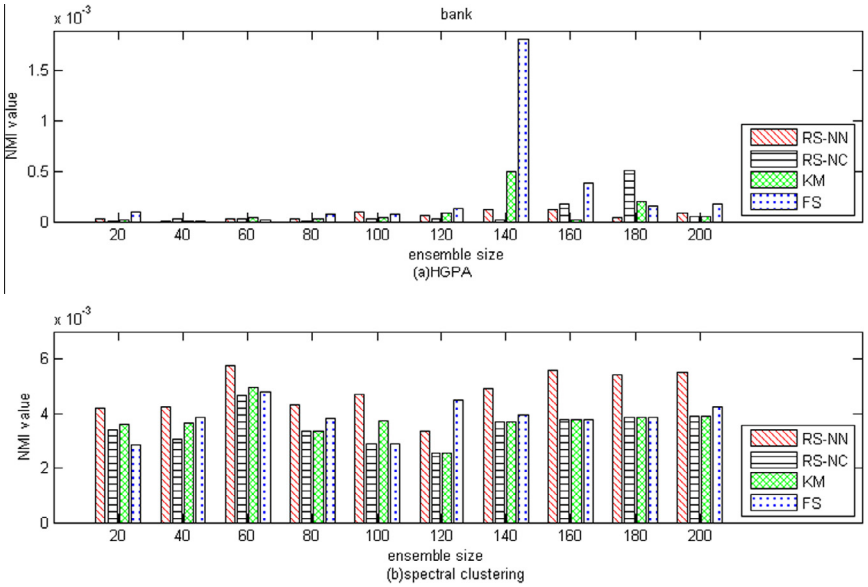


Fig. 33. *NMI* value of different generation methods using spectral clustering/HGPA on bank.

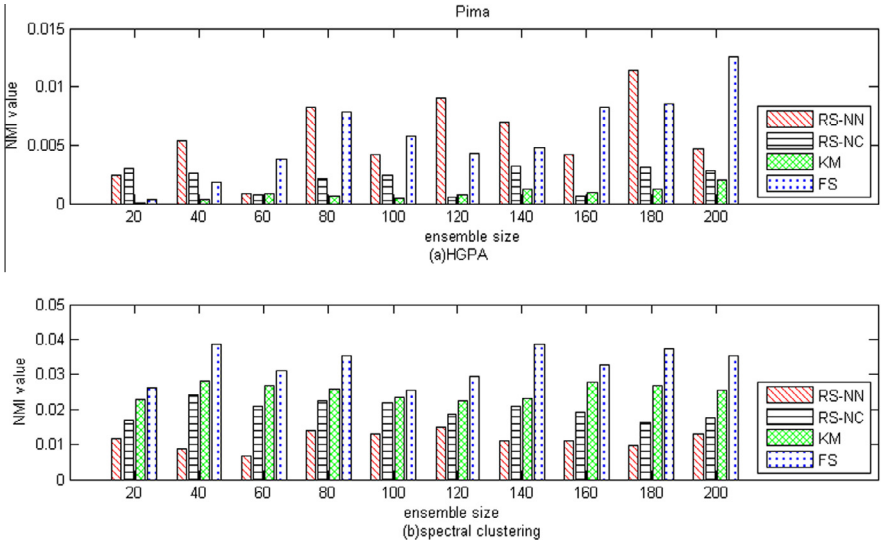


Fig. 34. *NMI* value of different generation methods using spectral clustering/HGPA on Pima.

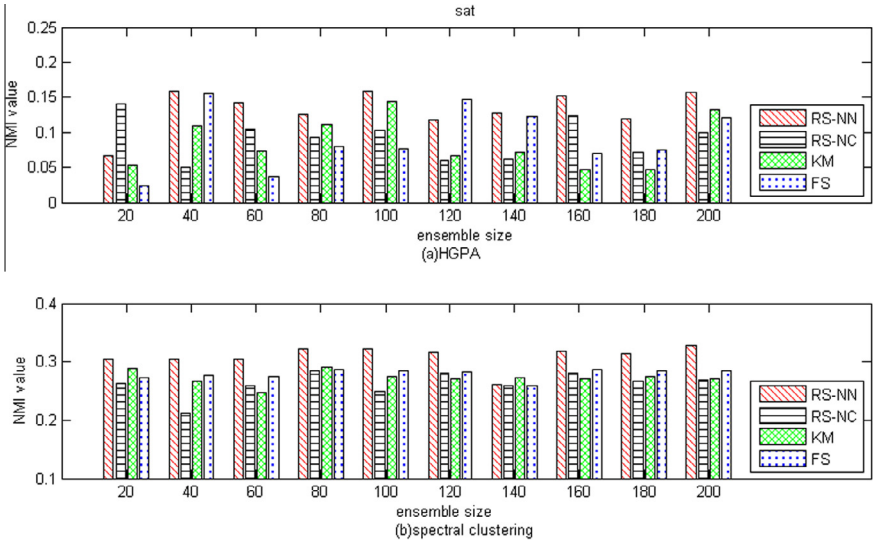


Fig. 35. *NMI* value of different generation methods using spectral clustering/HGPA on sat.

selection of k and consensus functions is also an issue worthy of study.

6. A dual random sampling method

In Figs. 2(a–c)–18(a–c), compared with the RS-NC and KM, FS method (the green curve) achieves comparable performance on eight datasets (hearts, ionosphere, CHART, SPECTFheart, SPECTheart, heart, bank and sat) and even better performance on other seven datasets (segmentation, Thyroid, wine, Soybean, glass Lung and Pima). A possible reason is that FS method can get relatively high diversity while its quality is not very high. (see Figs. 2(d,e)–18(d,e)). It generally reminds us that the performance can be further improved if we combine RS-NN and FS methods together. Hence we propose a dual random sampling method (FS-RS-NN for short) which uses both RS-NN method and FS method when we construct the subset of the original datasets. For comparison, we also combine RS-NC with FS method (FS-RS-NC for short).

The detailed algorithm is summarized as follows in Algorithm 2.

Algorithm 2. FS-RS-NN

Input:
S: The original dataset (the number of cases is N)
R: Predefined ensemble size
K: The range of k values
For i = 1 to R
(i) Get a subsample $s(i)$ by drawing N examples at random but with replacement from S and removing the duplicate examples
(ii) Get a subsample $f(i)$ by perform random sampling with replacement on the feature set of $s(i)$ and removing the duplicate features
(iii) Apply k-means on the subsample $f(i)$ with a k value randomly selected from K to the size of $f(i)$
(iv) Use nearest neighbor method to label the absent examples in the clustering solutions
(v) Construct similarity matrix(i) of the ensemble solutions
End
Compute the average similarity matrix
Apply spectral clustering/CSPA/HGPA to the average similarity matrix to obtain a final partition P^*
Output: The final consensus partition

Similarly we will get FS-RS-NC method if we use nearest centroid method to label the absent examples in Algorithm 2.

Figs. 38(a–c)–54(a–c) show the performance of the dual random sampling methods, FS and RS-NN on the seventeen datasets as the size of ensemble number varies. The consensus function used in this experiment is CSPA. Figs. 38(d)–(d) and Figs. 38(e)–54(e) show the overall diversity and quality, respectively. Each point in the graph is obtained by averaging the results of ten runs.

From Figs. 38(a–c)–54(a–c), we find that FS-RS-NN achieves improved performance on eight datasets (wine, glass, segmentation, Thyroid, WDBC, SPECTheart, heart and Pima) and comparable performance on the other nine datasets compared with RS-NN. Note that FS-RS-NN can get even higher diversity than RS-NN (see Figs. 38(d)–54(d)). This is also the principal reason that FS-RS-NN has got the better performance.

Figs. 39, 41, 49 and 51 show that when the method is very dominant on the diversity, the final results of FS-RS-NN have obvious advantages even if its quality is far below the RS-NN. It indicates that, compared with the quality, the diversity has more important influence on the final performance. Both of RS-NN and FS have low diversity on wine and glass datasets, but FS-RS-NN has significantly high diversity. It also illustrates the advantages of this dual random sampling method on achieving higher diversity.

On the other hand, the quality also has influence on the result. From Figs. 44 and 53, we observe that, compared with RS-NN, FS-RS-NN has no obvious advantage on diversity while its quality is relatively high, and the *NMI* value is still significantly higher over different ensemble size.

Accordingly, the phenomenon showed in Fig. 45 is easy to understand. The diversity of FS-RS-NN has no obvious advantages on diversity while its quality is far below the RS-NN, thus its performance is slightly inferior to the latter. In addition, the performance of FS-RS-NN is not very good on Iris and ionosphere probably due to the fact that the FS method has low diversity and low quality. We can observe this from Fig. 9.

Besides, the additional time cost of FS-RS-NN is another noteworthy issue. Compared with RS-NN, FS-RS-NN method has added the process of feature selection. Thus, it has more time cost in generating the new data subset. However, the new subset is much smaller in size than that is used in RS-NN as it has fewer features. Consequently, this will save the running time of k-means clustering. The comparison on the time cost of one run of sampling and clustering is shown in Table 2, where the *Sample-time* denotes

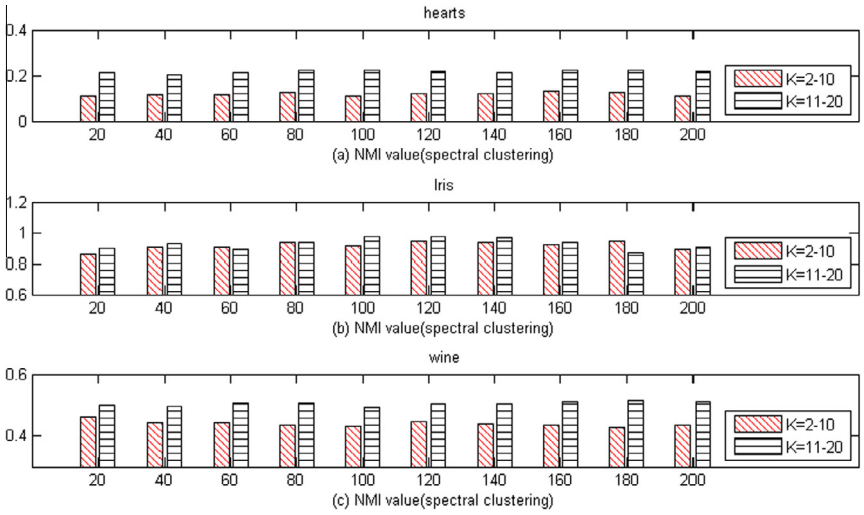


Fig. 36. Performance of RS-NN with spectral clustering.

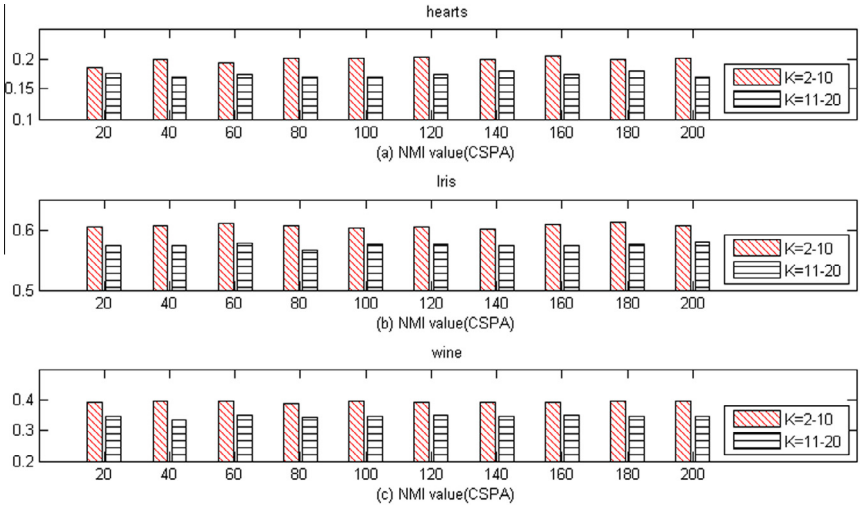


Fig. 37. Performance of RS-NN with CSPA.

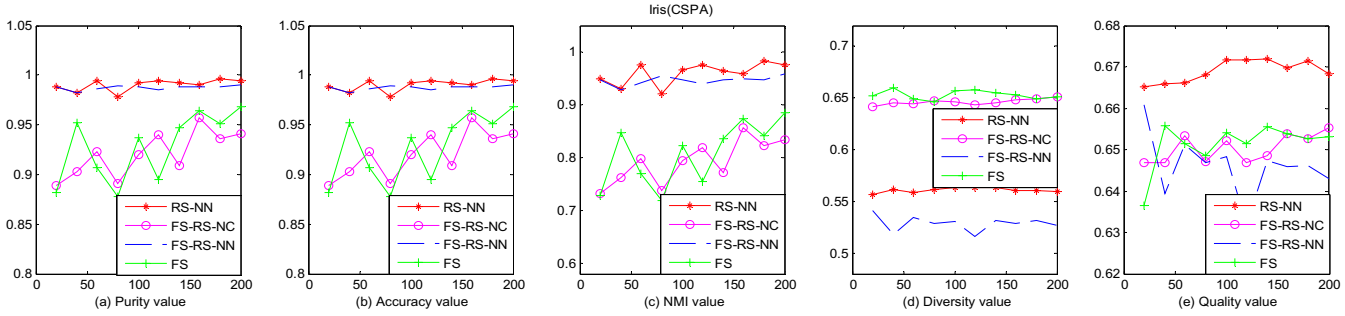


Fig. 38. Performance comparison on Iris.

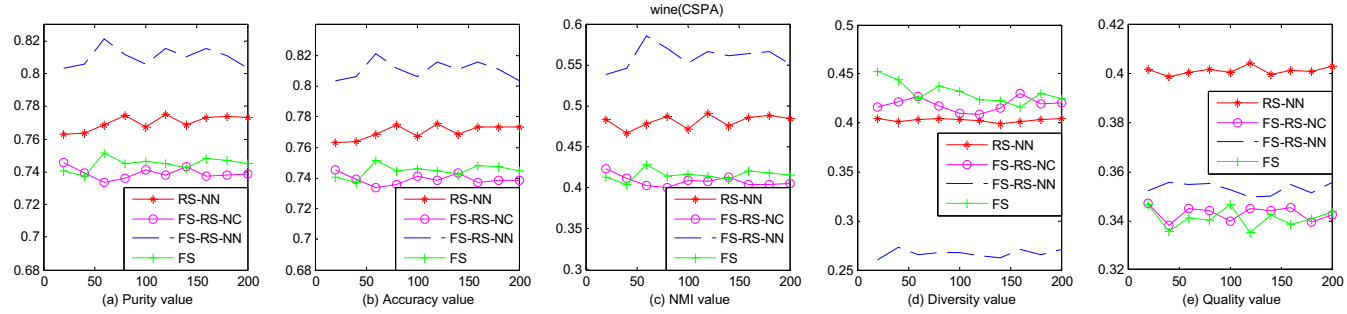


Fig. 39. Performance comparison on wine.

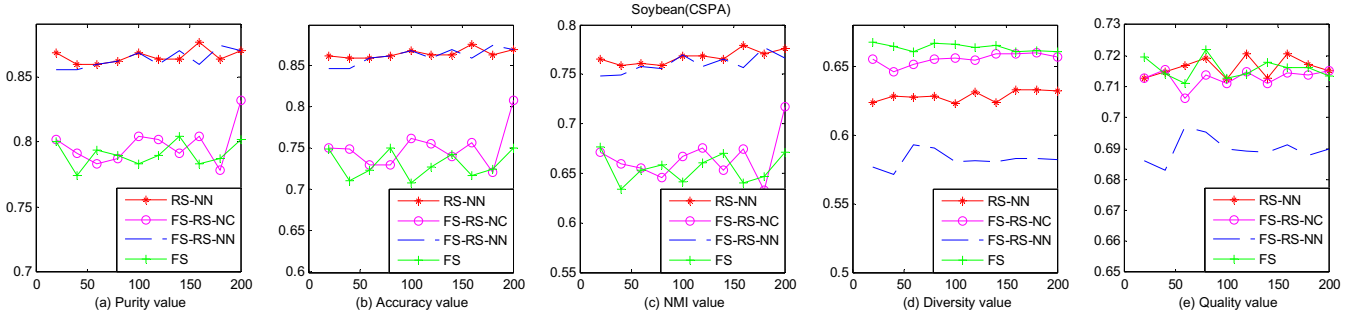


Fig. 40. Performance comparison on Soybean.

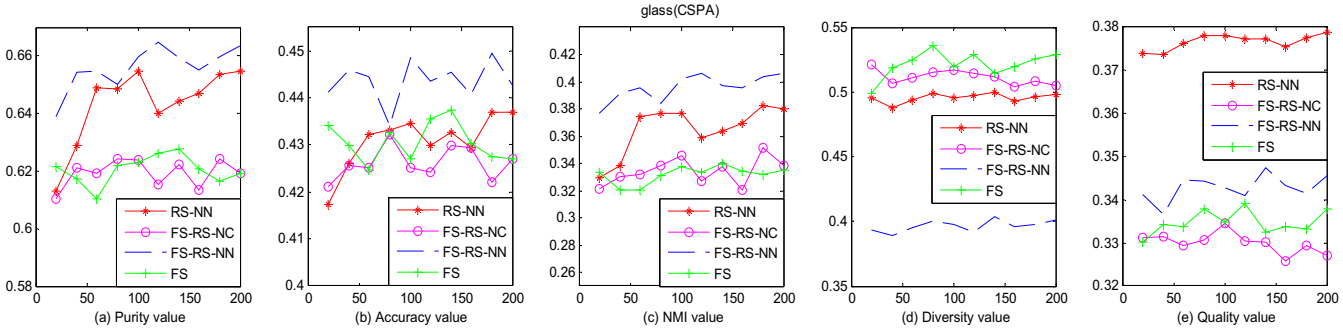


Fig. 41. Performance comparison on glass.

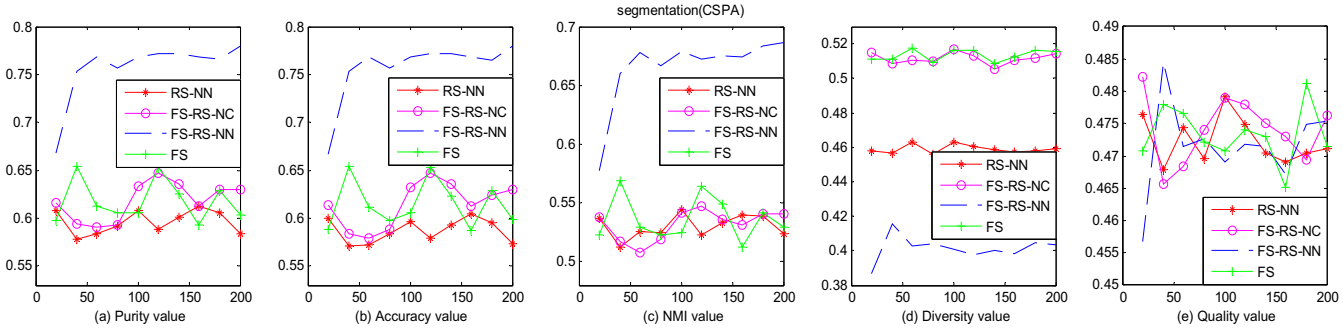


Fig. 42. Performance comparison on segmentation.

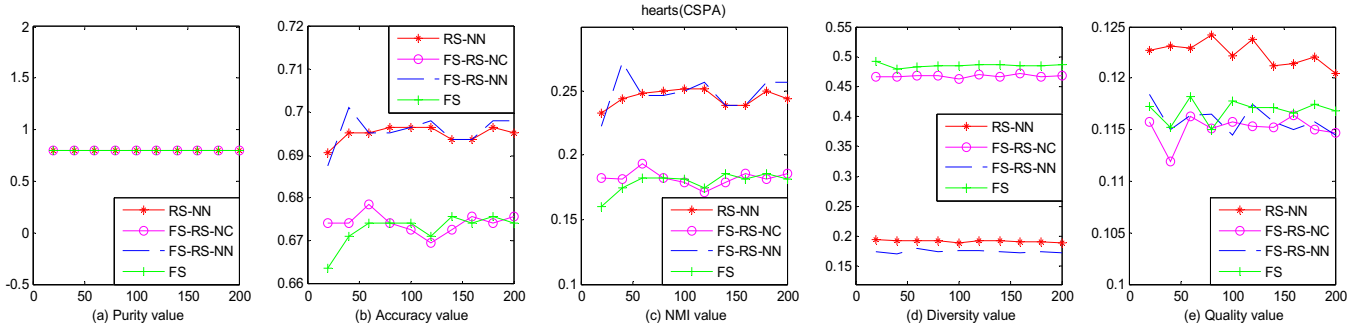


Fig. 43. Performance comparison on hearts.

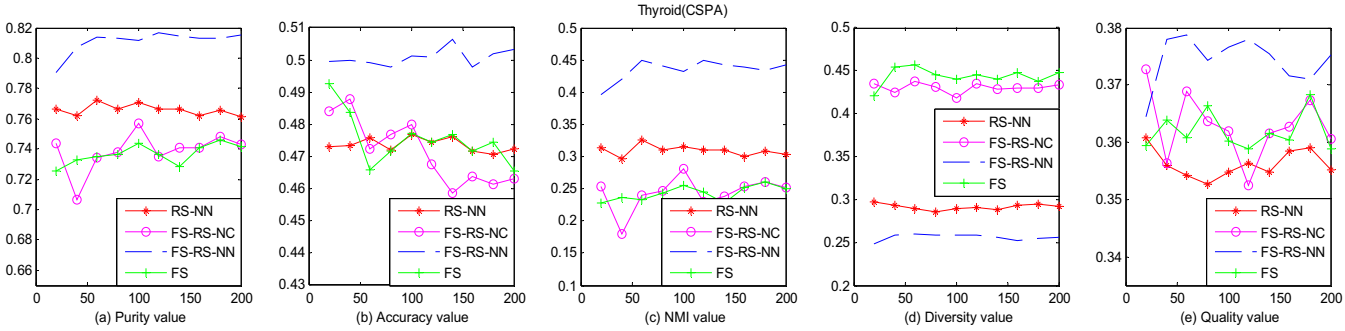


Fig. 44. Performance comparison on Thyroid.

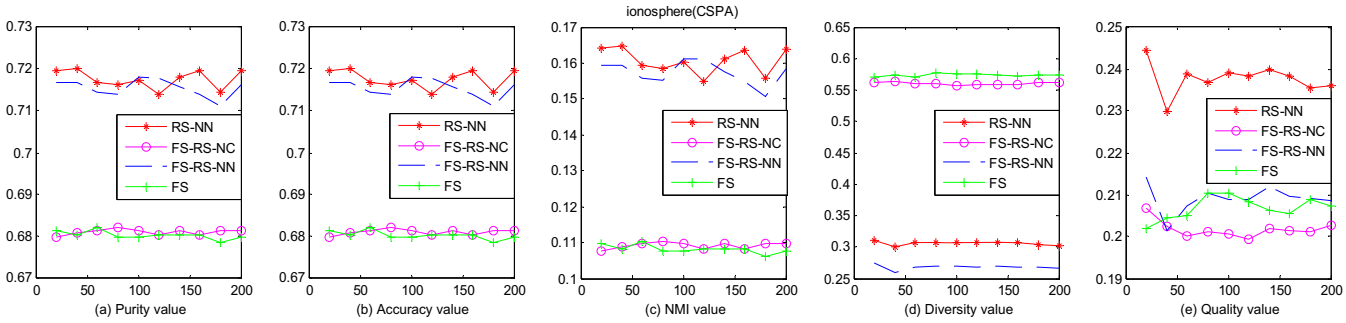


Fig. 45. Performance comparison on ionosphere.

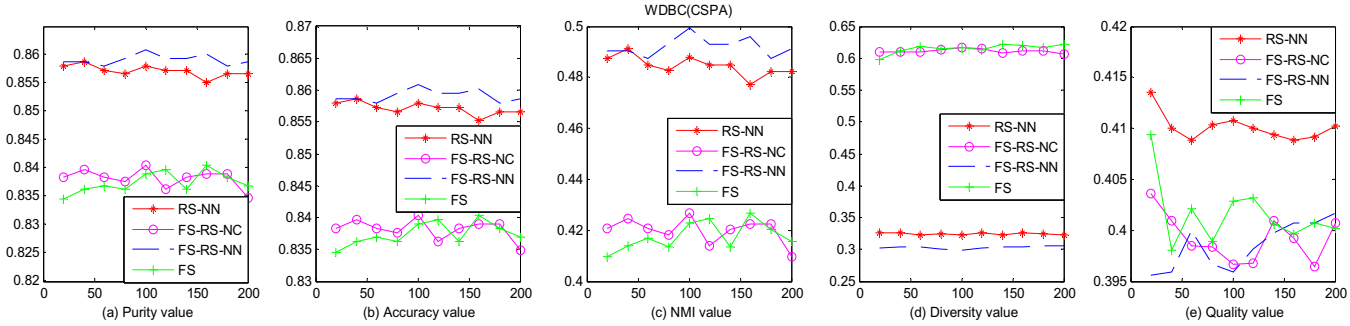


Fig. 46. Performance comparison on WDBC.

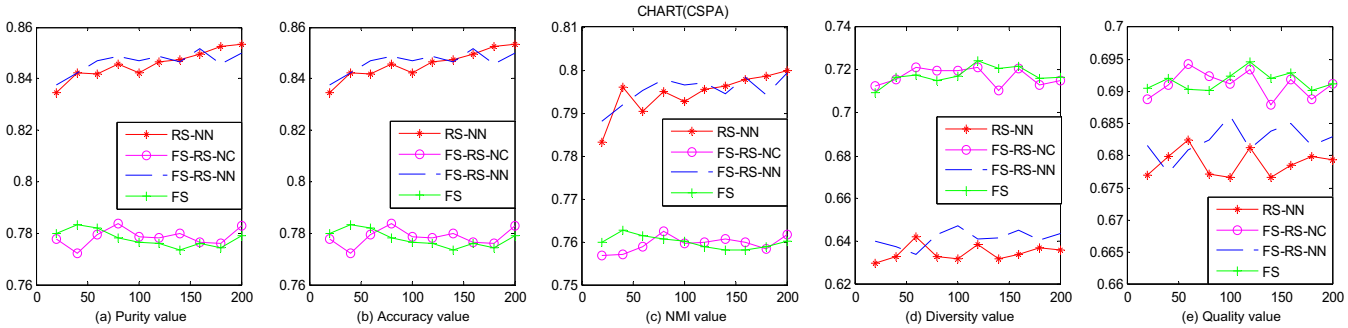


Fig. 47. Performance comparison on CHART.

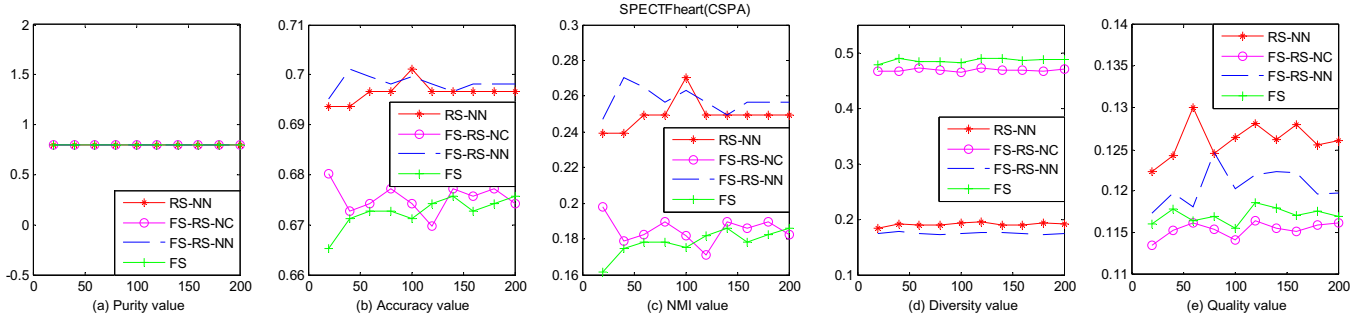


Fig. 48. Performance comparison on SPECTFheart.

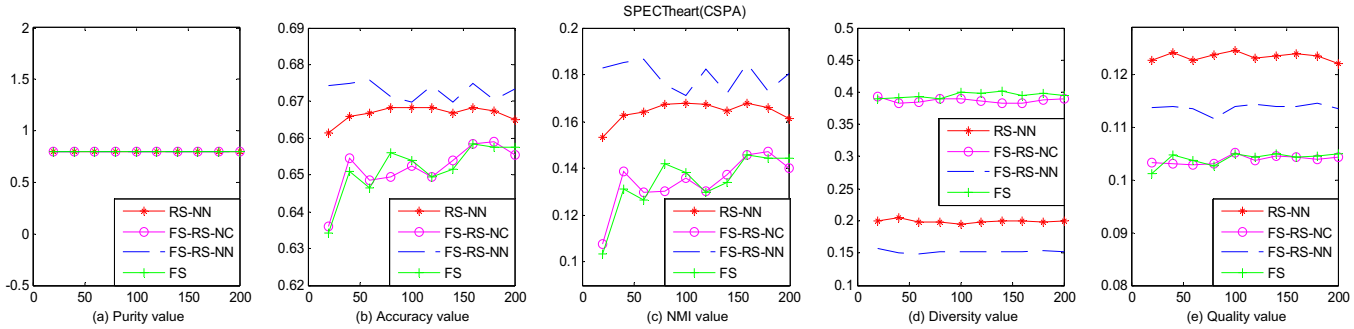


Fig. 49. Performance comparison on SPECTheart.

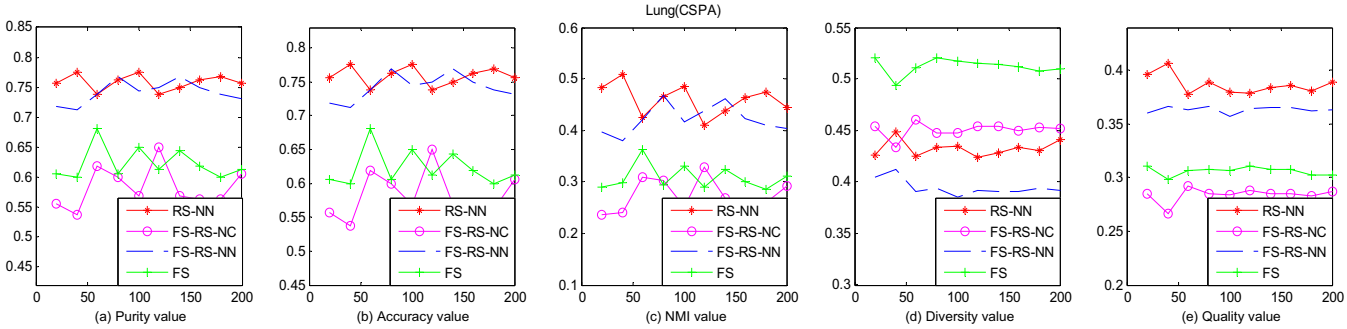


Fig. 50. Performance comparison on Lung.

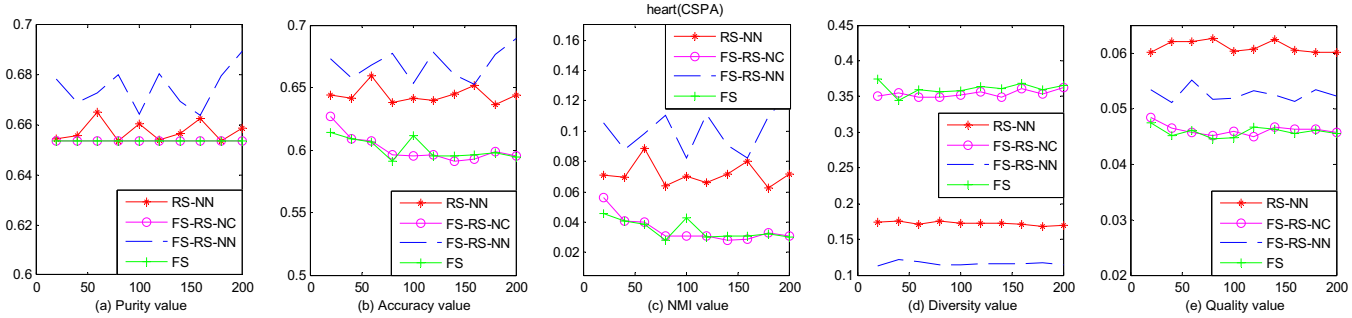


Fig. 51. Performance comparison on heart.

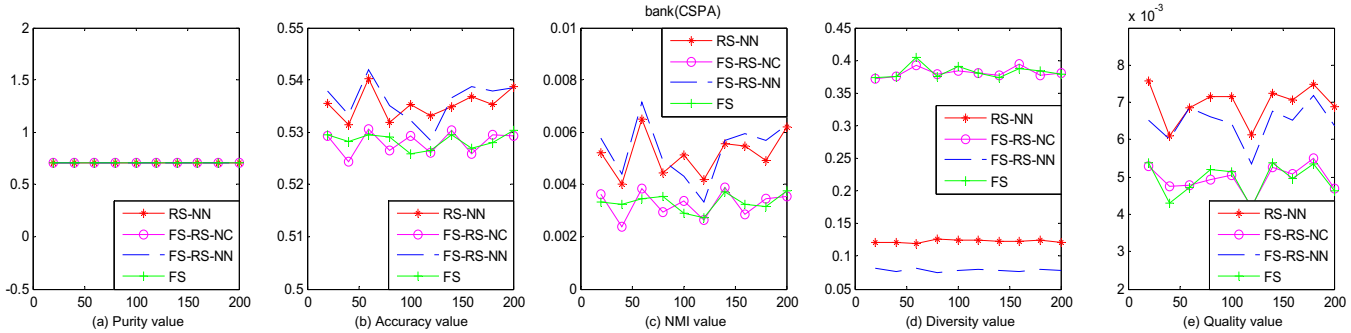


Fig. 52. Performance comparison on bank.

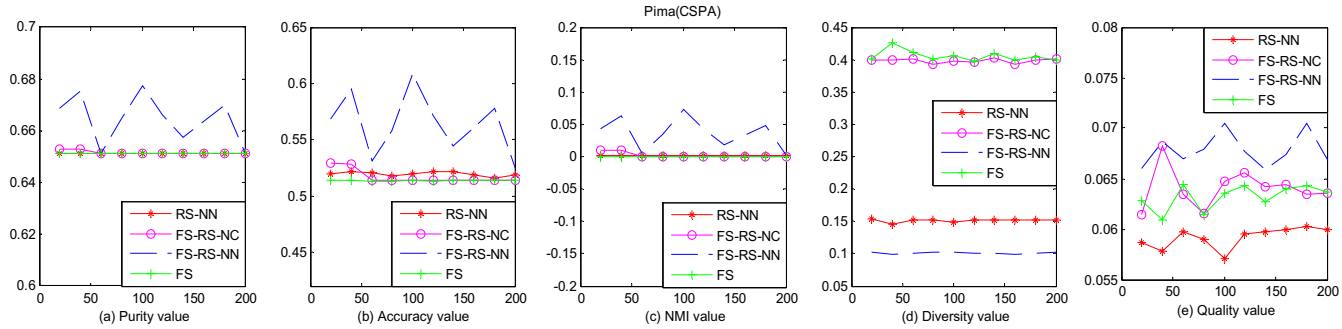


Fig. 53. Performance comparison on Pima.

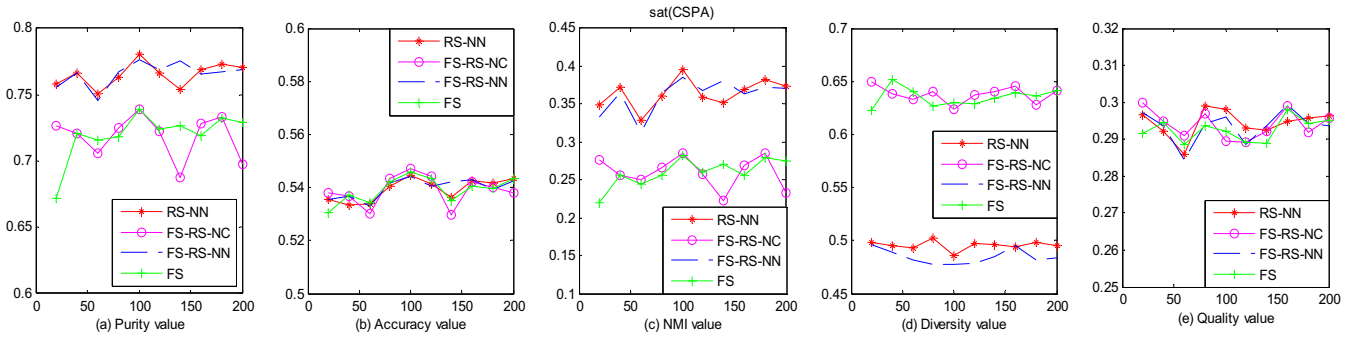


Fig. 54. Performance comparison on sat.

Table 2
Comparison of the time consumption on 17 datasets.

	Sample-time(10^{-4} s)		Total-time (s)	
	RS-NN	FS-RS-NN	RS-NN	FS-RS-NN
Iris	1.84	1.99	0.0031	0.0028
wine	2.51	3.00	0.0042	0.0040
Soybean	0.82	1.5	0.0026	0.0024
glass	2.99	3.56	0.0051	0.0050
segmentation	46	47	0.0343	0.0252
hearts	3.48	4.16	0.0189	0.0136
Thyroid	2.68	2.88	0.0052	0.0049
ionosphere	4.35	4.91	0.0172	0.0120
WDBC	7.37	7.94	0.0226	0.0160
CHART	7.85	8.97	0.0115	0.0092
SPECTFheart	3.45	4.22	0.0154	0.0096
SPECTheart	3.68	4.22	0.0089	0.0070
Lung	0.55	1.57	0.0030	0.0027
heart	6.13	6.59	0.0102	0.0089
bank	156	158	0.0477	0.0475
Pima	11	11	0.0114	0.0097
sat	652	658	0.1901	0.1696

the time cost in the construction of the new subset of data, and the *Total-time* is the time consumption of the overall process including the clustering. Note that each number in the table is obtained by averaging over 100 runs. We can easily find that FS-RS-NN is slightly faster than RS-NN.

7. Conclusions

In this paper, we propose a new cluster generation method based on random sampling, i.e., random sampling method which uses the nearest neighbor method to fill the category information of the missing samples. We evaluated its performance in comparison with the traditional k-means, FS and RS-NC. Experimental results indicate that the FS method always achieves more diverse partitions while RS-NC method gets better results on the quality of every partition. Further, to introduce more diversity, we propose a dual random sampling method which combines RS-NN and FS methods. The new method can obtain higher diversity and acceptable quality, and achieve better or comparable results on most datasets without additional time costs.

There are several avenues for future research. First, experiments demonstrate that both quality and diversity are crucial while diversity has more important influence on the final performance than quality. How to choose the solutions with high diversity and good quality and how to weigh the importance between them, namely, cluster ensemble selection is an interesting research direction.

Second, our work has presented two new ensemble generation methods. An interesting question is whether a high diversity can be reached by using ensembles generated by different methods (e.g., RS-NN, FS, k-means).

Third, our current work deals with ensemble generation in particular ensemble generation with random sampling and random projection along with K-means clustering. A natural research question is can the proposed approaches be used with other types of clustering algorithms such as hierarchical clustering algorithms?

Our study also demonstrates that the performance of ensemble clustering is contingent on both ensemble generation methods and consensus functions. In addition, the influence of different k values on the diversity requires in-depth investigation.

Finally, our proposed RS-NN can be considered as an integration of instance selection and clustering with nearest neighbors. In our future work, we will consider other instance selection methods such as instance selection based on hubness (i.e., instance that are the nearest neighbors of many other instances) and investigate those methods into our ensemble generation framework.

Acknowledgment

This work is supported by the Natural Science Foundation of China under Grant No. 61202144 and No. 61203282, the Natural Science Foundation of Fujian Province under Grant No. 2012J05125, the Jiangsu 973 Scientific Project (BK2011023), the Key Laboratory of System Control and Information Processing, Ministry of Education of Shanghai Jiao Tong University under Grant No. SCIP2012007, the National Natural Science Foundation of China (61272419), the US National Science Foundation under Grants DBI-0850203, CNS-1126619, and IIS-1213026, and the U.S. Department of Homeland Security under grant Award Number 2010-ST-062000039.

References

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery Journal, 14*(1), 63–97.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics, 19*(9), 1090–1099.

Fern, X. Z., & Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning* (pp. 186–193).

Fern, X. Z., & Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the 21st international conference on machine learning* (pp. 281–288).

Fern, X., & Lin, W. (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining, 1*(3), 128–141.

Filkov, V., & Skiena, S. (2004). Integrating microarray data by consensus clustering. *International Journal of Artificial Intelligence Tools, 13*(4), 863–880.

Fischer, B., & Buhmann, J. M. (2003). Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(4), 513–518.

Fred, A. L. N., & Jain, A. K. (2002). Data clustering using evidence accumulation. In *Proceedings of international conference on pattern recognition* (pp. 276–280).

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clustering using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 835–850.

Greene, D., Tsymbal, A., Bolshakova, N., & Cunningham, P. (2004). Ensemble clustering in medical diagnostics. In *Proceedings of the 17th IEEE symposium on computer-based medical systems* (pp. 576–581).

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion, 7*(3), 264–275.

Ilc, N., & Dobnikar, A. (2012). Generation of a clustering ensemble based on a gravitational self-organising map. *Neurocomputing, 96*, 47–56.

Karypis, G., Aggarwal, R., Kumar, V., & Shekhar, S. (1999). Multilevel hypergraph partitioning: Applications in VLSI domain. *IEEE Transactions on Very Large Scale Integration, 7*, 69–79.

Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing, 20*(1), 359–392.

Kuncheva, L., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Proceedings of IEEE international conference on systems, man and cybernetics* (pp. 1214–1219).

Li, T., Ding, C., & Jordan, M. I. (2007a). Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of 7th IEEE international conference on data mining* (pp. 577–582).

Li, T., Ding, C., & Jordan, M. I. (2007b). Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Seventh IEEE international conference on data mining, ICDM 2007* (pp. 577–582).

MacQueen, J. (1967). Some methods for classifications and analysis of multivariate observations. In *Proceedings of 5th Berkeley symposium on mathematical statistics and probability* (pp. 281–297).

Minaei-Bidgoli, B., Topchy, A., & Punch, W. F. (2004). Ensembles of partitions via data resampling. In *Proceedings of international conference on information technology* (pp. 188–192).

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*, 91–118.

Neumann, D. A., & Norton, V. T. (1986). Clustering and isolation in the consensus problem for partitions. *Journal of Classification, 3*, 281–298.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research, 3*, 583–617.

Topchy, A., Jain, A. K., & Punch, W. (2003). Combining multiple weak clusterings. In *Proceedings of IEEE international conference on data mining* (pp. 331–338).

Topchy, A., Jain, A. K., & Punch, W. (2004). A mixture model for clustering ensembles. In *Proceedings of SIAM conference on data mining* (pp. 379–390).

Topchy, A., Jain, A. K., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(12), 1866–1881. 2005.

Vega-Pons, S., Correa-Morris, J., & Ruiz-Shulcloper, J. (2010). Weighted partition consensus via kernels. *Pattern Recognition, 43*(8), 2712–2724.

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence, 25*(03), 337–372.

Wakabayashi, Y. (1998). The complexity of computing median of relations. *Resenhas, 3*, 311–323.

Xu, R., & Wunsch, D. II, (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*, 645–678.

Yoon, H.-S., Ahn, S.-Y., Lee, S.-H., Cho, S.-B., & Kim, J. H. (2006a). Heterogeneous clustering ensemble method for combining different cluster results. *Lecture Notes in Bioinformatics, 3916*, 82–92.

Yoon, H.-S., Lee, S.-H., Cho, S.-B., & Kim, J. H. (2006b). A novel framework for discovering robust cluster results. *Lecture Notes in Artificial Intelligence, 4265*, 373–377.

Yu, Z. W., Li, L., You, J., Wong, H. S., & Han, G. Q. (2012c). SC3: Triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(6), 1751–1765.

Yu, Z. W., Wong, H. S., You, J., Yu, G. X., & Han, G. Q. (2012b). Hybrid cluster ensemble framework based on the random combination of data transformation operators. *Pattern Recognition, 45*, 1826–1837.

Yu, Z. W., You, J., Wong, H. S., & Han, G. Q. (2012a). From cluster ensemble to structure ensemble. *Information Science, 198*, 81–99.

Zheng, L., Li, T., & Ding, C. (2010). Hierarchical ensemble clustering. In *Proceedings of 2010 IEEE international conference on data mining (ICDM 2010).*

Zhu, S., Wang, D., & Li, T. (2010). Data clustering with size constraints. *Knowledge-Based Systems, 23*(8), 883–889.

Generating Textual Storyline to Improve Situation Awareness in Disaster Management

Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, Ning Xie
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, U.S.A.
Email: {wzhou005,cshen001,taoli,chens,nxie}@cs.fiu.edu

***Abstract*—Hurricane Sandy affected the east coast of U.S. in 2012 and posed immense threats to businesses, human lives and properties. In order to minimize the consequent loss of a catastrophe like this, a critical task in disaster management is to understand situation updates about the disaster from a large number of disaster-related documents, and obtain a big picture of the disaster's trends and how it affects different areas. In this paper, we present a two-layer storyline generation framework which generates an overall or a global storyline of the disaster events in the first layer, and provides condensed information about specific regions affected by the disaster (i.e., a location-specific storyline) in the second layer. To generate the overall storyline of a disaster, we consider both temporal and spatial factors, which are encoded using integer linear programming. While for location-specific storylines, we employ a Steiner tree based method. Compared with the previous work of storyline generation, which generates flat storylines without considering spatial information, our framework is more suitable for large-scale disaster events. We further demonstrate the efficacy of our proposed framework through the evaluation on the datasets of three major hurricane disasters.**

Keywords: Textual Storyline, Situation Awareness, Disaster Management

I. Introduction

Natural disasters such as hurricanes, earthquakes and tsunamis cause inestimable physical destruction, loss of life and property around the world every year. For example, Hurricane Sandy affected the east coast of U.S. in 2012 and posed immense threats to businesses, human lives, and properties. In order to minimize the consequent loss of the disasters, a critical task in disaster management is to efficiently analyze and understand the disaster-related situation updates. This requires effective information gathering methods to operate on a myriad of web documents, e.g., news and reports that are related to the disasters. The domain experts expect to obtain condensed information about the detailed disaster event description, e.g., the evolutionary tendency of the disaster with respect to different locations [1]. However, it is often a non-trivial task to generate a big picture of the disaster events due to the flood of web documents.

To tackle this problem, various types of document understanding systems have been developed over the last decade. These systems include (1) summarization-based systems [2], [3], [4], [5], [6] that choose from multiple documents a subset of sentences conveying the principle idea; (2) topic detection and tracking systems [7] aiming to group documents into different clusters as events and monitor future events related to the corresponding topic; and (3) timeline generation systems [8], [9] that create summaries to present the evolution of an event by leveraging temporal information attached to or extracted from the documents. These systems are able to alleviate the so-called *information overload* problem to some extent; however, they suffer from several limitations that may affect the quality of the summarized results. First, most of them focus on summarizing an event via topic evolution over the time, but ignore the spatial information which is important especially for large-scale disaster events. For instance, for a hurricane which affects several states of U.S., a domain expert may be interested in how these regions are affected, and how the hurricane evolves over different geo-spatial regions. Second, these systems usually generate a single layer summarization or storyline to reflect topic changes over the entire event. However, due to the spatial factor, the information evolution over a disaster event is intrinsically hierarchical. In most cases, domain experts are often interested in not only the general picture of a disaster, but also how it affects a particular region.

In this paper, we propose a storyline generation framework that addresses the aforementioned limitations by generating a two-layer storyline that consists of global storylines for cross-location disaster events on the first layer and location-specific storylines for individual events on the second layer. Specifically, in our framework, a disaster event is initially summarized from a large set of documents (e.g., news and reports) with a big picture showing how the disaster affects different regions. It can then be zoomed into a specific location for more detailed location-specific event summarization. In the cross-location layer, integer linear programming is employed to summarize the event via a list of representative locations, each of which is associated with a short description. On the location-specific layer, a Steiner-tree based approach is applied to generate a storyline for each specific location. A demo of our system can be found at http://bigdata-node01.cs.fiu.edu/HurricaneStoryline/.

In summary, the contributions of this work are three-fold:

- We present a novel two-layer summarization framework to summarize multiple disaster-related documents. The first layer provides an overall summary of the disaster events, while the second layer gives condensed information on how specific locations/regions were affected by the disaster.

- We consider both temporal and spatial factors when generating summaries for the disaster events, and these two factors enable us to reason on the evolution of events over time and locations. The generated summaries can be naturally represented as a storyline.
- We conduct quantitative experiments and case studies on crawled web documents related to three major hurricane disasters, and the results demonstrate the efficacy of our proposed framework in generating readable and understandable summaries.

The rest of the paper is organized as follows. After discussing related work in Section II, we first define our problem in Section III. In Section IV, an overview of our proposed framework is introduced. Detailed descriptions of how to generate a global storyline and a local storyline are presented in Section V and Section VI, respectively. We evaluate our system in Section VII and finally conclude our work and discuss potential extensions of the proposed framework in Section VIII.

II. Related Work

In this section, we highlight some previous research results that are most relevant to this work in the following three directions: multi-document summarization, topic detection and tracking, and storyline generation. We will also discuss several useful disaster situation-specific tools.

Multi-document summarization is a mechanism which addresses the information overload problem by compressing a given collection of documents into a concise summary. In general, it can be categorized into extractive and abstractive summarization [10]. Extractive summarization [11] selects important sentences from the original documents to form a summary, while abstractive summarization [11] paraphrases the corpus using new sentences. The latter usually employs natural language generation techniques such as information fusion, sentence compression and reformulation. Our work is more related to extractive summarization. Various multi-document summarization methods have been proposed over the last decade, including centroid-based [12], graph-based [13], [5], knowledge-based [1], [14], and etc. Other methods, such as non-negative matrix factorization, latent semantic analysis, and sentence-based topic models, have also been applied to generate the summaries by selecting semantically important sentences in the documents [15], [16]. Most existing extractive summarization methods generate short summaries by selecting sentence from the input; however, they often ignore the implicit temporal, spatial and structural information possibly presented in the documents.

Topic detection and tracking (TDT) is a research program initiated by DARPA (Defense Advanced Research Projects Agency) for finding and following the new events in streams that broadcast news stories[1]. It consists of three major technical tasks: tracking known events, detecting unknown events, and segmenting a news source into stories. Many promising approaches have been proposed and identified during the TDT evaluation, in particular within the information retrieval and natural language processing communities [7], [17], [18]. However, previous research efforts only focused on detecting the flat structure of events, and fail to consider the hidden hierarchies of topics.

Storyline generation aims to obtain a sequence of summaries that describe how an event evolves over time, and has attracted great attention recently. For example, Google News Timeline clusters incoming articles into groups based on topics and lists the generated groups in chronological order. Alonso et al. [19] proposed a framework for generating temporal snippets to improve user search experience. These methods consider the temporal information as references and represent the results in chronological order. Recently, Wang et al. [9] proposed a framework that integrates text, image, and temporal information to generate storyline-based summaries to reflect the evolution of the given topic. Lin et al. [20] presented a framework for generating storylines from microblogs for user input queries. Shahaf et al. [8] proposed a methodology called *metro map* for creating structural summaries of documents by optimizing several objectives (e.g., relevance, coherence, coverage and connectivity) simultaneously. Jiang et al. [21] proposed an temporal event summarization solution to summarize the temporal dynamics of the event sequences using the inter-arrival information. Unlike these existing systems, our framework takes into account the spatial information and generates storyline-based summaries to reflect the evolution of a given topic over different geo-spatial regions.

Disaster Situation-specific Tools: Commercial systems such as Web EOC and E-Team are usually used by Emergency Management departments located in urban areas [22], [23]. Recently Ushahidi provides a platform to crowd source news stories and crisis information using multiple channels and prepares visualization and interactive maps [24] and GeoVISTA monitors tweets to form situation alerts on a map-based user interface according to the geo-locations associated with the tweets [25]. These situation-specific tools provide query interfaces, GIS and visualization capabilities to support user interaction and query [26]. However, they do not generate textual storylines to improve the situation awareness.

III. Problem Definition

To summarize what is happening in the vicinity of a given disaster, we present a storyline of the disaster in the form of a two-layer graph of events.

Definition An *event* is represented by a tuple (t, l, s) where t is the time that the event occurs, l is the location and s is the textual description about the event. For example, (08/27/2011, New York City, "The five main New York City-area airports will be closed to arriving flights") represents an event in Hurricane Sandy.

The problem of generating a storyline can be defined as follows:

Input: A collection of documents related to a disaster.

Output: A two-layer storyline consists of the most representative events summarizing the evolution of disaster-relevant topics. The **first layer** (or the upper layer) is a chain of events $(o_1, \ldots, o_n)$, as the global temporal and spatial evolution of a disaster, therefore also referred as the global storyline. An

[1] http://projects.ldc.upenn.edu/TDT/

event of the upper layer o_i can be further expanded in the **second layer** (or the lower layer) to a connected tree of events as the temporal and topic evolution locally for a specific location of o_i.

A global storyline, which is a chain of events, describes how the disaster moves over time by the location attribute of the events and how the disaster affects different areas by the description attributes. The chain structure is used under the assumption that a disaster at any time should have only one geo-spatial center, which should move continuously over time. Such an assumption is valid for most of the natural disasters like hurricanes, storms, and blizzards, but not for the man-made disasters like cyber attacks. In our future work, we will explore more complicated evolution structures of different disaster types. For local storyline generation, we follow previous work of storyline generation [9] to use a tree structure as the storyline to capture more topics in the topic evolution, allowing multiple topics to coexist at the same time.

IV. System Framework

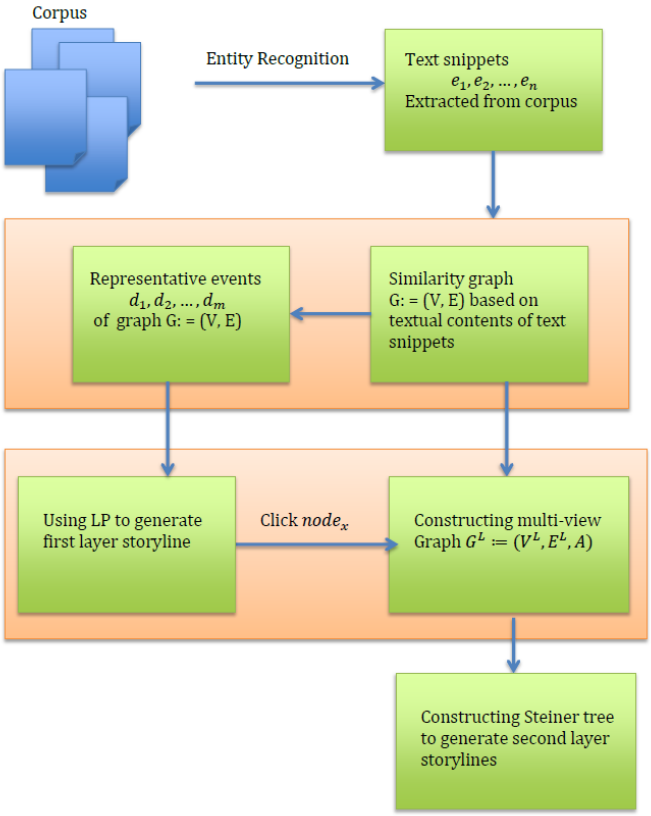


Fig. 1. The High-level System Overview

Figure 1 shows our system framework. Given a collection of documents related to a disaster, we first extract text snippets as sentences with time and location phrases, which are identified by Stanford NER [27]. Time phrases are normalized by SUTime [28] to timestamps and location phrases are mapped to geocodes by Google API[2]. Together with its timestamp and geocode, a snippet approximately describes an event.

[2]https://developers.google.com/maps/documentation/geocoding

In our framework, the extracted text snippets are first organized as a similarity graph, followed by two layers of processing, corresponding to the two layers of the output. In the first layer, a minimum dominating set algorithm is employed on the snippet graph to find several representative events, on top of which an integer linear programming method is then proposed to find a chain of events reflecting the overall spatial evolution of the disaster as the global storyline. We visualize the global storyline on a map using Google map APIs.

If a user is interested in certain area and click it on the map, the map will be zoomed-in the clicked area and display the local storyline of the area. To do this, a sub-graph of the overall similarity graph is first induced and augmented to a multi-view graph. The same minimum dominating set algorithm is first applied to the sub-graph for finding representative events, and then followed by a Steiner tree algorithm to make the selected events temporally smooth and coherent.

V. Global Storyline Generation

A. Text Snippet Graph Construction

Although each text snippet can be considered as an event, many of those are redundant. To remove the redundancy and obtain a set of representative events, we construct a graph $G = (V, E)$ with the given text snippets as the vertex set V, and add an edge between each pair of snippets which are likely to refer to the same event. Specifically, for two nodes $v_i, v_j \in V$, we first convert these two text snippets into two feature vectors as n-gram bags-of-words, then compute the cosine similarity between these two feature vectors. $e_{ij} = (v_i, v_j) \in E$ if and only if both the similarity of v_i and v_j is greater than a similarity threshold parameter α, and their distance calculated by their geocode is less than a distance threshold parameter $radius$. Note that the latter constraint takes the spatial smoothness of events into consideration.

B. Identifying Events via Dominating Set

We identify the set of representative events in the original snippets with minimum redundancy by solving the minimum dominating set problem. A vertex u of a graph dominates another vertex v of the graph, if u and v are joined by an edge in the graph. A subset of S of the vertex set of an undirected graph is a dominating set if for each vertex u, either u is in S or a vertex in S dominates u. The *Minimum Dominating Set* (MDS) problem is to find a dominating set with minimum size. MDS has been previously used to model multi-document summarization problem [5]. In our case, we use the MDS of text snippets to capture the representative events from the text snippets of disaster event descriptions.

The MDS problem is known to be NP-hard but an efficient greedy algorithm by Johnson [29] is known to achieve an approximation ratio of $H(d+1)$, where d is the maximum degree of the graph and $H(n) = \sum_{i=1}^{n} \frac{1}{i}$ is the harmonic function.[3] The greedy algorithm is described in Algorithm 1 and was also used in [5].

[3]Johnson's greedy algorithm was initially designed for the Set Cover problem, but it is well-known that there is an L-reduction between MDS and Set Cover.

Algorithm 1 Greedy MDS Approximation Algorithm

INPUT: Graph $G = (V, E)$, MDS upper bound W
OUTPUT: dominating set S

$S \leftarrow \emptyset$
$T \leftarrow \emptyset$
while $|S| < W$ and $S \neq V(G)$ **do**
  for $v \in V(G) - S$ **do**
    $s(v) \leftarrow |N(v) \setminus T|$
  end for
  $v^* \leftarrow \arg\max_v s(v)$
  $S \leftarrow S \cup \{v^*\}$
  $T \leftarrow T \cup N(v^*)$
end while

C. Storyline Generation by Connecting Dominating Objects via Linear Programming (LP)

Using Algorithm 1, we generate the dominating set of $G(V, E)$, m text snippets $d_1, \ldots, d_m$, as the representative events. Without loss of generality, the set of events are assumed to be in chronological order. To generate a global storyline capturing the major location change of the disaster, we select a sequence of nodes $o_1, o_2, \ldots, o_l$ from the representative events in chronological order. Intuitively, the generated storyline should also be in spatial coherence, reflecting the continuous location change of the disaster over time. Since a disaster is likely to affect adjacent areas in a similar fashion, the storyline should be coherent in content as well.

Based on the above discussions, we model the storyline generation problem using integer linear programming. To select a chain of nodes from $d_1, \ldots, d_m$, we use variables $node\text{-}active_i \in \{0, 1\}, i = 1 \ldots m$ to indicate whether d_i is included in the selected chain, and $next\text{-}node_{ij} \in \{0, 1\}, i, j = 1 \ldots m$ to indicate that d_i and d_j are two successive nodes (i.e., a transition) in the chain. The objective function aims to maximize the storyline's content coherence which is defined as the minimal similarity between two successive nodes along the storyline as shown below:

$$Coherence(o_1, o_2, \ldots, o_n) = \min_{i=1,2,\ldots,n-1} similarity(o_i, o_{i+1}).$$

We further impose the following set of constraints to model storyline's spatial coherence.

Chain Constraints: It should be guaranteed the consistency of variables $node\text{-}active_i$ and $next\text{-}node_{ij}$, and that the selected nodes should compose a chain in chronological order.

// A node has at most one in-edge and at most one
// out-edge

$$\forall_j : \sum_i next\text{-}node_{i,j} \leq node\text{-}active_j, \quad (1)$$

$$\forall_i : \sum_j next\text{-}node_{i,j} \leq node\text{-}active_i. \quad (2)$$

// The number of active transitions is equal to the
// number of active nodes minus one

$$\sum_i node\text{-}active_i - \sum_{i,j} next\text{-}node_{i,j} = 1. \quad (3)$$

// The chain is ordered chronologically:

$$\forall_{i>j} : next\text{-}node_{i,j} = 0. \quad (4)$$

// A transition of two node can not be active if
// there exists an active node between them.

$$\forall_{i<k<j} : next\text{-}node_{i,j} \leq 1 - node\text{-}active_k. \quad (5)$$

Length Constraints: The selected chain should be in a reasonable length ranged between pre-defined minimum length threshold $\mathcal{L}_{min}$ and maximum length threshold $\mathcal{L}_{max}$.

$$\mathcal{L}_{min} \leq \sum_i node\text{-}active_i \leq \mathcal{L}_{min}. \quad (6)$$

Location Smoothness Constraints: We require both pairwise and triple-wise smoothness of location change on the selected chain. Let $\mathcal{D}_{i,j}, i, j = 1, \ldots, m$ be the distance based pairwise location relationship between d_i and d_j, and $\mathcal{D}_{i,j} = 1$ if distance between d_i and d_j is less than a pre-defined distance parameter, $\mathcal{D}_{i,j} = 0$ otherwise. For triple-wise smoothness, let $\mathcal{A}_{i,j,k}$ be the angle based triple-wise location relationship, and $\mathcal{A}_{i,j,k} = 1$ indicates the angle constructed by three successive nodes d_i, d_j and event k is not an acute one, otherwise $\mathcal{A}_{i,j,k} = 0$. By not including in the chain three successive nodes of which the angle is acute, we excludes the back-and-forth events from the storyline and smooth the location change.

// Distance of two successive nodes should be
// within some range

$$\forall_i : \sum_j (1 - \mathcal{D}_{i,j}) \cdot next\text{-}node_{i,j} \leq 0. \quad (7)$$

// Three successive nodes can not construct
// an acute angle

$$\forall_{i,j,k} : next\text{-}node_{i,j} + next\text{-}node_{j,k} \leq 1 + \mathcal{A}_{i,j,k}. \quad (8)$$

Minimal Similarity Constraints: Let $\mathcal{S}_{ij}, i, j = 1 \ldots, m$ be the cosine similarity between d_i and d_j. we can use the following constraints to find the similarity of the minimum similar transition $min\text{-}edge$ among active transitions.

$$\forall_{i,j} : min\text{-}edge \leq 1 - (1 - \mathcal{S}_{i,j}) \cdot next\text{-}node_{i,j} \quad (9)$$

The Objective Function: Besides to maximize minimal similarity between two successive nodes along the storyline, we also try to make storyline as long as possible, so the objective function has the following form

$$\text{Maximize: } min\text{-}edge + \lambda \cdot l, \quad (10)$$

where λ is a coefficient parameter.

Although integer linear programming is an NP-hard problem, there are efficient approximation algorithms and implementations such as IBM CPLEX[4], which is used for optimization in this paper.

VI. Local Storyline Generation

A global storyline presents a general high-level picture of how a disaster affects different areas when it hits these areas. To show how the disaster affects a specific area locally for a longer time period during preparation and recovery, we allow users to zoom-in to a node $node_x$ of the global storyline. Once a user clicks the node $node_x$, a new graph $G^L(V^L, E^L)$ will be constructed, which is an induced sub-graph of $G(V, E)$, where V^L includes all text snippet nodes which are close to $node_x$ according to their associated geocodes. For the graph $G^L(V^L, E^L)$, we employ the storyline generation method proposed in [9] to generate a storyline for the selected area.

A. Augmented Multi-view Graph Construction

Definition A *multi-view graph* is a triple $G = (V, E, A)$, where V is a set of vertices, E is a set of undirected edges, and A is a set of directed edges.

Different from the global storyline generation where the temporal and spatial information of text snippets are modeled by integer linear programming, here we incorporate temporal information in an augmented multi-view graph $G^L = (V^L, E^L, A)$ from $G^L = (V^L, E^L)$, where A is a set of directed edges for temporal relationship between events. To define edges in A, we introduce two additional parameters $\tau_1, \tau_2, 0 < \tau_1 < \tau_2$. For every pair of nodes o_i, o_j in V, we draw an arc from o_i to o_j if $\tau_1 < t_j - t_i < \tau_2$, where t_i, t_j are the timestamps of o_i and o_j, respectively.

B. Generating Storylines via Directed Steiner Tree

Similar to generating global storylines, after extracting a dominating set of $G^L = (V^L, E^L)$ which represent the main content topics, we need to generate a storyline capturing the temporal and structural information of the local event descriptions. To tackle this problem, we use the concept of Steiner Tree. A *Steiner tree* of a graph G with respect to a vertex subset X is the edge-induced subtree of G that contains all the vertices in X with minimum cost, where the cost is often measured by the size of the tree.

Problem: Given a directed graph $G = (V, A)$, a set X of vertices (called *terminals*), and a root $v_0 \in X$ from which every vertex of X is reachable in G, find the subtree of G rooted at v_0 containing X with the smallest total vertex weight.

This problem is known to be NP-hard since the undirected version is already NP-hard. While the undirected version has been well studied, much less work has been done on direct4ed version [30]. An intuitive solution for this problem is to find the shortest path from the root to each of the terminal and then merge the paths. Of course, this does not guarantee the optimal solution.

[4]http://www.ibm.com/software/commerce/optimization/cplex-optimizer/

We make use of an algorithm due to Charika et al. [30]. The algorithm takes a level parameter $i \geq 1$. In addition, it takes as input the target terminal set Y, the root r, and the required number of nodes to cover, k. When $i = 1$, it leads to the intuitive solution: i.e., selecting the top k shortest path from the root to k nodes and return the union of those paths. Let the length of every arc $(u, v) \in A$ is 1. We will make initial call of $A_i(k, v_0, X)$ with X is the dominating set calculated by Algorithm 1 based on graph G, v_0 is the event among X with the earliest timestamp, and k is $|X|$, the size of X. We interpret the output tree as a local storyline evolving from the root event to all the other dominating events. For a constant i, the algorithm is known to run in polynomial time and produces an $O(k^{\frac{1}{i}})$-approximate solution [30].

Algorithm 2 $A_i(G, k, r, X)$

INPUT: $G = (V, A)$: directed multi-view graph
X : target vertex set X
$r \in X$: the root X
$k \geq 1$: the target size X
OUTPUT: T: a Steiner tree rooted at r covering at least k vertices in X

```
T = ∅
while k > 0 do
    T_best ← ∅
    cost(T_best) ← ∞
    for each vertex v, (v_0, v) ∈ A, and k', 1 ≤ k' ≤ k do
        T' ← A_{i-1}(k', v, X) ∪ {(v_0, v)}
        if cost(T_best) > cost(T') then
            T_best ← T'
        end if
        T ← T ∪ T_best
        k ← k − |X ∩ V(T_best)|
        X ← X \ V(T_best)
    end for
end while
return T
```

VII. System Evaluation

A. Datasets

TABLE I. Statistics of the datasets.

keyword	#documents	#text snippets
Hurricane Katrina	800	1572
Hurricane Sandy	795	2253
Hurricane Irene	691	2186

We collect datasets from Bing News Search[5] using keywords about three major hurricanes in the last ten years (i.e., Hurricane Katrina, Hurricane Irene, and Hurricane Sandy) to evaluate our storyline generation system. For the search results returned from Bing News Search, we extract the text content from the corresponding web pages. Basic statistics about the datasets are shown in Table I, and some examples of extracted text snippets are shown in Table II.

B. Summarization Performance of Global Storylines

To evaluate the quality of global storylines generated by our proposed framework, a human labeler manually composed

[5]http://news.bing.com

TABLE II. EVENTS EXAMPLE EXTRACTED FROM DOCUMENT USING ENTITY RECOGNITION

content	time	location
This photo made available by the New Jersey governor's office shows flooding and damage in Seaside Heights, N.J. on Oct. 30, 2012 after super-storm Sandy made landfall in the state.	2012-10-30	New Jersey — Seaside Heights N.J.
October 22, 2012 - Sandy develops into a tropical storm in the Caribbean Sea.	2012-10-22	Caribbean Sea
October 24, 2012 - Hurricane Sandy makes landfall near Kingston, Jamaica, with winds of 80 mph.	2012-10-24	Kingston Jamaica
By Patrick Clark September 26, 2013 Business owners pile muddy furniture outside their building off Canon Avenue in Manitou Springs, Colo.	2013-09-26	Manitou Springs Colo.

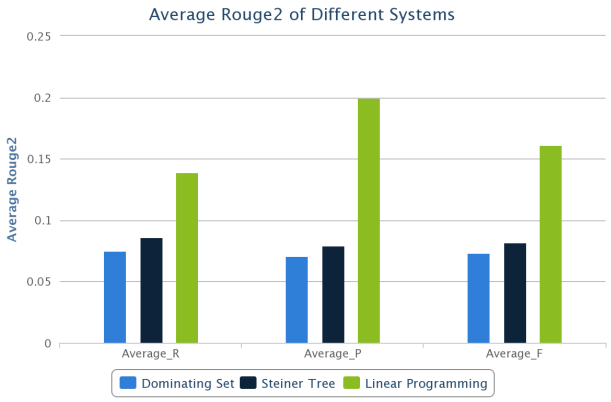


Fig. 2. Average Recall, Precision, F-1 of ROUGE-2.

global storylines for the three hurricane disasters, which are compared with system-generated storylines using ROUGE [31] toolkit (version 1.5.5). ROUGE is widely applied by DUC for summarization performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-SU.

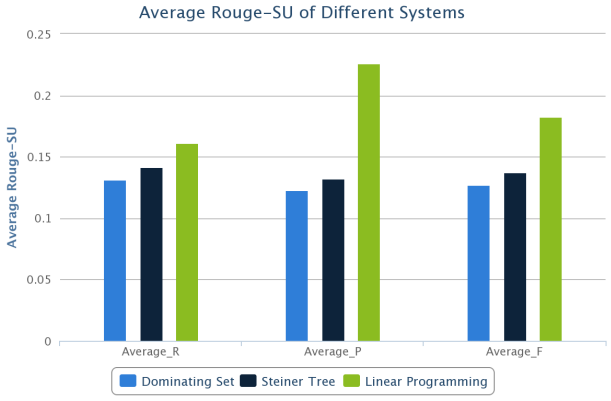
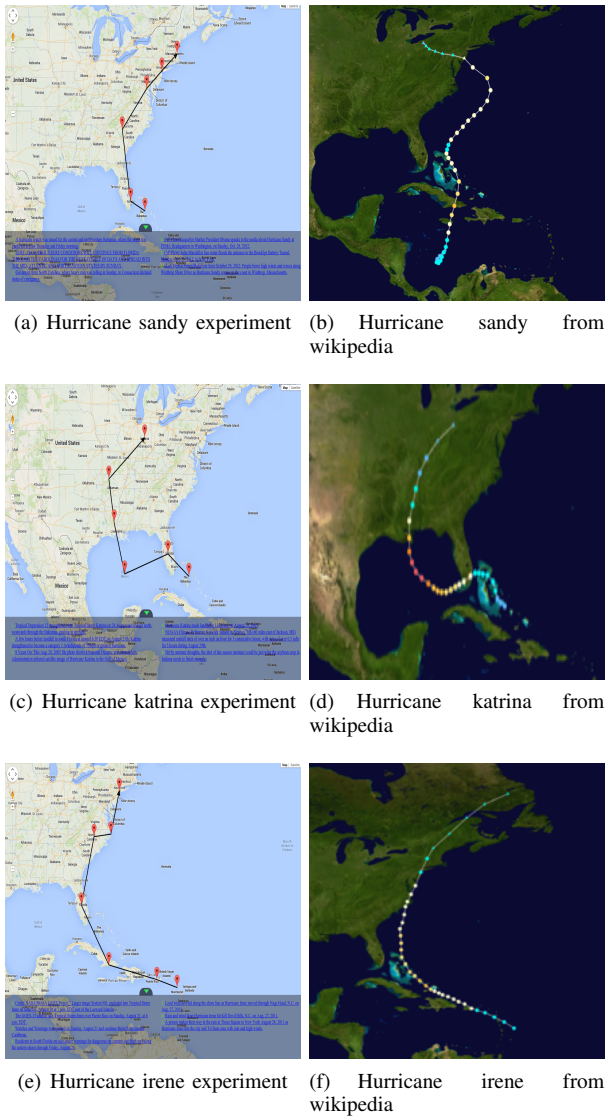


Fig. 3. Average Recall, Precision, F-1 of ROUGE-SU4.



(a) Hurricane sandy experiment (b) Hurricane sandy from wikipedia

(c) Hurricane katrina experiment (d) Hurricane katrina from wikipedia

(e) Hurricane irene experiment (f) Hurricane irene from wikipedia

Fig. 4. Experimental result of Hurricane Sandy, Katrina and Irene compared to Wikipedia.

ROUGE-N is an n-gram recall computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}, \quad (11)$$

where n is the length of the n-gram, and ref stands for the reference summaries. $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries, and $\text{Count}(\text{gram}_n)$ is the number of n-grams in the reference summaries. ROUGE-SU4 is based on skip-bigram plus unigram, where skip length is 4.

We compare the global storylines generated by our proposed method considering geo-spatial information with the results from the following methods:

1) The Steiner tree based storyline generation [9], which does not consider geo-spatial information;
2) The Dominating set based summarization method [5], which is a standard multi-document summarization

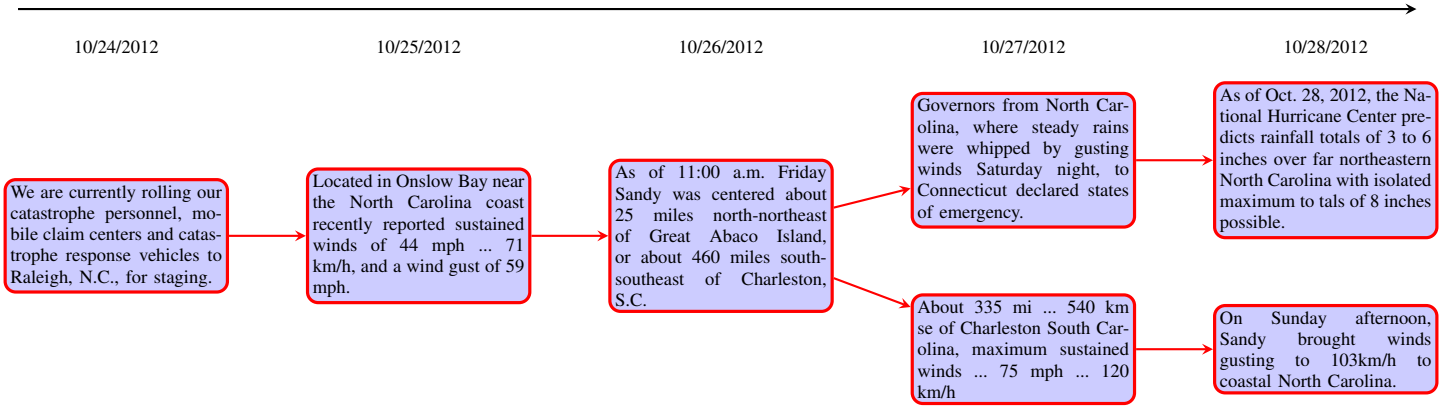


Fig. 5. An illustrative example of the local storyline for the area of the Carolinas during Hurricane Sandy.

method.

Figure 2 and Figure 3 show the performance comparison of the three methods using ROUGE-2 and ROUGE-SU4, respectively.

We can observe that the Streiner tree based storyline generation method outperforms the pure multi-document summarization method that does not incorporate the temporal information. Our proposed storyline generation method, which considers both the temporal and spatial information, performs the best among all three methods.

C. A Case Study

A case study is conducted to demonstrate the effectiveness of the storylines generated using our proposed method. We draw the global storyline generated by our proposed method using Google Map API (shown on the left sub-figures in Figure 4) and compare it with the storm paths downloaded from Wikipedia (shown on the right sub-figures in Figure 4).

We can observe that the paths in our generated storylines are similar with the ground truth. The differences are: 1) in addition to show the real paths, our generated storylines can reflect more information about how the hurricanes affect different areas; and 2) the generated storylines not only shows how hurricanes move but also present text descriptions about the status updates and damages they cause along the movement. With the geo-temporal storyline, users can easily capture the overall situation evolution of a disaster.

Figure 5 shows an illustrative example of a local storyline when we are interested in a specific area like Carolina during Hurricane Sandy. We can see how Hurricane Sandy affects the area during the period of time and covering different topics like wind and rain.

VIII. Conclusion

In this paper, we present a storyline framework for summarizing multiple disaster-related documents to generate a two-layer hierarchical storyline to improve situation awareness during or after disasters. We organize the storyline as a two layer hierarchical structure to naturally describe a large-scale disaster. Especially both temporal and spatial factors are considered in the global storyline generation capturing spatial evolution of the disaster over time.

In our future work, we will first explore more complicated evolution structures of different disaster types for storyline generation. We will also extend our framework to incorporate more disaster types like earthquakes and other man-made disasters. To make our system more practical in a real-time disaster environment, we will include Twitter streams as another data source.

Acknowledgment

The work was supported in part by the National Science Foundation under grants HRD-0833093, CNS-1126619, and IIS-1213026, the U.S. Department of Homeland Security under grant Award Number 2010-ST-062000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and Army Research Ofce under grant number W911NF-1010366 and W911NF-12-1-0431.

References

[1] L. Li and T. Li, "An empirical study of ontology-based multi-document summarization in disaster management," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 2, 2014.

[2] J. Li, L. Li, and T. Li, "Multi-document summarization via submodularity," *Applied Intelligence*, vol. 37, no. 3, pp. 420–430, 2012.

[3] D. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, no. 6, pp. 919–938, 2004.

[4] H. Saggion, K. Bontcheva, and H. Cunningham, "Robust generic and query-based summarisation," in *EACL*, 2003.

[5] C. Shen and T. Li, "Multi-document summarization via the minimum dominating set," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 984–992.

[6] F. Wei, W. Li, Q. Lu, and Y. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," in *Proceedings of SIGIR*, 2008.

[7] J. Allan, *Topic detection and tracking: event-based information organization*. Springer, 2002, vol. 12.

[8] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 899–908.

[9] D. Wang, T. Li, and M. Ogihara, "Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs." in *AAAI*, 2012.

[10] I. Mani, "Automatic summarization," *Computational Linguistics*, vol. 28, no. 2, 2001.

[11] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational linguistics*, vol. 28, no. 4, pp. 399–408, 2002.

[12] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," in *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*. Association for Computational Linguistics, 2000, pp. 21–30.

[13] G. Erkan and D. R. Radev, "Lexpagerank: Prestige in multi-document text summarization." in *EMNLP*, vol. 4, 2004, pp. 365–371.

[14] L. Li, D. Wang, C. Shen, and T. Li, "Ontology-enriched multi-document summarization in disaster management," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 819–820.

[15] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 307–314.

[16] C. Shen, T. Li, and C. H. Ding, "Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases." in *AAAI*, 2011.

[17] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas, "Relevance models for topic detection and tracking," in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 115–121.

[18] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Information Retrieval*, vol. 7, no. 3-4, pp. 347–368, 2004.

[19] O. Alonso, R. Baeza-Yates, and M. Gertz, "Effectiveness of temporal snippets," in *WSSP Workshop at the World Wide Web ConferenceWWW*, vol. 9, 2009.

[20] C. Lin, C. Lin, J. Li, D. Wang, Y. Chen, and T. Li, "Generating event storylines from microblogs," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12, 2012, pp. 175–184.

[21] Y. Jiang, C.-S. Perng, and T. Li, "Natural event summarization," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 765–774.

[22] E. A. Inc, "Webeoc," http://www.esi911.com/home.

[23] NC4, "E-teams," http://www.nc4.us/ETeam.php.

[24] Ushahidi, "http://www.ushahidi.com/," 2012.

[25] GeoVISTA, http://www.geovista.psu.edu.

[26] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, and S.-C. Chen, "Data mining meets the needs of disaster information management," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 5, pp. 451–464, 2013.

[27] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[28] A. X. Chang and C. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *LREC*, 2012, pp. 3735–3740.

[29] D. Johnson, "Approximation algorithms for combinatorial problems," in *Proceedings of STOC*, 1973.

[30] M. Charikar, C. Chekuri, T.-y. Cheung, Z. Dai, A. Goel, S. Guha, and M. Li, "Approximation algorithms for directed steiner problems," in *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1998, pp. 192–200.

[31] C. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of HLT-NAACL*, 2003.

Expert Systems with Applications 41 (2014) 5115–5124

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Social network user influence sense-making and dynamics prediction

Jingxuan Li [a,c], Wei Peng [b], Tao Li [a,*], Tong Sun [b], Qianmu Li [c], Jian Xu [c]

[a] School of Computer Science, Florida International University, 11200 SW 8th St, Miami, FL, USA
[b] Xerox Innovation Group, 800 Phillips Rd, Webster, NY, USA
[c] Nanjing University of Science and Technology, Nanjing 210094, PR China

ARTICLE INFO

Keywords:
Social media
Influential users
Dynamic information diffusion
Continuous-Time Markov Process

ABSTRACT

Identifying influential users and predicting their "network impact" on social networks have attracted tremendous interest from both academia and industry. Various definitions of "influence" and many methods for calculating influence scores have been provided for different empirical purposes and they often lack the in-depth analysis of the "characteristics" of the output influence. In addition, most of the developed algorithms and tools are mainly dependent on the static network structure instead of the dynamic diffusion process over the network, and are thus essentially based on descriptive models instead of predictive models. Consequently, very few existing works consider the dynamic propagation of influence in continuous time due to infinite steps for simulation. In this paper, we provide an evaluation framework to systematically measure the "characteristics" of the influence from the following three dimensions: (i) *Monomorphism vs. Polymorphism*; (ii) *High Latency vs. Low Latency*; and (iii) *Information Inventor vs. Information Spreader*. We propose a dynamic information propagation model based on *Continuous-Time Markov Process* to predict the influence dynamics of social network users, where the nodes in the propagation sequences are the users, and the edges connect users who refer to the same *topic* contiguously on time. Finally we present a comprehensive empirical study on a large-scale twitter dataset to compare the influence metrics within our proposed evaluation framework. Experimental results validate our ideas and demonstrate the prediction performance of our proposed algorithms.

1. Introduction

1.1. Identifying influential users

Social network analysis has been gaining attention from different domains, including economics, anthropology, biology, social psychology, physics, etc.

The rapid growth of the online social network sites (e.g. Facebook, Twitter, LinkedIn, and Google+) and their publicly available data acquiring API has led the prosperity of social network analysis research these days. One of most popular topics of the social network analysis is identifying influential users and their "network impact". Knowing the influence of users and being able to predict it can be leveraged for many applications. The most famous application to researchers and marketers is viral marketing (Domingos & Richardson, 2001; Kempe, Kleinberg, & Tardos, 2003; Richardson & Domingos, 2002), which aims at targeting a group of influential users to maximize the marketing campaign ROI (Return of Investment). Other interesting applications include search (Adamic & Adar, 2005), expertise/tweets recommendation (Song, Tseng, Lin, & Sun, 2006, 2007, 2010), trust/information propagation (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004; Golbeck & Hendler, 2006), and customer handling prioritization in social customer relationship management.

1.2. Limitations of current research efforts

There are two main limitations of current research efforts on identifying influential users in social network analysis: one is on the characteristics of influence, and the other is on the influence models and measures.

1.2.1. Characteristics of influence

Various definitions of "influence" and many methods for calculating influence scores have been provided for their own empirical purposes, or applications. Since they often lack the in-depth analysis of the "characteristics" of the output influence. it is difficult to adapt or choose them for other applications.

* Corresponding author. Tel.: +1 305 348 6036; fax: +1 305 348 3549.
E-mail addresses: jli003@cs.fiu.edu (J. Li), wei.peng@xerox.com (W. Peng), taoli@cs.fiu.edu (T. Li), tong.sun@xerox.com (T. Sun), qianmu@njust.edu.cn (Q. Li), dolphin.xu@njust.edu.cn (J. Xu).

http://dx.doi.org/10.1016/j.eswa.2014.02.038
0957-4174/

1.2.2. Influence models and measures

Currently most applications and tools compute user influence based on their static network properties, such as, the number of friends/followers in the social graph, the number of posted tweets/received retweets/mentions/replies in the activity graph, or users' centrality (e.g. PageRank, Betweeness-centrality, etc.).

A few works investigate adoption behaviors of social network users as the dynamic influence propagation [1] or diffusion process (Rogers, 2003). The adoption behaviors refer to some activities or topics (tweets, products, Hashtags, URLs, etc.) shared among users implicitly and explicitly such as users forwarding a message to their friends, recommending a product to others, joining some groups with the similar musical favor, and posting messages about the same topics, etc. According to the diffusion theory, the information cascades from social leaders to followers. In most diffusion models, propagators have certain probabilities to influence their receivers, and the receivers also have certain thresholds to be influenced. Finding the social leaders or the users who can maximize the influence coverage in the network is the major goal of most diffusion models.

Some drawbacks of existing social network influence models based on either static networks or the "influence maximization" diffusion process are: (1) The static influence scores are not actionable for users. For example, marketers do not know what will be the difference if targeting users with influence scores of 30 or 80. (2) Most existing models are descriptive models rather than predictive models. For example, the number of friends or the centrality score of a given user describes his/her underlying network connectivity. The number of tweets that a user posted or get retweeted indicates the trust/insterest that his/her followers have on his/her tweets. All these measures/models are descriptive and very few models are able to predict users' future influence. (3) Existing "influence maximization" diffusion process is often modeled by discrete-time models such as Independent Cascade Model or Linear Threshold Model. Because the real world diffusion process is continuous-time, it is difficult to define an appropriate time step t for discrete-time models.

1.3. Content of the paper

The aforesaid limitations motivate our study on social network user influence and dynamics prediction in this paper. In particular, to address the first limitation, we take an initial step to introduce three dimensions of influence: (i). *Monomorphism vs. Polymorphism*; (ii). *High Latency vs. Low Latency*; and (iii). *Information Inventor vs. Information Spreader*, for understanding the characteristics of influential users calculated from various methods. These three dimensions provide an evaluation framework to systematically measure the influence.

To address the second limitation, we propose a dynamic information diffusion model based on the *Continuous-Time Markov Process* (CTMP) to predict the influence dynamics of social network users. CTMP assumes that the number of activations from a given node is following an exponential distribution over the time. This can be often seen in the real-world data (Kwak, Lee, Park, & Moon, 2010). Fig. 1 shows that the average number of topic adoptions decreases exponentially over the time. Hashtags receive more adoptions compared with URLs, and the number of Hashtag adoptions decreases more slowly. Furthermore, transition rates q are calculated and treated as the transition probabilities (or activation probability) of the embedded Markov chain in CTMP. Then the transition probability $P(t)$ can be computed from q, given any time t. In this paper, the nodes in the propagation sequences are the users, and the edges connect users who refer to the same topic contiguously on time. Topics here particularly refer to Hashtags (expressed as # followed by a word) and short URLs (e.g. bit.ly, TinyURL, etc.) on twitter, which is one of the most popular microblog services, was launched since July 13, 2006. Hashtags and URLs are both unique identifiers tagging distinct tweets with certain "topic" labels. We regard the temporal sequences of Hashtags and URLs as the diffusion paths, where the topics are reposted subsequently. Although retweeting is not included in our paper as a diffusion approach, it is implicitly considered because the retweets would usually contain the same Hashtags and URLs as in the original tweets. Our experimental results on a large-scale twitter dataset show that our proposed diffusion model outperforms other influence models for viral marketing. It also demonstrates a promising prediction performance on estimating the number of influenced users within a given time.

1.4. Paper contribution and organization

A preliminary study of the work has appeared at the 15th Asia–Pacific Web Conference in 2013 (Li, Peng, Li, & Sun, 2013). In that conference paper, the study focuses on the proposed influence model – IDM-CTMP, and shows its advantages over two baselines, which are not necessarily continuous-time models. In this journal submission, (1) we propose three "dimensions" of users' influence in the social network to help others understand different aspects of influence; (2) we conducted comprehensive experiment to systematically measure users' influence and compare different influence models over three proposed dimensions; (3) two heuristic continuous-time influence models are defined as baselines to further show the advantages of our proposed model.

In summary, the contributions of this paper are listed below.

1. We introduce three dimensions on application perspectives and provide an evaluation framework to systematically measure the influence and compare different influence models (See Section 5.3).
2. Comprehensive experiments are conducted on various extracted networks (mentions, retweets, replies), as well as temporal propagation paths from the large-scale twitter data (See Section 5).
3. Two heuristic influence models considering the topic diffusion in continuous time are defined as baselines (See Section 3) to highlight the strengths of our proposed dynamic information diffusion model based on the *Continuous-Time Markov Process*.

The remainder of this paper is organized as follows. Before discussing about any influence models, we propose three dimensions of social influence in Section 2. After, in Section 3, we first give the definition of the temporal influence network, introduce some existing influence models, and propose two heuristic dynamic influence models. In Section 4, we propose an information diffusion model based on the *Continuous-Time Markov Process*. Experimental results are demonstrated in Section 5. In particular, we discuss the three dimensions of influence and present a comprehensive empirical study on a large-scale twitter dataset to compare the influence metrics (including both the dynamic influence metrics and well-known static influence metrics) within our proposed evaluation framework in Section 5.3. We evaluate the prediction power of our proposed information diffusion model in Section 5.4. Related work on influence modeling is reviewed in Section 6. Finally Section 7 concludes the paper.

2. Three dimensions of influence

Everyone is talking about how to identify users with high influence, because it is believed that influential users can help with

[1] In this paper, we use "information/influence propagation", "information/influence diffusion", and "information cascade", interchangeably to represent the same concept.

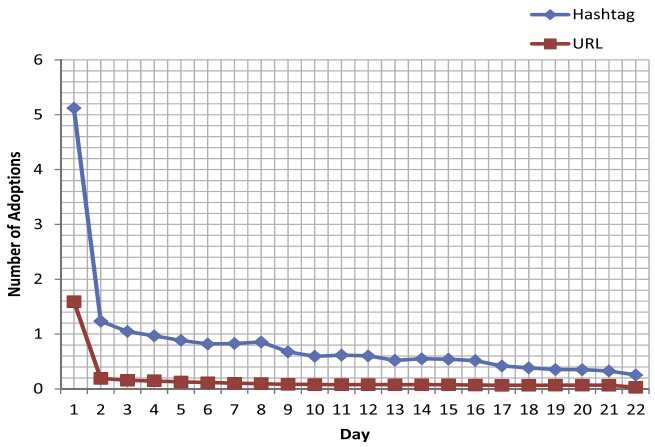


Fig. 1. The average number of topic adoptions over the time on our twitter dataset.

many applications, e.g., Viral Marketing. Question is: what does influence exactly mean in the context of social media?

In this section, social media users' influence is discussed from three dimensions.

2.1. Monomorphism VS. Polymorphism

The concept of Monomorphism vs. Polymorphism is borrowed from the diffusion of innovations (Rogers, 2003). In our paper, users with high monomorphism usually focus on a constant set of topics, while users with high polymorphism post a variety of topics over the time. Knowing this property of social media users could benefit applications with different purposes. For example, high monomorphism influencers should be ranked higher than high polymorphism influencers in expert recommendation applications. However, the high polymorphism influencers would be more desirable to users aiming for general information gathering.

To determine whether a user is monomorphic or polymorphic is difficult, nevertheless, we suppose if a user is monomorphic, his/her posted topics should be similar across two different time periods; on the other hand, if a user is polymorphic, his/her topics would be different across two different time periods. In our experiment, we compare two time periods – the 12-day training period and 10-day testing period specified in previous sections. For each user, two topic vectors (consisted of Hashtags/URLs) from these 12-day data and 10-day data are extracted. Then the cosine similarity is measured between these two topic vectors as the topic similarity. The high topic similarity indicates the high monomorphism.

2.2. High Latency VS. Low Latency

As for the second dimension - *High Latency vs. Low Latency*, here latency means, once a user posted a topic, the time delay before the next posts about the same topic would appear. Influencers with a low latency often receive immediate topic "adoption". Thus they should be picked as viral marketing "seeds" when marketers want to quickly test customer response.

Different topics may result in different adoption latencies. For example, influencers interested in "Machine Learning" might generally have a higher latency than ones interested in "Justin Bieber". Instead of regarding the average time difference between the user's original post and the next topic adoption as the latency, which may be highly affected by the type of topic, we define the latency as follows:

$$Latency(v) = |\{\tau_1 | aveDiff(\tau_1) < firstDiff(v,\tau_1), \tau_1 \in T(v)\}| - |\{\tau_2 | aveDiff(\tau_2) >= firstDiff(v,\tau_2), \tau_2 \in T(v)\}|, \quad (1)$$

where $T(v)$ denotes all the topics posted by user v, τ_1 and τ_2 represent topics from $T(v)$, $aveDiff(\tau_1)$ is the average interarrival time between every pair of neighboring posts about topic τ_1, and $firstDiff(v,\tau_1)$ indicates the time taken for the follower to make the first adoption right after v posts topic τ_1. A large value of $Latency(v)$ means a high latency.

2.3. Information Inventor VS. Information Spreader

The third dimension about *Information Inventor vs. Information Spreader*, is to measure the diffusion power of influencers. Information inventors are innovators who are usually the information source, the first group of people to adopt products/brands, or new trend leaders. Information Spreaders are people who are able to spread topics to a lot of social media users. It is quite obvious that the third property dimension of influential users is very useful for viral marketing. The targeted seed users for viral marketing should be both information inventors and information spreaders.

Rather than identifying who are the information inventors and information spreaders, we measure the inventing ability of each user as:

$$Inv(v) = \frac{\#(new\ topics\ started\ by\ v)}{\#(tweets\ by\ v)}. \quad (2)$$

The term "new topics" indicates the Hashtags/URLs that are first posted in twitter. The spreading ability can be computed by using the definition of Time-Window Diffusion Size in Section 3.2.1.

3. Influence network and influence models

3.1. Influence network

A social graph can be denoted as $G(V,E)$, where V represents social network users, and E is the set of edges/relations between users. The follower-followee graph is one type of social graphs, where the edges indicate following relations. *Activity graphs* are another type of social graphs, which are extracted from users tweeting behaviors. The typical twitter activity graphs are tweet-retweet graph, tweet-reply graph, and mention-mentioned graph. In our paper, we run well-known user influence models (e.g., degree-centrality, PageRank) on these three activity graphs in our comparative study. Both the follower-followee graph and activity graphs are directional *Influence Networks*, where the influence flows from users to people who follow them, or people who retweet their tweets, or people who reply their tweets, or people who mention their names. The influence network can be denoted as $G(V, E_{influence})$, where V denotes social network users, and the edge $V_i \leftarrow V_j$ in $E_{influence}$ means V_j is influenced by V_i.

The above networks can be viewed as static networks, which do not demonstrate the dynamic propagation process over the time. In order to analyze how topics are passing on social networks progressively, we construct a temporal influence network by considering the continuous time. Given a Hashtag/URL (*topic*), a group of users can be ordered based on the time when they post this topic. As shown in Fig. 2, user i is linked to user j if they post the same topic contiguously and user j follows/friend with user i. The number on the top of each arrow is the time taken to transfer a topic from a user to another user.

Definition 1 (*Temporal influence network*). The temporal influence network is $G(V,E,T(E))$, where $V = \{V_0, V_1, \ldots, V_n\}$ contains all users who posted at least one Hashtag or URL, $E = \{V_i \leftarrow V_j | V_i$ posted a topic earlier than $V_j\}$, where edges can be constrained to only exist between followers and followees or between friends. So the propagation is along the paths from

followees to followers over continuous time. The function $T(V_i \leftarrow V_j) = \{t_{ij}^0, t_{ij}^1, \ldots, t_{ij}^l\}$. $t_{ij}^m \in \{t_{ij}^0, t_{ij}^1, \ldots, t_{ij}^l\}$ is the time difference between user i posting a topic and user j posting the same topic.

There can be multiple entries in $T(V_i \leftarrow V_j)$ since user i and user j can post the same set of topics or one topic at multiple times. Note that we aggregate all topics together to form this temporal influencer network in this paper. One natural extension is to categorize these Hashtags/URLs into topics so that topic-sensitive influential users can be computed from each topic-dependent network G_{topic_i}.

3.2. Influence models

Degree Centrality and PageRank, as two most well-accepted influence models, are computed on static networks. The static networks here refer to the three activity networks we specified in previous subsection. The Degree Centrality is defined as the number of inlinks incident upon a node/vertex. The essential idea of PageRank is to define a link analysis method to evaluate a user's influence, so that not only the immediate information flow is incorporated, but also the information flow after that would be considered. According to PageRank, a user is "authoritative" if he/she has a lot of inlinks from other "authoritative" users.

Different from the above mentioned influencer models, we propose two straightforward dynamic influence models on the temporal influence network to incorporate the temporal information.

3.2.1. Time-window diffusion size

Definition 2 (*Time-window diffusion size*). The diffusion size of a user u over a topic c, $DS_{u,c}$, is the number of other users posting the same topic c after user u within a pre-defined time range. The aggregated diffusion size over all the topics of a user is $DS_u = \sum_c DS_{u,c}$.

It is worth noticing that the influence computed here is based on a pre-defined time range, specifically, this method grants us the ability of identifying the comparative influential users within a pre-defined time range. We can see that the users with a large time-window diffusion size tend to post topics at the beginning of fast and large cascade of the topics.

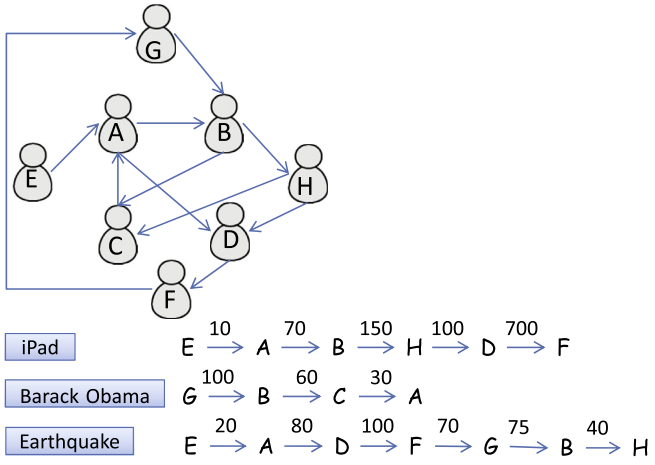


Fig. 2. The example of temporal influence network construction.

3.2.2. Temporal closeness centrality

Definition 3 (*Temporal distance*). The temporal distance $d_{temporal}(V_i, V_j)$ between two users V_i and V_j is the least time difference $min(T(V_i \leftarrow V_j))$ w.r.t. the set of topics posted by both V_i and V_j where $T(V_i \leftarrow V_j)$ is defined in Definition 1.

In order to measure the *reach-ability* of a user, the temporal closeness centrality is given by:

$$TCC_u = \frac{\sum_{v \in V \setminus u} d_{temporal}(u, v)}{n - 1}, \tag{3}$$

where n is the number of all users in the temporal influence network. It is worth pointing out that: sometimes a user u never goes to v since no topic diffuses from user u to v. In such a case, we treat the temporal distance between u and v as $n \cdot Max_{i,j \in V, i \neq j} T\{V_i \leftarrow V_j\}$. Users with low temporal closeness centrality often post topics close to fast and large cascade of the topics.

4. Information diffusion model based on Continuous-Time Markov Process

The aforementioned influence models are either based on static activity networks or descriptive models (instead of predictive models) building on the temporal influence network. The descriptive models answer questions such as "How many followers that user A has?" and "How many followers post the topic 'ipad' after user A?", etc. In this section, we introduce our proposed predictive Information Diffusion Model based on Continuous-Time Markov Process, abbreviated as IDM-CTMP for convenience. IDM-CTMP is able to answer the following question, "In the next month, how many users would post the topic 'ipad' estimably if user A posts it now.", or even a harder question "In order to make a maximal number of people to talk about our product in the next week, who are the seed users we should target?". Note the influential users discovered by IDM-CTMP maximize not only the information *coverage*, but also the *rate* of information cascade given a certain period of time.

4.1. Model formulation

A trending topic (a Hashtag/URL) is propagated by social network users within the temporal influence network defined in Definition 1. Suppose $X(t)$ denotes the user who posts a specific topic at time point t, $X = X(t), t \geqslant 0$ forms a Continuous-Time Markov Process (CTMP) (Anderson & James, 1991), in which the user who will discuss this topic next only depends on the current user given the whole history of the topic propagation. Formally, this markov property can be defined by:

$$\begin{aligned} P_{ij}(t) &= P\{X(t+\gamma) \\ &= j | X(\gamma) = i,\ X(\mu) = x(\mu),\ 0 \leqslant \mu < \gamma\} \\ &= P\{X(t+\gamma) = j | X(\gamma) = i\}, \end{aligned} \tag{4}$$

where $P_{ij}(t)$ is the transition probability from i to j within time t, i is the current user who discusses the trending topic, j is the next user who posts the topic following i, and $x(\mu)$ denotes the history of the topic propagation before the time point γ. We assume that the transition probability $P_{ij}(t)$ does not depend on the actual starting time of the propagation process, thus the CTMP is time-homogeneous:

$$\begin{aligned} P_{ij}(t) &= P\{X(t+\gamma) = j | X(\gamma) = i\} \\ &= P\{X(t) = j | X(0) = i\}. \end{aligned} \tag{5}$$

In order to estimate the diffusion size of user i given a pre-defined time window t, we need to compute the transition

probability from user i to all the other users, then determine the number of users being affected by i at the end of the time window. The diffusion size of user i over time t based on CTMP can be defined as

$$DS_{i,t} = \sum_j P_{ij}(t) \cdot n_i, \tag{6}$$

where n_i is the number of times that user i occurs at time t. It can be estimated by supposing that it linearly increments on t. However, it is impractical to estimate the transition probability matrix $P(t)$ with infinite possible t. Thus instead of estimating $P(t)$ directly, we calculate the transition rate matrix Q, and then $P(t)$ can be estimated from Q.

4.2. Estimation of transition rate matrix

The transition rate matrix Q is also called the infinitesimal generator of the Continuous-Time Markov Process (Dynkin, 1965). It is defined as the derivative of $P(t)$ when t goes to 0. The entry q_{ij} is the transition rate to propagate a topic from user i to user j. The sum of the rows in Q is zero, with $\sum_{j,j\neq i} q_{ij} = -q_{ii}$.

$$q_{ij} = \lim_{t\to 0} \frac{P_{ij}(t)}{t} = P'_{ij}(0) \quad (i \neq j). \tag{7}$$

Note that q_{ij} reflects a change in the transition probability from user i to user j. q_i, namely out-user transition rate in this paper, is equal to $-q_{ii}$. It indicates the rate of user i propagating topics to any other users. As shown in Fig. 1, the average number of topic adoptions decreases exponentially over the time. Thus, in order to compute q_i, we assume that the time for user i to propagate a topic to all the other users is following an exponential distribution as observed for many users in our data, where the rate parameter is q_i. It is known that the expected value of an exponentially distributed random variable T_i (in this case, the topic propagation time for user i) with rate parameter q_i is given by Feller (2008):

$$E[T_i] = \frac{1}{q_i}. \tag{8}$$

Thus q_i is one divided by the mean of $\cup_j(T(V_i \leftarrow V_j))$, which is defined in the temporal influence network.

According to the theory of Continuous-Time Markov Process, if a propagation occurs on user i, the probability that the other user j would post the topic forms an embedded Markov chain (Karlin & Taylor, 1975). The transition probability is S_{ij}, and $\sum_j S_{ij} = 1$ $(i \neq j)$ and $S_{ii} = 0$. One important property is that $q_{ij} = q_i S_{ij}$. Then, the transition rate from user i to j can be estimated by:

$$q_{ij} = \sum_m q_i^2 \cdot exp(-q_i \cdot t_{ij}^m), \tag{9}$$

where m is the number of topics diffused from use i to j, and t_{ij}^m denotes the transition time from user i to j on the mth topic.

4.3. Estimation of transition probability matrix

Now we obtain all the entries of the transition rate matrix Q. Next, we will specify how to derive the transition probability matrix $P(t)$. The well accepted Kolmogorov's Backward Equations (Gardiner, 1985) in the Continuous-Time Markov Process can be utilized:

$$P'_{ij}(t) = q_i \times \sum_{i\neq k} P_{ik}(t) \times P_{kj}(t) - q_i \times P_{ij}(t). \tag{10}$$

By performing some algebraic operations, the above equation can be written as the following matrix form:

$$P'(t) = QP(t). \tag{11}$$

The general solution for this equation is given by:

$$P(t) = e^{Qt}. \tag{12}$$

$P(t)$ is a stochastic, irreducible matrix for any time t. We approximate it using Taylor expansion, so that $P(t)$ can be estimated by Huang, Yang, Huang, and Ng (2004):

$$P(t) = e^{Qt} = \lim_{n\to\infty}(I + Qt/n)^n. \tag{13}$$

We raise the power of $(I + Qt/n)$ to a sufficiently large n.

5. Experiment

5.1. The dataset description

Twitter provides Streaming APIs which allow high-throughput near-realtime access to various subsets of twitter data. It samples the statuses (including the tweets and the authors) from the Firehose stream of public statuses which is the full feed of all public tweets. Our paper uses Twitter Gardenhose streaming API, which is said to sample 10% of all public tweets. Hashtags beginning with # in tweets represent keywords or topics. URLs add more detailed topic information to tweets, shortened via the services such as bit.ly or tinyurl.com. Hashtags and URLs enable twitter users to create and follow a thread of discussion. They are regarded as unique identifiable topics in our paper. Hashtags and URLs of each tweet can be extracted from its metadata fields, embedded in the crawled raw twitter data.

We continuously collected 22-day twitter data, ranging from March 2 to March 24, 2011. The first 12-day data is used for our training purpose, and the remaining 10-day data is for testing and validation. We removed all non-English tweets to focus on only English twitter world. Tweets posted by users with less than 20 followers are also removed. These twitter users post close to 10% of all tweets, and supposedly they are very less likely to be influential users. Finally we have a total of 48,113,490 tweet records (a tweet record may include the tweet and the corresponding retweets, thus more than one tweet) in the 12-day training data. As for the 10-day test data, we also removed tweets without Hashtags or URLs, which result in 27,237,631 tweet records, about 10% of original data. We also filter the tweets with more than three Hashtags or URLs, which tend to be spam tweets as introduced in Kwak et al. (2010). Finally, we obtained totally 78,858,046 tweets, in which there are 9,431,404 unique users, 3,209,330 Hashtags, and 21,107,164 URLs.

5.2. Correlations between different metrics

In order to understand the difference between various influence metrics: the number of mentions, the number of replies, the number of retweets, the PageRank of mentions, the PageRank of replies, the PageRank of retweets, Time Window Diffusion Size, Temporal Closeness, and IDM-CTMP, we measured the overlap and Spearman's correlation between every two influential user rank lists obtained from the above metrics. Although our proposed IDM-CTMP is a dynamic metric which outputs different user rank lists given different time ranges, in our comparative study we fix the time range to be 10 days (from day 13 to day 22). The social graph is not added as a constraint on IDM-CTMP in the experiment to test its performance even without any network structure information. The empirical result shows little correlation between most pairs of rank lists except the correlation between the number/PageRank of mentions and the number/PageRank of replies, and the

correlation between the Time Window Diffusion Size and Temporal Closeness.

5.3. An evaluation framework to measure influence: three dimensions of influencer properties

To understand the properties of various user influence metrics, we conducted experiments to systematically compare them on three dimensions: (i). *Monomorphism vs. Polymorphism*; (ii). *High Latency vs. Low Latency*; and (iii). *Information Inventor vs. Information Spreader*.

5.3.1. Monomorphism VS. Polymorphism

Fig. 3 shows that a large number of users have ever used a limited number of Hashtags/URLs, while only a few users utilized a large quantity of Hashtags/URLs. That means a very few twitter users post a wide range of topics.

As we introduced in Section 2.1, the cosine similarity of two topic vectors from the first 12-day training period and 10-day testing period for a specific user is used as the topic similarity. The high topic similarity indicates that this user has high monomorphism.

To compare 9 different user influence rank lists, we choose the top 10,000 users and the bottom 10,000 users from each rank list. The average topic similarity of the top 10,000 and the bottom 10,000 users for each rank list across the specified two time periods is computed.

The comparison results are presented in Fig. 4. From the results, we can observe that users with high degree centralities of mentions and retweets have higher monomorphism. Especially the gap between the top 10,000 and the bottom 10,000 users based on the number of mentions is the largest. Looking into the data, users with high mentions like justinbieber, charliesheen, JonasBrothers, and XSTROLOGY, usually focus on a constant set of topics. Note that the top ranked users by our proposed method IDM-CTMP have relatively high polymorphism. Thus they tend to post a variety of topics over the time. In order to explain this phenomenon, we can think of the characteristics of the dataset and the IDM-CTMP method. The dataset is sampled from the real world tweets, which contain all kinds of topics, thus it covers topics from different areas. IDM-CTMP aims at identifying influential users who are able to diffuse topics to many other users no matter from which areas those topics come. Therefore, polymorphic users tend to be ranked higher by IDM-CTMP since they can diffuse more topics to more users in the social network.

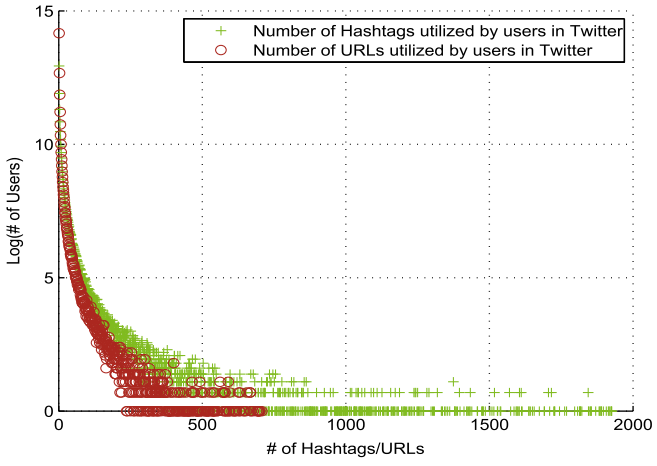


Fig. 3. Number of Hashtags/URLs utilized by users in twitter.

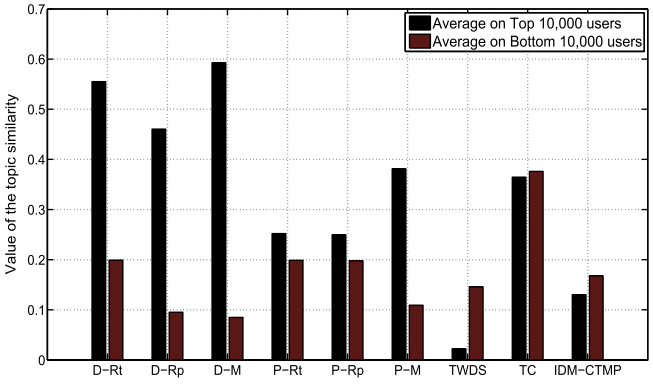


Fig. 4. The average topic similarity of top 10,000 users and bottom 10,000 users from 9 user influence rank lists. "D" denotes Degree, "P" denotes Pagerank, "Rt" denotes Retweet, "Rp" denotes Reply, "M" denotes Mention, "TWDS" is Time-Window Diffusion Size, and "TC" means Temporal Closeness.

5.3.2. High Latency VS. Low Latency

To measure if the top ranked influential users have a higher post diffusion latency or not, we first calculate the latency score as shown in Eq. (1) for each user. Then 10,000 users with the lowest latency are extracted to compare with the top 10,000 users from each user influence rank list using the Spearman's correlation. The correlation results are shown in Fig. 5. It shows that the top ranked influential users from IDM-CTMP have the lowest latency than other metrics. The reason behind this observation is that IDM-CTMP tries to maximize not only the diffusion coverage but also the diffusion speed.

5.3.3. Information Inventor VS. Information Spreader

Similar to previous experiments, 10,000 users with the highest information inventing ability score and 10,000 users with the highest information spreading ability score are extracted to compare with the top 10,000 users from 10 influential user rank lists using the Spearman's correlation.

The comparison results are shown in Fig. 6. In this experiment we add "INV", which is the metric defined by Eq. (2). Note all the user influence metrics are calculated from the first 12-day training data, while the ground truth is computed from 10-day test data. Thus the correlation is not 1 for "INV" and "TWDS" compared with the top 10,000 inventing ability users and the top 10,000 spreading ability users, respectively.

We can observe that influential users from all the traditional static metrics do not have high inventing ability and spreading ability. As we expected, "Inv" influencers have high inventing ability but low spreading ability. On the other hand, "TWDS" influencers have high spreading ability but low inventing ability. Our proposed method IDM-CTMP achieves both high inventing ability and high spreading ability because (1) as we described in Section 5.3.1, IDM-CTMP can identify high polymorphic users who post many different topics, it is not difficult to conclude that part of those topics are "invented" by those users with high inventing capability; (2) IDM-CTMP focuses on users who can diffuse topics to many other users, in other words, high spreading ability.

It is worth noticing that so far IDM-CTMP has shown its advantages for viral marketing application because its derived top ranked influential users tend to be innovators, obtain quick topic adoption, and spread topics widely and fast.

5.3.4. Comparisons of top ranked influential users on twitter

In this section, we list the top 10 influential users identified by various methods in Table 2. The first three columns on the left are based on existing Degree Centrality influence model over

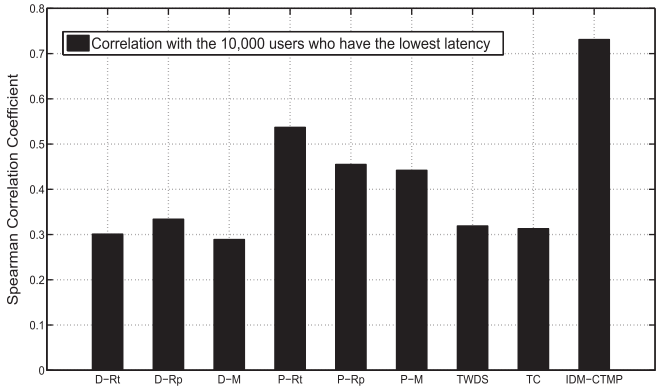


Fig. 5. The correlation between top ranked 10,000 influential users based on different influence metrics and 10,000 users with the lowest latency. "D" denotes Degree, "P" denotes Pagerank, "Rt" denotes Retweet, "Rp" denotes Reply, "M" denotes Mention, "TWDS" is Time-Window Diffusion Size, and "TC" means Temporal Closeness.

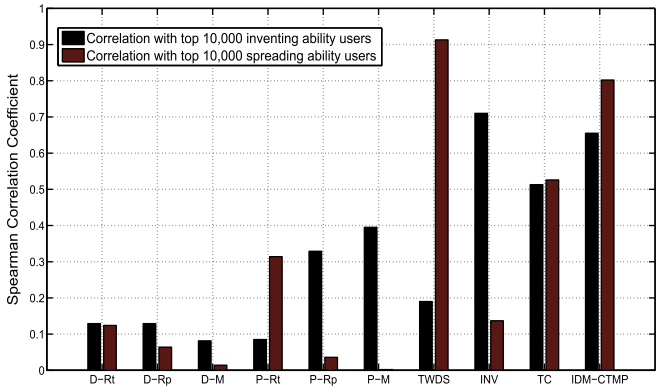


Fig. 6. The comparison results of top 10,000 users from 10 influence rank lists against top 10,000 inventing ability users and top 10,000 spreading ability users. Notice that "D" denotes Degree, "P" denotes Pagerank, "Rt" denotes Retweet, "Rp" denotes Reply, "M" denotes Mention, "TWDS" is Time-Window Diffusion Size, "INV" is Inventing Ability metric, and "TC" is Temporal Closeness.

"Retweet", "Reply" and "Mention" activities respectively; the next three columns are based on the PageRank influence model, which is then followed by the two baseline approaches: Time-Window Diffusion Size and Temporal Closeness Centrality. Our approach is listed at the last column named "IDM-CTMP". Several observations can be drawn by comparing and analyzing the top influential user lists in Table 1:

1. Degree Centrality influence model over three activity graphs (namely retweet, reply and mention) consistently picks out the "celebrities" (e.g. "justinbieber", "charliesheen"), who tend to have a large number of followers (or fans). Even though these celebrities may not "tweet" often, but even one or two tweets could still drive significant activities (i.e. retweet, reply and mention) among their immediate followers or fans; not to mention the case in which they "tweet" a lot. Meanwhile, in our experiment settings, Degree Centrality model aims to capture those users who are being frequently retweeted, replied or mentioned. As a result, the Degree Centrality model, which focuses on the first hope of influence (between the celebrity and his/her followers), is more suited for finding out influential celebrities.
2. Similar to the PageRank algorithm for ranking web pages, the PageRank influence model identifies the top influential users not just from the one-hop influence, but based on the influential users from whom they receive influence and then spread their influence in the network that beyond their direct followers. In the PageRank's "being-retweeted" and "being-replied" influence networks, the most influential users are not necessarily the celebrities with many followers, but the users who are highly interactive and responsive. For instance, "XboxSupport" and "waze" are twitter user accounts that frequently reply/being-replyed and tweet/being-retweets with their brand customers who seek help for their questions or issues. It is not difficult to imagine scenarios in which the proposed solution to certain customer issues can be spread further in the network beyond their direct followers. Another interesting observation in PageRank-Mention column is that two controversial political figures "glennbeck" and "algore" are listed among the top influential users. This could coincide with some heated political debates during our experimental data gathering time period. These political debates tend to debate on pre-known topics (such as conservative or liberal views on environment, same-sax marriage, etc.), but the influences spread across the network widely rather than among only direct followers.
3. The top influential users identified by the Time-Window Diffusion Size mostly tend to be some regular guy (or "nobody") who has a small set of followers and followees. But those guys may happen to tweet often on a small set of topics during a particular time-window and their posts get spread to many users who are not necessarily their direct followers. It is very difficult to justify that these people are "influential" in a reliable or consistent manner. Please also be noted that none of these influential users overlaps with the results of our method (IDM-CTMP), which indicates that our method is able to properly filter these users out.
4. Our method, IDM-CTMP, has the ability to identify some of Twitter accounts that representing popular news media (e.g. "washingtonpost", "nytimes", "BBCWorld") along with a well-known influential figure in technology innovation and entrepreneurship (e.g. "GuyKawasaki"). Since our experiment data was collected during March of 2011, during which there was an earthquake and tsunami event in Japan, and also coincidently Guy Kawasaki was on the promotional social-media tour for his new book "The Art of Enchantment". These noticed events have triggered some time-sensitive "new" or "unprecendented" or "bursted" topics. On the other hand, the news media also tend to be either "information innovator" (first mention the topic) or "information spreader" (diffuse news more reliably than regular people) or both at the same time. These correlations prove our method is able to detect not only time-sensitive new topics, but also considering both "innovator" and "spreader" factors. Furthermore, there is some overlap among the top influential users between Temporal Closeness and our IDM-CTMP models. Many of top influential users in Temporal Closeness column belong to news media accounts. This overlap further indicates our IDM-CTMP also favors the low latency of influence spreading.

5.4. Predicting spreading size using IDM-CTMP

In the previous subsections, we transformed IDM-CTMP into a static influence metric and compare it with other existing static influence metrics. However, IDM-CTMP is a dynamic metric and a predictive model. It is able to predict how many users would adopt some topic given a certain period of time after a user posts it. In our experiment, we first train IDM-CTMP model on the first 12-day training data, then calculate the spreading coverage of each user for each day from day 1 to day 22, including both training and testing periods.

Table 1
The top 10 influential users lists obtained by different methods.

Degree Retweet	Degree Reply	Degree Mention	PageRank Retweet	Pagerank Reply	Pagerank Mention	TimeWindow DiffusionSize	Temporal Closeness	IDM- CTMP
justinbieber	justinbieber	justinbieber	AntonioPires	XboxSupport	106andpark	bananatay	RT_com	washingtonpost
charliesheen	charliesheen	Xstrology	13eatSmith	FART_ROBOT	drdrew	Ciaramedlies	TrendingUSA	GuyKawasaki
YouTube	officialjaden	ZodiacFact	XboxSupport	bot_marley	glennbeck	FrostBight	_iMarriedaWhore	nytimes
JonasBrothers	ochocinco	EpicTweets	waze	ro_bot_dylan	simonpegg	JarrodEder	HumorORtruth	BBCWorld
XSTROLOGY	CodySimpson	TheNoteboook	GibbsRules	yodaism	algore	Neoley	SatelliteShow	kidsjoycom
AddThis	chrisbrown	algore	alfian_007	RedScareBot	priyankachopra	naira_24	GIRLTHINGS	Drudge_Report
foursquare	KimKardashian	damnitstrue	FART_ROBOT	waze	MatchupChats	anasathia	iRespectFemales	cnnbrk
chrisbrown	JASMINEVILLEGAS	charliesheen	jojokejohn	saferprint	Nea1968	kimchidiction	cnnbrk	HuffingtonPost
damnitstrue	rioferdy5	ihatequotes	saferprint	IAmJacksBot	Bikertwitts	mrPerd	eienKATTUN	Metro_TV
officialjaden	selenagomez	WowTeenagers	YouTube	for_a_dollar	BestAt	WhiteAddict	takaosaito	mashable

5.4.1. The ground truth

In order to evaluate the prediction performance of IDM-CTMP and present its feasibility in real world applications, we need to provide the ground truth. Suppose user u posts a topic τ at time t_1, and subsequently n users post the same topic τ till time t_2, then the ground-truth spreading size of u from t_1 to t_2 with regards to topic τ, denoted as $DS_{u,\tau}^{t_1\sim t_2}$, is n. For example, we know that user B first posts a topic #ipad in day 2, afterwards there are 10 users posting #ipad in day 2, and 20 users posting #ipad in day 3. Then the ground-truth spreading size of B is 0 for day 1, 10 for day 2, and 20 for day 3.

After a user's spreading sizes over different topics in a particular period of time are computed, we can obtain the average spreading size over all topics in that time period by dividing the number of involved topics:

$$DS_u^{t_1\sim t_2} = \frac{\sum_\tau DS_{u,\tau}^{t_1\sim t_2}}{\#(\tau)}, \tag{14}$$

where DS is the spreading size, u denotes a user, $t_1 \sim t_2$ indicates a time window, and τ is a topic.

5.4.2. Baselines

To our best knowledge, this is the first attempt to predict the continuous-time spreading coverage of social network users. Therefore, we employ the *Autoregressive Integrated Moving Average* (ARIMA) model (Mills, 1991), which is widely used for fitting and forecasting the time series data in the area of statistics and econometrics, as one baseline. This model can first fit to time series data (in our case, a user's spreading sizes of different days in the history), then predict this user's spreading size in the future. Thus, the spreading sizes of first 12 days are used to build the ARIMA model. Then, it predicts the entire 22 days. Note that the optimal ARIMA is always selected based on Akaike information criterion (AIC) and Bayesian information criterion (BIC) for comparison.

In addition to ARIMA, one of the basic information diffusion models – *Independent Cascade* (IC) (Kempe et al., 2003) is used as the second baseline. In the IC model, a user u who mentions a topic in his/her tweets at the current time step t is treated as a new activated user. An activated user has one chance to "activate" each of his/her neighbors (i.e., make them adopt this topic) with a certain probability. If a neighbor v posts the same topic after time step t, it is said that v becomes active at time step $t+1$. If v becomes active at time step $t+1$, u cannot activate v in the subsequent rounds.

In order to apply the IC model to calculate users' spreading sizes, the activation probability for every pair of users needs to be estimated. Specifically, for a user u, we first obtain the spreading size of each of his/her topics during the first 12 days, thus we can get average spreading size over all of his/her topics. Then, the daily average spreading size (DDS) is computed from dividing the average spreading size by 12 days. Finally, $1/DDS$ is taken as the activation probability of u and each of his/her neighbors.

Besides the abovementioned two baselines, we also compare our IDM-CTMP with two recent works from Goyal, Bonchi, and Lakshmanan (2010), and Saito, Kimura, Ohara, and Motoda (2010) (please refer to Section 6 for introduction to these two works). We name these two methods as "Goyal-model" and "Saito-model" respectively.

5.4.3. Prediction

To compare the performance of IDM-CTMP, we choose 10,000 top users computed by IDM-CTMP given the entire 22 days and 10,000 random users. Three well-known metrics for measuring prediction accuracy are utilized in our experiment for evaluation: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and MASE (Mean Absolute Scaled Error).

The average values of three metrics for IDM-CTMP, ARIMA, IC, Goyal-model, and Saito-model are listed in Tables 2 and 3. It can be seen that our proposed method IDM-CTMP performs (1) better than baseline methods ARIMA and IC, because ARIMA fits the overall trend of the time series data and does not consider the underlying network cascading causing the change of the spreading sizes. The basic IC model needs predefined time step, which is set to be 1 day. It might be too large to capture the real-time topic propagation. However, if setting it to be small, it would take long time to run. The parameter estimation of the basic IC model assumes the constant activation probability for all neighbors, which could be another reason of poor performance; (2) better than Goyal-model and Saito-model mainly because it models dynamic probabilities instead of static ones.

In Fig. 7, we plot the ground truth spreading sizes and the predicted spreading sizes of different models for both top 5 users and 5 random users. Note that the plot is mainly for illustrating how prediction results of IDM-CTMP fit the ground truth. In order to make it more readable, we skip the results of Goyal-model and Saito-model. We can observe that even though the predicted results by IDM-CTMP are not exactly as same as the ground truth, most of the predicted curves fit very close to the true curves. In particular, most of the "peaks" and "valleys" can be well captured by our proposed method. However, ARIMA and IC does not perform well, missing many "peaks" and "valleys" and having wrong predictions.

Table 2
The comparison over the 10,000 top users.

Methods	MAE	RMSE	MASE
IDM-CTMP	3.290	4.231	0.714
ARIMA	4.369	5.470	1.294
IC	5.858	7.209	2.355
Goyal-model	4.831	6.112	1.818
Saito-model	4.412	5.861	1.773

Table 3
The comparison over the 10,000 random users.

Methods	MAE	RMSE	MASE
IDM-CTMP	1.686	2.055	0.702
ARIMA	2.026	2.855	0.764
IC	3.928	4.834	2.091
Goyal-model	3.130	4.118	1.987
Saito-model	2.817	4.005	1.629

6. Relate work

A number of recent works have addressed the matter of user influence on social network. Many of them regard user influence as their network metrics. Kwak et al. (2010) found the difference between three influence measures: number of followers, page-rank, and number of retweets. Cha, Haddadi, Benevenuto, and Gummadi (2010) also compared these three measures, and discovered that the number of retweets and the number of mentions are correlated well with each other while the number of friends does not correlated well with the other two measures. Their hypothesis is that the number of followers of user may not be a good influence measure. Weng, Lim, Jiang, and He (2010) regarded the central users of each topic-sensitive subnetwork of the follower-and-followee graph as influential users. Other work such as Ghosh and Lerman (2010), Romero, Galuba, Asur, and Huberman (2010), Agarwal, Liu, Tang, and Yu (2008) and Tang, Sun, Wang, and Yang (2009) mined users influence from their static network properties derived from either their social graphs or activity graphs.

Various dynamic diffusion models have also been proposed to discover the influential users. They are shown to outperform influence models based on static network metrics (Richardson & Domingos, 2002; Ghosh & Lerman, 2010). A lot of work in this direction are devoted to *viral marketing*. Domingos and Richardson (2001) and Richardson and Domingos (2002) were the first to mine customer network values for 'influence maximization' for viral marketing in data mining domain. The proposed approach is a probability optimization method with the hill-climbing heuristics. Kempe et al. (2003) further showed that a natural greedy strategy can achieve 63% of optimal for two fundamental discrete-time propagation models - Independent Cascade Model (IC) and Linear Threshold Model (LT). Many diffusion models assume the influence probabilities on the edges or the probability of acceptance on the nodes are given or randomly simulated. Goyal et al. (2010) proposed to mine these probabilities by analyzing the past behavior of users. Saito, Kimura, Ohara, and Motoda (2010) and Saito et al. (2010) extend IC model and LT model to incorporate asynchronous time delay. Model parameters including activation probabilities and continuous time delay are estimated by Maximum Likelihood. Our proposed diffusion model is different from the above discussed models: (1) We model the dynamic probabilities of edge diffusion and node threshold changing over the time, rather than computing the static probabilities. (2) Our model is a Continuous-Time diffusion model instead of a discrete-time diffusion model. Although Saito et al. also proposed Continuous-Time models, the fundamental diffusion process of their models are following LT and IC models. For example, in asynchronous IC, an active node can only infect one of its neighbors in one iteration, while our proposed models does not assume iterations so that nodes can be activated at any time without resetting the clock in the new iteration. Moreover, the models proposed by Saito et al. supposed only one initial active user and focused on model parameter estimation, not much on prediction. The experiments are evaluated on simulated data from some real network topology. Our proposed model estimates the

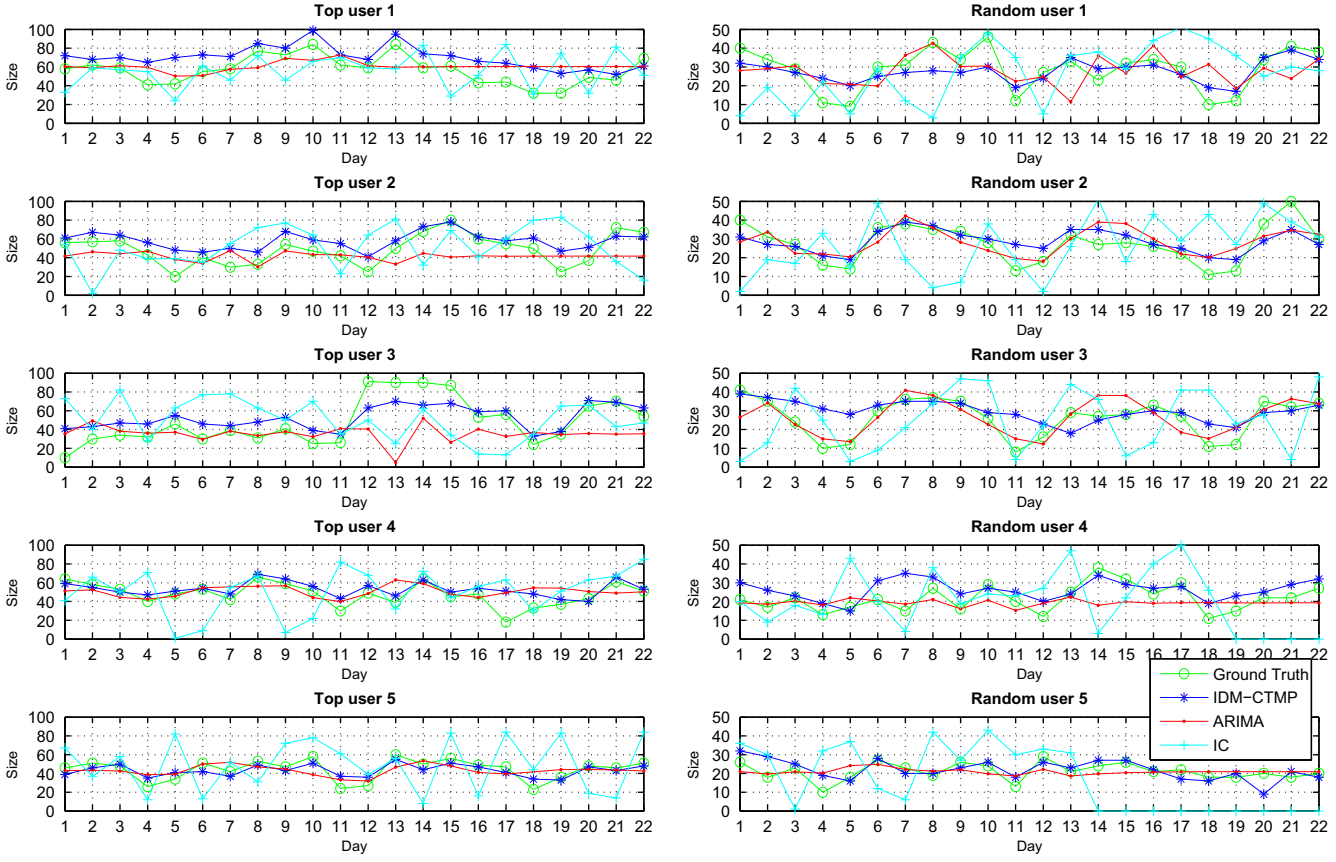


Fig. 7. The comparison between the predicted spreading size of top ranked 5 users (left side) and randomly picked 5 users (right side) by IDM-CTMP and baseline against the ground truth.

model parameters from the real large-scale social network data, allows many initial active users asynchronously or simultaneously to influence other users, and predicts the real diffusion sizes in the future.

In addition, most of influence models are basically descriptive models instead of predictive models. Bakshy, Hofman, Mason, and Watts (2011) studied the diffusion tree of URLs on twitter, and train a regression tree model on a set of user network attributes, user past influence, and URL content to predict users' future influence. Our work is quite different from the work of Bakshy et al. in the following aspects: (1) They predict users average spreading size in the next month based on the data from the previous month. However, the dynamic nature of word-of-mouth marketing determines that the influence coverage vary over the time. Thus our work aims at predicting the spreading size of each individual user within a specific given date, so we can answer "what is the spreading size of user A within 2 hours, 1 day, or 1 month, etc.". (2) Their work is built on top of a regression model, while our work proposes a real-time stochastic model. The input and output from these two models are quite different. (3) Besides URLs diffusion, we also study the diffusion of Hashtags on twitter, which usually have longer lifetime.

Continuous-Time Markov Process (CTMP) has been used in web-page or document browsing. Huang et al. (2004) adopted it to model the web user visiting patterns. Liu et al. (2008) also utilized Continuous-Time Markov Process to model user web browsing patterns for ranking web pages. Song, Chi, Hino, and Tseng (2007) employed CTMP to mine document and movie browsing patterns for recommendation. To the best of our knowledge, our work is the first to construct influence diffusion model based on CTMP for spreading coverage prediction and user influence on social networks. We are also the first to introduce three intuitive criteria for users to compare and choose different influence models.

7. Conclusion

In this paper, we propose IDM-CTMP, an information diffusion model based on Continuous-Time Markov Process. IDM-CTMP is able to predict the influence dynamics of social network users, i.e., it can predict the spreading coverage of a user within a given period of time. We also define two other dynamic influence metrics, and empirically compare different influence metrics on three dimensions of influence: (i). *Monomorphism vs. Polymorphism*; (ii). *High Latency vs. Low Latency*; and (iii). *Information Inventor vs. Information Spreader*. Our experiment results show that the IDM-CTMP metric favors the users with high inventing ability, high spreading ability, and low topic adoption latency. In addition, IDM-CTMP achieves very promising performance as its predicted spreading size demonstrated can fit closely to the ground truth.

Acknowledgment

The work is partially supported by the US National Science Foundation under Grants DBI-0850203, CNS-1126619, and IIS-1213026, the U.S. Department of Homeland Security under Grant Award Number 2010-ST-062000039, Army Research Office under Grant No. W911NF-1010366 and W911NF-12-1-0431, and the Jiangsu 973 Scientific Project (BK2011023), and the National Natural Science Foundation of China (61272419 and 61300053).

References

Adamic, L. A., & Adar, E. (2005). How to search a social network. *Social Networks, 27*, 2005.
Agarwal, N., Liu, H., Tang, L., & Yu, P. (2008). Identifying the influential bloggers in a community. In *WSDM* pp. 207–218.
Anderson, W., & James, W. (1991). *Continuous-time Markov chains: An applications-oriented approach* (Vol. 7). New York: Springer-Verlag.
Bakshy, E., Hofman, J., Mason, W., & Watts, D. (2011). Everyone's an influencer: Quantifying influence on twitter. In *WSDM* pp. 65–74.
Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *AAAI (ICWSM)*.
Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending content from information streams. In *CHI* pp. 1185–1194.
Davitz, J., Yu, J., Basu, S., Gutelius, D., & Harris, A. (2007). iLink: Search and routing in social networks. In *SIGKDD*.
Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *SIGKDD* (pp. 57–66). ACM.
Dynkin, E. (1965). Markov processes.
Feller, W. (2008). *An introduction to probability theory and its applications* (Vol. 2). John Wiley & Sons.
Gardiner, C. (1985). *Handbook of stochastic methods*. Berlin: Springer.
Ghosh, R., & Lerman, K. (2010). Predicting influential users in online social networks. In *CoRR, abs/1005.4882*.
Golbeck, J., & Hendler, J. (2006). Inferring binary trust relationships in web-based social networks. *ACM TOIT, 6*, 497–529.
Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *WSDM* pp. 241–250.
Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *WWW* (pp. 491–501). ACM.
Huang, Q., Yang, Q., Huang, J., & Ng, M. (2004). Mining of web-page visiting patterns with continuous-time markov models. *PAKDD*, 549–558.
Karlin, S., & Taylor, H. (1975). In *A first course in stochastic processes* (pp. 474–502). New York: Academic Press.
Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *SIGKDD* pp. 137–146.
Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *WWW* pp. 591–600.
Li, J., Peng, W., Li, T., & Sun, T. (2013). Social network user influence dynamics prediction. In *Web technologies and applications* (pp. 310–322). Springer.
Liu, Y., Gao, B., Liu, T., Zhang, Y., Ma, Z., He, S., & Li, H. (2008). Browserank: Letting web users vote for page importance. In *SIGIR* pp. 451–458.
Mills, T. (1991). *Time series techniques for economists*. Cambridge University Press.
Richardson, M., & Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *SIGKDD* pp. 61–70.
Rogers, E. M. (2003). *Diffusion of innovations* (Vol. 27). Free Press.
Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2010). Influence and passivity in social media. In *CoRR, abs/1008.1253*.
Saito, K., Kimura, M., Ohara, K., & Motoda, H. (2010). Efficient estimation of cumulative influence for multiple activation information diffusion model with continuous time delay. In *PRICAI* (pp. 244–255). Springer-Verlag.
Saito, K., Kimura, M., Ohara, K., & Motoda, H. (2010). Generative models of information diffusion with asynchronous timedelay. *JMLR – proceedings track, 13*, 193–208.
Song, X., Tseng, B. L., Lin, C. -Y., & Sun, M. -T. (2006). Personalized recommendation driven by information flow. In *SIGIR* pp. 509–516.
Song, X., Chi, Y., Hino, K., & Tseng, B. (2007). Information flow modeling based on diffusion rate for prediction and ranking. In *WWW* pp. 191–200.
Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. In *SIGKDD* pp. 807–816.
Weng, J., Lim, E., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *WSDM* pp. 261–270.

Info Vis

Article

Information Visualization
2014, Vol. 13(2) 134–158

DOI: 10.1177/1473871612468877
ivi.sagepub.com
SAGE

Characterizing the intelligence analysis process through a longitudinal field study: Implications for visual analytics

Youn-ah Kang and John Stasko

Abstract

While intelligence analysis has been a primary target domain for visual analytics system development, relatively little user and task analysis has been conducted within this area. Our research community's understanding of the work processes and practices of intelligence analysts is not deep enough to adequately address their needs. Without a better understanding of the analysts and their problems, we cannot build visual analytics systems that integrate well with their work processes and truly provide benefit to them. In order to close this knowledge gap, we conducted a longitudinal, observational field study of intelligence analysts in training within the intelligence program at Mercyhurst College. We observed three teams of analysts, each working on an intelligence problem for a 10-week period. Based on the findings of the study, we describe and characterize processes and methods of intelligence analysis that we observed, make clarifications regarding the processes and practices, and suggest design implications for visual analytics systems for intelligence analysis.

Keywords

Intelligence analysis, visual analytics, qualitative user study

Introduction

Visual analytics applies to many domains and problem areas, but one area of particular study since the beginning of the field has been intelligence analysis. Intelligence analysis is a cognitively demanding process, one that seems ideal for the application of visual analytics tools. Accordingly, a growing number of systems have been built for it.[1–4]

Research in human–computer interaction also teaches us to deeply analyze and understand end users and their problems in order to design appropriate computational solutions. We question whether visual analytics systems, including some of our own, have been based on a deep enough understanding of the discipline. Relatively few studies of intelligence analysts, their tasks, and their work processes exist. Notable exceptions[5–8] provide initial insights into the field, but we have frequently interacted with analysts who feel that their practices are misunderstood and that visual analytics systems often fail to address their most important problems.

To address these concerns and to learn more about the analysis process, we conducted a longitudinal, observational field study of intelligence analysis on real-world problems. Unfortunately, getting access to working professional analysts is challenging. Even if they are available, it is difficult or impossible to study them for an extended period of time while they work on real tasks without having some type of special

School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

Corresponding author:
Youn-ah Kang, Google Inc, 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA.
Email: kang.younah@gmail.com

access that simply was not available to us. As an alternative, we studied analysts in training who are soon to become working professionals. More specifically, we studied groups of students from the Department of Intelligence Studies at Mercyhurst College as they conducted a long-term intelligence project.

We were given deep access to the students, the materials they examined, the tools they used, and their final intelligence products. We interviewed the teams multiple times and observed their group meetings. Additionally, we interviewed their instructor to learn his impressions of the process. Our goal was simply to better understand what these young analysts do, the challenges they face, and how we might be able to help them. Thus, the contributions of our research include a characterization of the processes and methods of intelligence analysis that we observed, clarification and reflection of several beliefs about intelligence analysis processes and practices, and resultant design implications for visual analytics systems for intelligence analysis.

Munzner[9] has argued of the importance and the need for more domain characterization research like this. She notes that such research is both difficult and time-consuming to do properly, but the visualization community could benefit greatly from it.

This article is an extended version of the conference paper.[10] We have added an expanded discussion of the tools and methods used by the analysts with a specific focus on their use of wikis. We discovered that wikis were used pervasively by the analysts for a variety of benefits. We explain how the analysts used wikis and the specific characteristics of wikis that assisted the analysis.

Background

One of the most widely used models in the visual analytics community is Pirolli and Card's[7] sensemaking model for intelligence analysis. While the model broadly characterizes processes used in the analysis activities and has guided the design of visual analytics tools, the model does not provide rich details of how intelligence analysts work in the real world. More empirical and descriptive explanations of the intelligence analysis process are required to provide appropriate visual analytics system solutions.

Several studies have captured and characterized the work practices and analytical processes of individual or collaborative analysis through a qualitative approach. Chin et al.[5] conducted an observational case study with professional intelligence analysts in which participants worked on real-world scenarios. The researchers revealed various characteristics of the analytical processes of intelligence analysts. Gotz et al.[11] also recognized the lack of studies examining analyst behavior and conducted a user study to explore the ways in which analysts gather and process information. Another study by Robinson[8] examined how analysts synthesize visual analytics results by studying domain experts conducting a simulated synthesis task using analytical artifacts. Based on the analysis of video coding results, he identified several characteristics in the process of collaborative synthesis. While these studies did not evaluate specific visual analytics tools or features per se, they provide valuable implications to inform design directions for future support tools.

Relatively few studies examine the analytical culture in general. These include a number of books[6,12–14] published from the intelligence analysis domain. These books provide insights into the complex analytical process as seen by those who practice it as well as an understanding of some critical aspects of the analysis.

Krizan,[12] in *Intelligence Essentials for Everyone*, provides a slightly revised version of the traditional intelligence cycle,[15] which contains several component functions including intelligence needs, collection activities, processing of collected information, and analysis and production. Quoting Dearth,[16] she states, "These labels, and the illustration below, should not be interpreted to mean that intelligence is a uni-dimensional and unidirectional process. In fact, 'the process is multidimensional, multi-directional, and—most importantly—interactive and iterative.'"

Clark[13] also describes the current intelligence process using his target-centric approach to intelligence analysis. Examining how intelligence should be done, he advocates an inclusive approach that includes all the stakeholders or individuals affected by the intelligence produced. In this approach, he argues that "the goal is to construct a shared picture of the target, from which all participants can extract the elements they need to do their jobs and to which all can contribute from their resources or knowledge." Compared to other models, this approach implies more interactivity throughout the analysis cycle.

Johnston,[6] an anthropologist, conducted an ethnographic study of the Central Intelligence Agency (CIA) for a year and identified variables that affect intelligence analysis and requirements for techniques and procedures to reduce analytical error. While he made useful recommendations to improve analytical performance, his approach was primarily intended to understand organizational culture and describe current community practices, rather than identifying leverage points for designing support systems.

While many researchers have suggested new intelligence models and tried to emphasize nonlinearity, existing intelligence process models still do not clearly and accurately describe what people do in intelligence analysis and fail to capture the nuance of the process

as practiced. In our study, we aim at an in-depth exploration of the intelligence analysis process, emphasizing the disconnect between theory and real-world practice. We also seek to deeply understand the analysis process with an eye toward designers, so that we can provide meaningful implications for developing technological support for analysts.

Methods

In order to investigate the intelligence analysis process in-depth, we conducted an observational study of teams of analysts conducting an in-class intelligence project. In the long-term (10-week) project, each team addressed a real intelligence problem proposed by a client. We observed three teams, monitoring their status and process throughout the project. At the end of the project, each team had to produce final deliverables and present their findings and analysis to decision makers.

Participants

We recruited three groups of students, one team of four undergraduate students (Team A) and two teams of five graduate students (Teams B and C), from the Department of Intelligence Studies at Mercyhurst College.[17] Mercyhurst's Intelligence Program, started in 1992, provides education for students who want to pursue a career as an intelligence analyst. It is recognized as one of the top programs for intelligence studies in the United States, offering a broad range of classes and degrees for students seeking a career as an analyst in national security, law enforcement, or the private sector.

We recruited students who were taking the courses named "Strategic Intelligence" (undergraduate) and "Managing Strategic Intelligence" (graduate), in which teams are required to conduct an analysis project over a 10-week term. The two courses are very similar with respect to the projects. The students all were close to graduation, with past internship experience, and most of them had already received job offers.

While these student teams clearly are not practicing professional analysts, there was not a significant difference between the way the students worked and the way real analysts work, according to the instructor. The analysis process used in the class was modeled directly after the process employed by the US National Intelligence Council to produce its strategic reports, the National Intelligence Estimates.[18] The instructor also intentionally stayed relatively detached from the students, acting as a mentor and limiting his supervision, so that the teams could autonomously work on the project. The teams were diverse in expertise on the subject matter, which is common for teams in the intelligence community (IC). One key difference from real-world practice was the relative absence of administrative and bureaucratic overhead affecting the student teams, as well as issues relating to security clearances. They operated in a much more "sanitary" environment than the real world.

Task

Different types of intelligence questions exist—we focused on one of the most common types, strategic intelligence. Strategic intelligence is "intelligence that is required for the formulation of strategy, policy, and military plans and operations at national and theater levels."[19] In other words, strategic intelligence is the intelligence necessary to create and implement a strategy, typically a grand strategy. It aims to provide ways to accommodate and/or coordinate a variety of variables. Strategic intelligence is exploratory and long term in nature.

The requirement for tasks within the class was that "the questions should be relevant and relatively important to the client's success or failure but outside their control." We served as a client/decision maker for Team A in order to observe the process even closer, whereas Teams B and C worked with external organizations. The specific issues each team addressed were the following.

Team A. The strategic assessment of potentially influential factors to the evolution of computer-mediated undergraduate and graduate distance education: What aspects of computer-mediated distance education will likely influence R1 institutions during the next 5, 10, and 20 years with specific, but not exclusive, emphasis on undergraduate education and computer science? As part of this overreaching question, this study will seek to address key components including:

- Enrollment figures
- Value of education
- Cost of education

Team B. Who are the key people, technologies, and organizations that likely currently have or will develop the potential to disrupt or replace traditional US national security IC analytical workflows and products with commercially available products available over the next 24 months? Criteria that will be used to identify these key players are as follows:

Those who are not beholden to the IC or US Government as primary sources of funding.
Those who have the potential to solve IC-like analytical problems.

Those who are looking at future-based events or actions that are outside the control of the forecaster/predictor.
Those who are focused on forecasting or predicting future courses of action.

Team C. What are the most consistent and identifiable characteristics displayed by potential insider threats to (a defense department)?

- An insider threat will be defined as an individual or collection of individuals employed directly or indirectly by the department who violate security or access control policies with the intent of causing significant damage to the department's personnel, operations, or information.
- Within the broad range of insider threats, special priority will be given to violent threats and improper diversion of information or physical assets.
    - Violent threats include actions that endanger department personnel, physical assets, and the department's protective service mission.
    - Improper diversion includes the sale, surrender, and/or sabotage of information or physical assets.

Throughout the study, we tried to minimize our intervention, and, furthermore, every decision on the intelligence process (e.g. what tools to use) was made by the study participants. The teams updated the status and the process of the project on a wiki site. At the end of the semester, they needed to produce a final report that synthesized analytical results and strategies of the entire analysis process.

Study protocol and procedures

The analyst teams conducted the project for 10 weeks—from the week of 1 September through the week of 10 November 2010 for Team A and from the week of 1 December to the week of 14 February 2011 for Teams B and C. Normally, strategic intelligence projects range from a couple of months to years; 10 weeks is short but within normal limits for strategic intelligence.

Before the project began, the external clients formulated a draft of their initial intelligence problem. In the first week of the project, the clients conducted a conference call with the analyst team to discuss the scope and requirements of the problem. During the next 2 weeks, the analysts refined the problem and wrote a formal statement of the intelligence question, which they call "Terms of Reference (TOR)." Upon approval from the decision makers, the teams began working on the problem, which took another 7 weeks.

The wiki platform was used as a workspace for analysts to document their process and findings, and we were able to monitor the wiki's status throughout the project period. The final reports of the projects were also documented on the wikis.

During the project period, we conducted two face-to-face meetings with each team—one in week 7 and the other in week 10. In the meetings, we interviewed each team as a group and the class instructor in order to learn more details about the project's status, process, difficulties, and future steps. Each interview took approximately an hour. While the interview was semi-structured, we followed an interview guide containing several key topics,[20,21] including the following:

- How do the analysts perceive their analysis process?
- What barriers and difficulties do they encounter?
- Tools and aids being used—where and why?
- Collaborative aspects in the analysis process.
- Where in the process can technology help?

We also observed two team meetings firsthand, which took about 3 h in total.

Data collection and analysis

Most of the process descriptions and produced artifacts were stored digitally. The teams reported methodologies, tools used, sources, as well as the findings on their own website (wiki). To further understand the process, we analyzed interview notes and audio recordings from the focus group interviews. We used the artifacts produced by the analysts, such as drawings, wiki pages, tables, and slides, as further data. Additionally, we had access to history logs of wiki page changes.

We transcribed each interview's audio recording and then coded the transcripts based on the grounded theory approach.[22] We began by identifying major themes and categories from the text. One emergent theme focused on the analysis process, including methodology and challenges encountered. Another theme was collaboration, focusing on how and to what degree the analysts collaborated and what types of collaboration existed. Throughout the coding process, we iteratively refined the categories. We then elaborated on supporting evidence from the data for each category through a deductive approach.

Results and findings

Overall analysis process

Through the project, we found that four component processes were essential to the overall analysis:

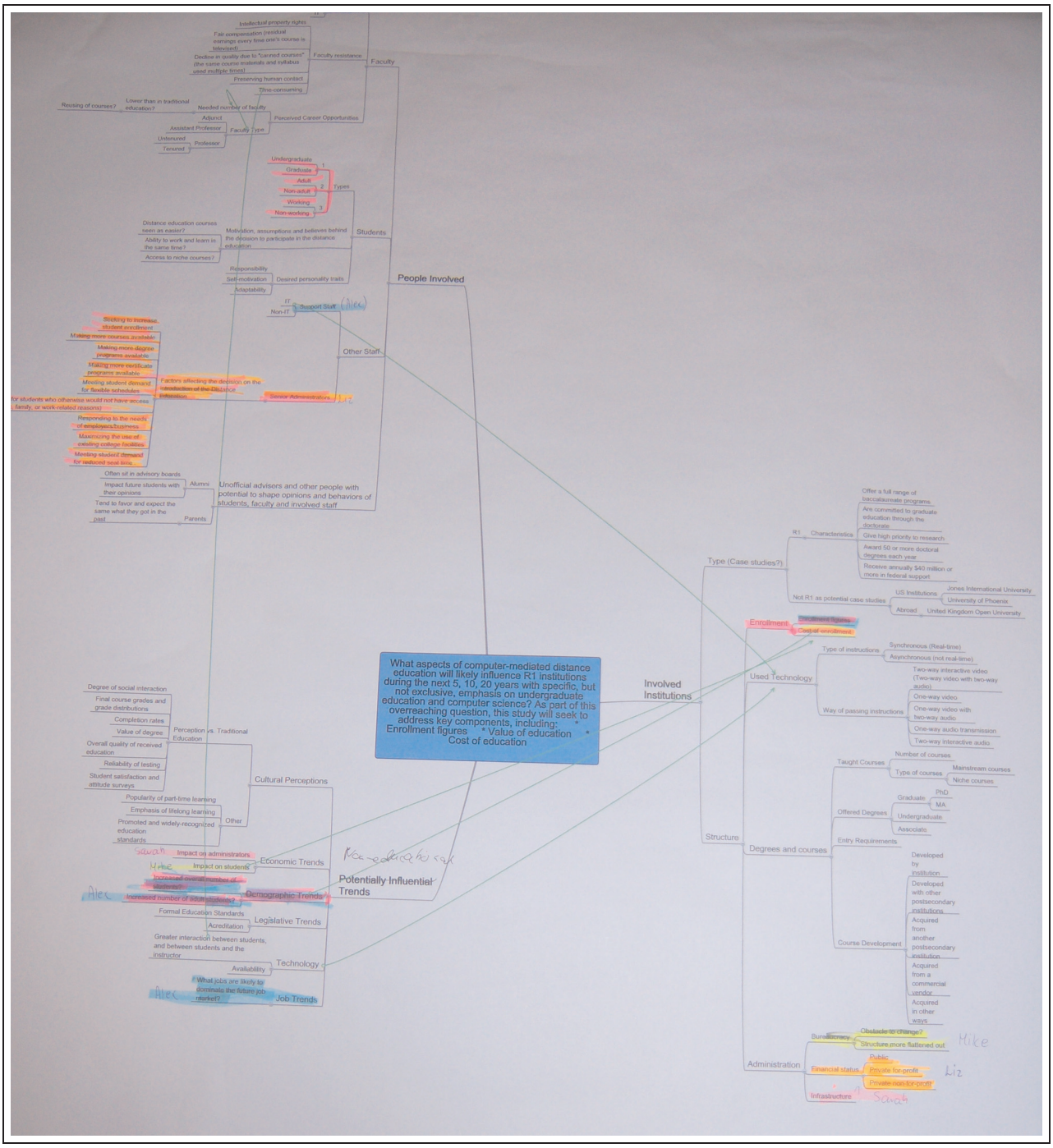


Figure 1. A conceptual model.
Source: Printed from Mindmeister.[23]

constructing a conceptual model, collection, analysis, and production. While the four processes did not occur linearly, this section describes the importance of each and how the analyst teams worked on each.

Phase 1: constructing a conceptual model. Once the teams and clients/decision makers finalized the requirements of the intelligence question, the teams started to build a conceptual model, which is a map of issues and concepts that the team will be investigating to address the problem. The conceptual model illustrates the areas the analysts need to research by helping them to visualize the question at hand. The question is placed in the center and then several high-level components of the question surround the question (Figure 1). Each component branches out and

creates a bigger map, from which the team gains an idea of the areas with less/more information that they need to research. This allows the team to focus on collecting a set of data with an appropriate scope.

While the significance of the conceptual model differs depending on the question and the team, it plays a key role for the team to understand the domain area and determine the direction of research. We were told that analysts often construct this conceptual model implicitly, rather than externalizing it, which we found quite interesting. The instructor commented:

> In most cases, it's implicit. People don't write it down. But it's the way they are actually doing it. There's a model in people's heads, and that's far more important than the data. There's research that says analysts' judgments are far more driven by the way they think about the problem than the data itself. So making the way you think about the problem explicit would allow analysts to identify whether they disagree about how to think about this model, and to merge their best thinking about this model. So the process happens, it's just the degree of to which it's made explicit, that is unusual.

Phase 2: collection. While working on the conceptual model, the teams also assigned areas/concepts to each member. Next, they collected information from various sources including online and offline sources (e.g. interviews with experts), which they call "all-source intelligence." While each analyst was responsible for collecting data about their assigned topic(s), the team shared their sources using Zotero,[24] a web browser plug-in for gathering and organizing source material. This allowed teammates to view the data like a common library—other team members might already have found information that they need. More specifically, the data collection process typically involved these steps:

- Once an analyst identifies needed information, they search the Internet using search engines and various keywords.
- The analyst also sets up Rich Site Summary (RSS) news feeds on websites of interest using Google Reader.
- Whenever they find useful sources, whether for their own topic or someone else's, the analysts place the link into a Zotero group library.

The collection process occurred from the very beginning to the final stages—the ninth week of the project.

Phase 3: analysis. The analysis phase exhibited various characteristics depending on the requirements and analytical methods used. In this phase, analysts processed data that they collected from many different sources in order to convert "information to knowledge." While Team A directly began writing short format analytical reports on each topic, Teams B and C used a more structured format (e.g. spreadsheets) to quantify information and rank the significance of each topic or entity. No matter which method they used, the initial analysis of each topic/entity was undertaken and written by one person in accordance with the assigned topic. However, everyone on the team could review and comment on others' work via the wiki pages. In all cases, the analysis phase was incorporated with the collection and the production phases.

Phase 4: production. Once individual collection and analysis were almost finished, the teams met and tried to synthesize findings from each part, which led to the "key findings"—the major product of the analysis. Production was an intensive reading/writing process in which the team collaborated tightly with each other. This stage was more to prepare a presentation for the decision makers. Team members repeatedly checked their sources and findings to make sure that they were consistent and logical.

Reiterating, while we separated these four components for the sake of clarification, the process was not simple, and it was not clear which phase the team was in throughout the project. Instead, the characteristics of the question and the analytical method chosen most influenced the process. In our study, we observed two different styles of intelligence analysis process. The difference in approaches resulted from the type of the intelligence question.

Intuitive analysis—Team A. Team A addressed potentially influential factors to online distance education in the near future. Because the requirements were rather broad and intuitive, the team decided to take a top-down approach, investigating meta-information sources such as research that forecasts future education trends.

Instead of using a specific analytical method, this team depended considerably on the conceptual model and used it as a guide throughout the entire project. They put significant time and effort into constructing it, revising the conceptual model until the seventh week of the project. Because of time constraints, they were not able to cover all the topics in the model. Through discussions, they chose a number of concepts they felt most worth exploring and divided up the concepts for each member.

After collecting and reading information for their designated topic, each analyst wrote a short format analytical report that synthesized the information.

MCIM : Entities

Entity Name	Unstructured Data	Unreliable Data	Incomplete Data	Detect Deception in Data	Data Extracted	Application Compatibility	Impact Score* (see Impact Matrix)	Total Score
Text Analytics	3	0	3	1	3	3	3	16
Dependency or Association Analysis	0	2	2	2	0	3	3	12
Network Analysis	0	2	2	2	0	3	3	12
Data profiling and transformations	2	2	1	2	0	3	1	11
Prescriptive Analytics	0	2	1	2	0	3	3	11
Sequential Pattern Analysis	2	0	2	0	2	3	2	11
Bayesian Analytics	0	2	0	2	0	3	3	10
Clustering or Segmentation	0	2	0	2	0	3	2	9
Optimization Models	0	1	1	1	0	3	3	9
Game theory	0	1	1	0	0	3	3	8
Regression Analysis	0	0	0	2	0	3	3	8
Basket Analysis	0	0	2	0	0	3	2	7
Classification	0	0	0	1	0	3	3	7
Simulations	0	0	1	0	0	2	3	6
Time series forecasting	0	0	0	0	0	3	3	6
Time series tracking	0	0	0	0	0	2	2	4
Behavior Learning Software-Machine Learning	3	2	3	2	3	3	3	19
Neural Networking	3	2	3	3	3	2	2	18
Complex Event Processing	3	2	2	3	2	2	2	16
Statistical Package for the Social Sciences (SPSS)	3	1	2	2	2	3	3	16
Temporal Analytics	3	1	3	1	3	3	2	16
Reality Mining	3	1	3	1	1	3	3	15

Entities

Figure 2. MCIM of people, technologies, and companies.
MCIM: multicriteria intelligence matrix.

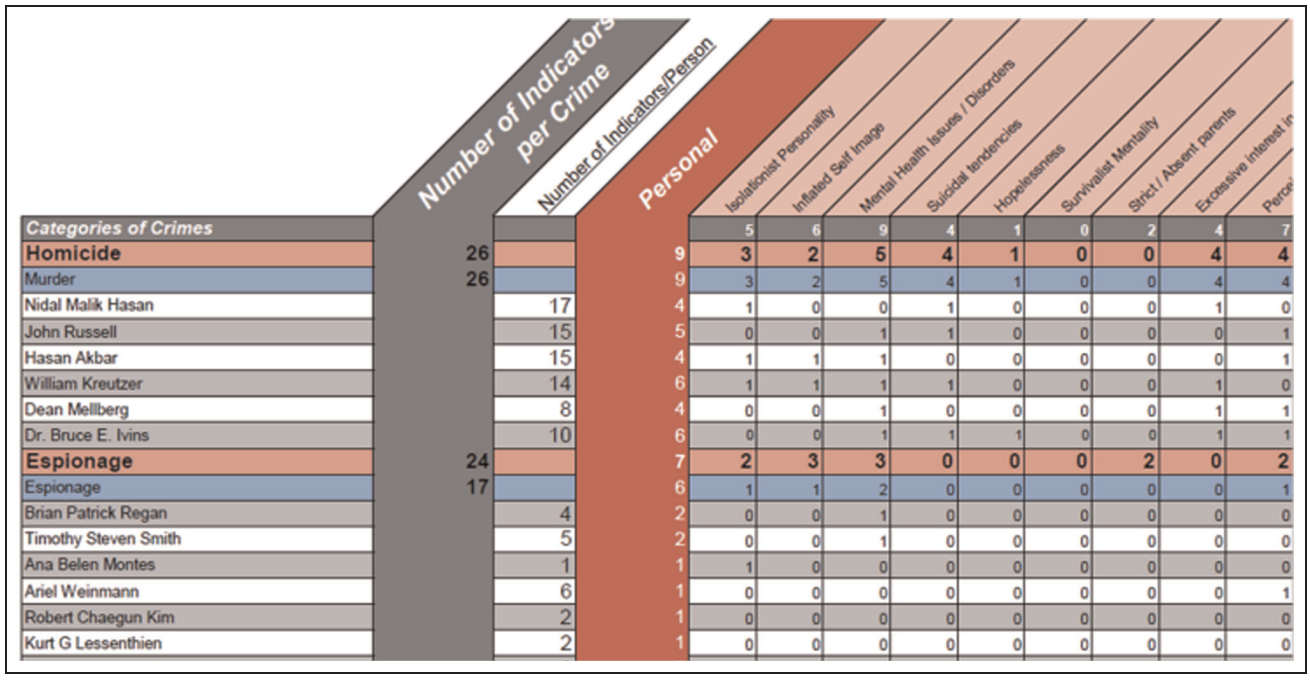
Most of the analysis simply involved reading. For a few topics that required careful weighing of alternative explanations, the team employed analysis of competing hypothesis (ACH).[14] While documenting results, everyone was able to review and edit others' drafts on the wiki page, and team members frequently discussed others' analysis (short write-ups) both online and face-to-face. Therefore, everyone was responsible for the reporting of each topic.

After working on the individual topics, the team met to write key findings together. This team invested considerable efforts in synthesizing their findings because their narrative was extremely important for their intuitive type of analysis.

Structured analysis—Teams B and C. Teams B and C used structured analysis with quantified information because their research questions tended to be more specific and required rank ordering of entities (e.g. top x indicators, key people/companies). Both teams built their conceptual model in the beginning as a base model. For these teams, however, the model was more of a collection plan rather than an actual conceptual model. Although they used the model to collect information and divide up the work, they did not refer to it for the remainder of the project. Instead, they started building a matrix in a spreadsheet to collect and analyze data from diverse sources. The matrix was rather a reinterpretation of the conceptual model, and each cell in the matrix indicated a collection requirement.

The purpose of the matrix was to evaluate each entity based on the criteria chosen and to identify the most influential ones, those of most interest to the decision maker. Team B, which was asked to identify key people, technologies, and companies that might affect IC products, created a matrix and chose criteria while collecting information (Figure 2). They identified 180 entities and graded each based on the criteria, noting the ones with highest scores. Team C, which was asked to identify indicators displayed by potential insider threats to a defense department, analyzed data from the 117 case studies about crimes using a matrix (Figure 3). They used it to compare the relationship between crimes and motivations, as well as crimes and indicators.

In both the teams, the matrix captured the conceptual model and how each team was thinking about the question. Filling in the cells was a time-consuming part as analysts needed to read and analyze each



Categories of Crimes	Number of Indicators per Crime	Number of Indicators/Person	Personal	Isolationist Personality	Inflated Self Image	Mental Health Issues / Disorders	Suicidal tendencies	Hopelessness	Survivalist Mentality	Strict / Absent parents	Excessive interest in	Perce
				5	6	9	4	1	0	2	4	7
Homicide	26		9	3	2	5	4	1	0	0	4	4
Murder	26		9	3	2	5	4	1	0	0	4	4
Nidal Malik Hasan		17	4	1	0	0	1	0	0	0	1	0
John Russell		15	5	0	0	1	1	0	0	0	0	1
Hasan Akbar		15	4	1	1	1	0	0	0	0	0	1
William Kreutzer		14	6	1	1	1	1	0	0	0	1	0
Dean Mellberg		8	4	0	0	1	0	0	0	0	1	1
Dr. Bruce E. Ivins		10	6	0	0	1	1	1	0	0	1	1
Espionage	24		7	2	3	3	0	0	0	2	0	2
Espionage	17		6	1	1	2	0	0	0	0	0	1
Brian Patrick Regan		4	2	0	0	1	0	0	0	0	0	0
Timothy Steven Smith		5	2	0	0	1	0	0	0	0	0	0
Ana Belen Montes		1	1	1	0	0	0	0	0	0	0	0
Ariel Weinmann		6	1	0	0	0	0	0	0	0	0	1
Robert Chaegun Kim		2	1	0	0	0	0	0	0	0	0	0
Kurt G Lessenthien		2	1	0	0	0	0	0	0	0	0	0

Figure 3. Case study matrix of crimes.

case/source to fill in one cell, addressing "the devil in the details."

However, this type of analysis required additional efforts in the production phase. Initially, the teams converted qualitative information from sources into quantitative information for rank ordering. Once they had completed the matrix, the teams needed to transform its data into a story, so that it could be made useful to decision makers.

Upon the completion of the projects, the instructor evaluated the teams' performances as being in the "top 10% of the projects over 8 years." He commented that all three teams performed the analysis well, and in one case, the decision maker briefed the head of his organization with the team's results.

Tools and methods used

The teams used various software tools and analytical methods to develop hypotheses, arrive at analytical estimates, and create written reports and multimedia products.

Wikispaces/Google Sites. The teams used a wiki platform (Teams A and B—Wikispaces, Team C—Google Sites) to exchange gathered information, aid administration, and share organizational details. The wiki sites became part of the final product, displaying the key findings, TOR, and all analytical reports.

Mindmeister (conceptual model). Mindmeister[23] is an online mind-mapping tool the teams used to build a conceptual model. A conceptual model provides a revisable platform to view the requirements and their components. As research and facts begin to support or refute initial ideas, main ideas become more solidified and focused.

Zotero. The teams used Zotero[24] as a source collection database. Downloaded as an Add-on to Mozilla Firefox, Zotero allows the analyst to search websites and save the sites in a database that is accessible through the Zotero website. The teams used the group library feature to place their sources in a single database.

Website evaluation worksheet. To evaluate the credibility of the online sources, all the teams used the Dax Norman Trust Scale.[25] This matrix allows scores to be applied based on the criteria such as clear bias, corroboration of information, and the analyst's overall perception of the source. Based on the sum of scores, the source can score a high, moderate, low, or not credible rating.

Decision matrix. A decision matrix is a decision-support tool allowing decision makers to address a problem by evaluating, rating, and comparing

different alternatives on multiple criteria. Teams B and C employed a modified version of a decision matrix appropriate to address their problems.

Analytical confidence. Each report includes an analytical confidence section that conveys to the decision maker the overall doubt connected with the estimative statement(s). While assessing the level of analytical confidence, the teams used Peterson's[26] method. Peterson identified seven factors that influence analytical confidence: the use of structured analytical methods, overall source reliability, source corroboration, level of expertise on subject, amount of collaboration, task complexity, and time pressure. In the analytical confidence section, the teams addressed these six factors as applicable to the particular estimate.

Social network analysis. Team C employed social network analysis using i2's Analyst's Notebook[2] to see relationships within the industry. The team analyzed the social network analysis based on betweenness and eigenvector scores.

ACHs. Team A used ACH for some problems. ACH is a simple model for assessing alternatives to a complex problem. It takes analysts through a process for making a well-reasoned, analytical judgment. ACH is particularly useful for issues that require careful weighing of alternative explanations of what has happened, is happening, or is likely to happen. It also helps analysts minimize some of the cognitive limitations.

Wiki usage in intelligence analysis

Since wikis were extensively used as a workspace for analysts in our study, we further examined their wiki usage throughout the process. In addition to the three wikis we observed in the study, we also examined five other wikis as a reference. This section details how and to what extent wikis were used in their intelligence process.

Wiki statistics. We looked at the usage statistics from Wikispaces and Google Sites to better understand the context of the work. Because Google Sites does not support the number of page views, only statistics from the two teams are considered for "the number of views."

Number of pages of analysis. For the teams, 37, 111, and 48 wiki pages were created, respectively. Each wiki page is counted as only one page no matter how long it is. When printed, however, each wiki page can be several pages in length. This number represents the number of pages in the finished projects and not the total number of pages created. In our study, some of the pages—approximately 37%—were created to help write the projects and then deleted in the final cleanup before presentation to a decision maker.

Number of files uploaded. A total of 81, 202, and 152 files were uploaded and used for the analysis. The files include images, Excel spreadsheets, charts, and Word documents.

Number of edits. While an "edit" can be fixing a spelling error or rewriting an entire wiki page, the number of edits can be a useful measure of how much contribution the analysts have made. In our study, the teams generated 1638, 2650, and 1655 edits or an average of 425 edits per analyst. If we look at the number by page, 30 edits were made for each page on average.

Number of views. Whenever an analyst clicks on a page, whether it is read or not, it counts as a view. Analysts, as a normal part of the analytical process, often examined and commented upon the work of others in their group. The analysts generated 14,665 and 21,524 page views or approximately 4021 page views per analyst. Each page on average was viewed 228 times by the analysts.

How analysts used wikis. As shown in the statistics, the analysts heavily used a wiki for the project. The teams' purpose for using the wikis was interesting as well.

In the projects we observed, the use of wiki sites started from the very beginning of the project. Once the teams met with clients and clarified the questions and requirements, they set up their own wiki site and uploaded TOR to a page. In this stage, the teams also invited clients to the wiki, so that the teams could communicate with their clients more effectively. Throughout the process, the teams used the discussion feature available in the wikis, enabling communication between analysts about specific tasks and the overall project.

In the early stages of a project, a wiki was primarily a repository for information. When the teams were actually working on a conceptual model, they used another tool (i.e. Mindmeister) and had more face-to-face meetings for discussion. Once they created an initial conceptual model, they put the model into the wiki (Figure 4). As the project evolved, the model helped analysts identify intelligence gaps, discover new information, and update current information, thus aiding in the evolution of the project. Wikis, therefore, helped

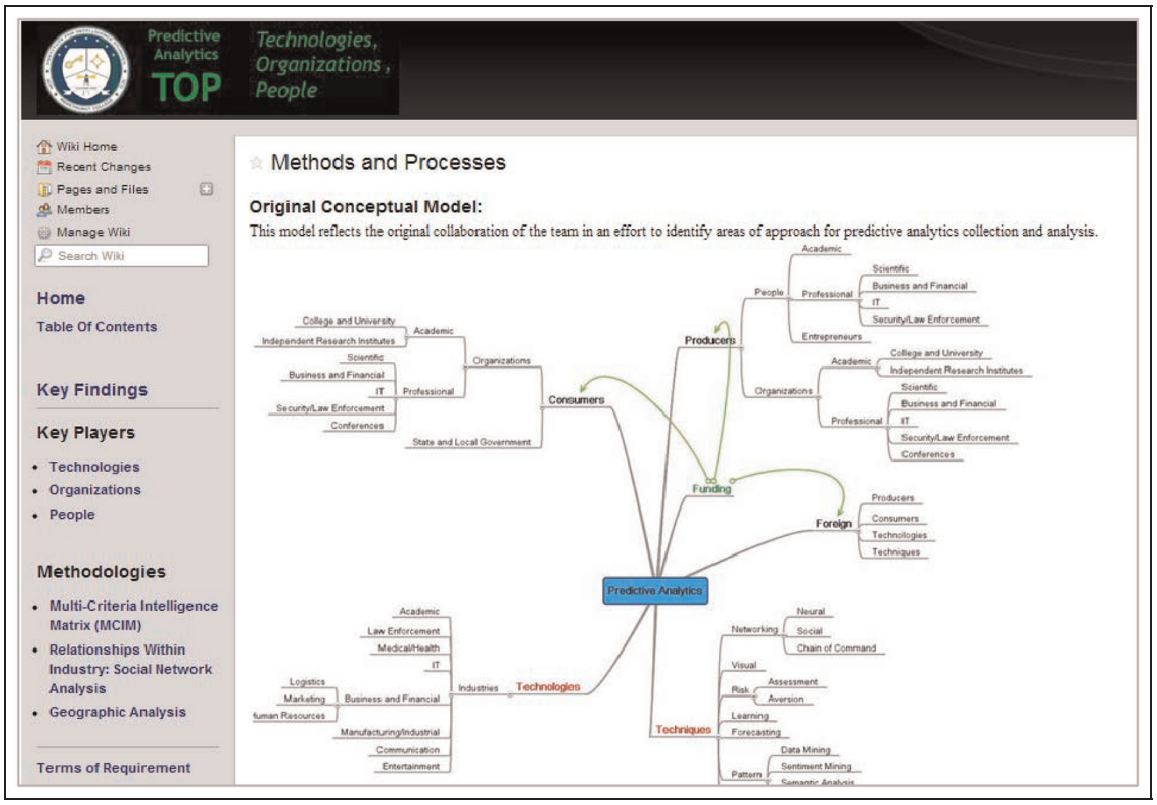


Figure 4. A conceptual model added to a wiki.

the team members stay familiar with the conceptual model throughout the project period.

Once the teams decided what to collect based on the conceptual model and started collection, the wiki served as a platform for rapid gathering of information from different sources. Although the teams used more sophisticated tools such as Zotero or Google spreadsheets for the majority of the collection effort, they still uploaded a selected number of highly useful resources to the wikis. Because wikis preserve information and make it easily findable with a simple search function, they seemed to be a good repository for resources. For most of the resources, analysts referenced the sources by hyperlinking to outside or internal sources, which is easily supported by wikis.

While the teams were collecting information, analysts started working on the analysis part. For intuitive analysis, a wiki was the main workspace for the analysis. Analysts created pages for their assigned topics and directly started writing into the pages. In the analysis phase, wikis provided transparency to both the teams and the clients. While working on their own pages, analysts could review others' work—each contribution and each modification in the process. Although a wiki alone did not lead to specific insights or findings, it was a medium for a better collaborative output. For structured analysis, the analysts primarily worked on Excel spreadsheets because they had to create matrices and fill out each cell. The analytical gains derived from using wikis were relatively small.

The production phase is all about wrapping up the project and putting on final touches to create a polished product. All the teams extensively used wikis in the production phase because they chose a wiki platform as a final deliverable. We observed many structural changes at the end of the project. Teams reorganized numerous pages they created, merged or deleted pages, and modified links to make the wiki easier to navigate without significant changes in the content. Because wikis are flexible and can easily adapt when requirements change or analysts prefer a different structure, wikis seemed to be ideal for production. Analysts also did a lot of formatting and editing on the wiki to polish the output.

Finally, the teams were able to create a final report on the wiki platform, documenting all of the analysis process, methods, resources, and key findings. Since a well-organized wiki format helped clients navigate the output, the teams created a table of contents for more effective navigation (Figure 5).

Wiki example: "National Intelligence Estimate". As mentioned earlier, we examined wikis that had been

Figure 5. Table of contents created for clients to navigate.

created in the past class projects in addition to the three wikis used in the projects we observed. To better illustrate how analysts organized the wikis and what kind of contents they put into the pages, we provide a detailed example of a wiki used for a "National Intelligence Estimate on infectious and chronic disease." While the wiki was created 5 years ago, it still has about 350 unique visitors per day (as of 22 February 2012).

In the project, a team of 26 graduate student analysts collaboratively worked on estimating important impacts/threats to US national interests resulting from infectious and chronic disease originating outside the United States over the next 10–15 years. The home page of the wiki contains a short description of the project, a navigational tutorial including a video, and special features such as terms and icons for readers to better understand the reports (Figure 6).

In the left pane, the default top menu has general actions related to the wiki such as "Pages and Files" and "Recent Changes," as well as the search box. The navigation pane on the bottom left lists important pages that the clients and other potential readers should visit, including "Home," "Terms of Reference," "Final Estimates," "National Interest Matrix," "Methods and Process," "Resources," and "Contact Information."

- In Terms of Reference, readers can find the key estimative question and secondary estimative questions of the projects, which define the scope of the analysis in detail.
- The Final Estimates page (Figure 7) contains the answer to the Key Estimative Question at the global, regional, national interest, and country levels. Each estimate page contains detailed explanation including an executive summary and discussion, ranging from one to five pages. The Global Estimate Video is also included in this page.
- National Interest Matrix is a page where the team embedded a matrix created in Excel spreadsheet. The matrix quantifies the impact of disease on US interests in each country and region round the world on a 5-point scale. As the team took a structured analysis approach, this matrix gives a good overview of how the team derived the Final Estimates.
- The Methods and Process page provides background information detailing the analytical processes that the team followed. It describes how the team developed the project to best answer the Key Estimative and Secondary Questions, allowing readers to understand the team's approach and procedures throughout the project period.
- The Resources page is a library of the works collected by the team throughout the project, including reports, journal articles, books, websites, and videos. The resources are also organized by geographic regions, so that readers can easily look up to the references.
- Finally, Contact Information lists all 26 analysts' email addresses and their roles in this project, so that readers could contact them with relevant questions and comments.

All the pages are connected through hyperlinks, allowing easy navigation between the pages. This particular wiki would have been over 1000 pages if it were a paper document. Using the wiki, the team seems to have effectively organized the contents of the report,

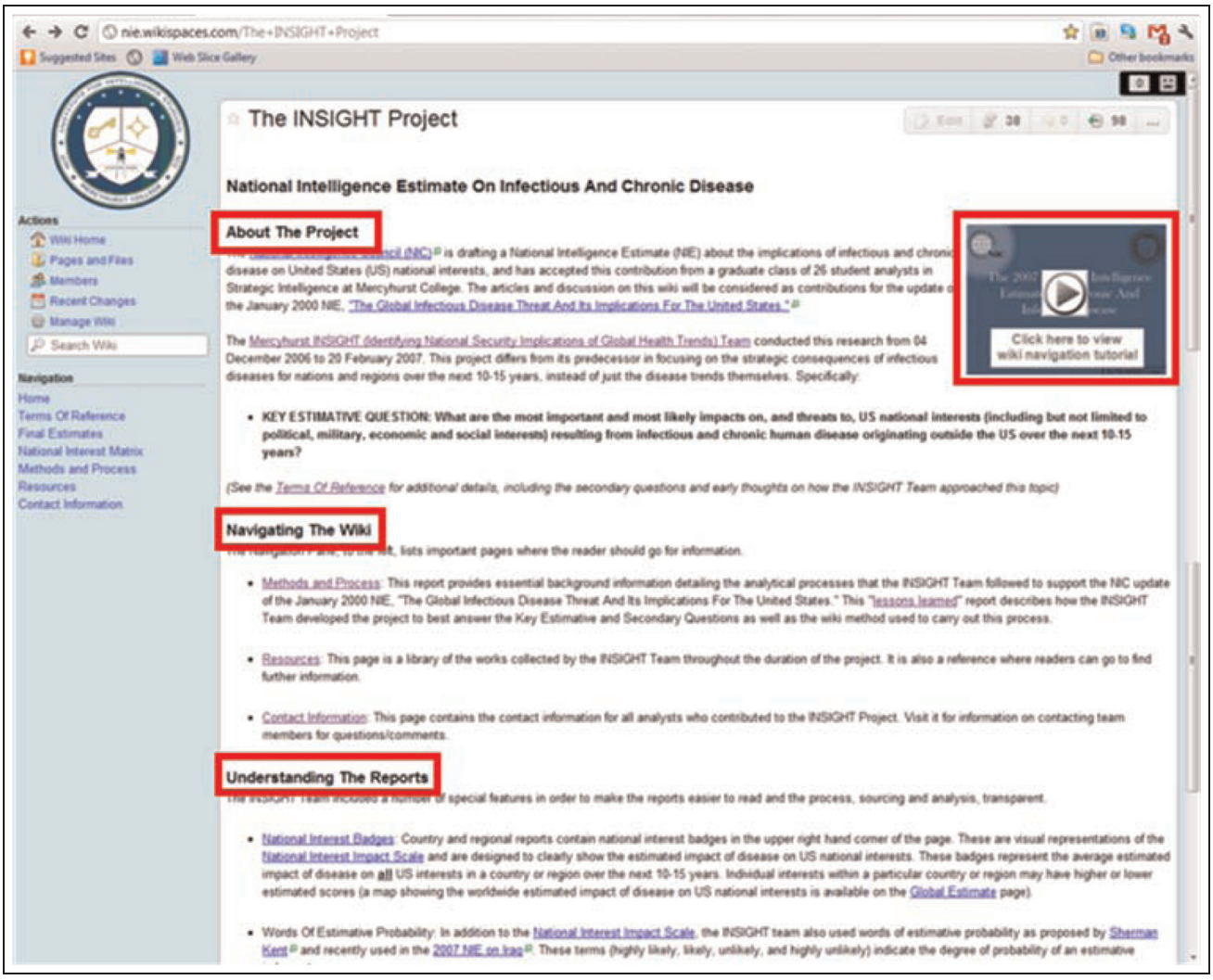


Figure 6. Home page containing "About the project," "Navigating the wiki," and "Understanding the reports."

so that readers can access the information they need more easily.

Understanding the intelligence analysis process

Observing analyst teams helped us to better understand their goals and processes. In particular, the study highlighted a number of misconceptions we harbored about the intelligence process. Other visual analytics researchers may or may not share these preconceived beliefs, but we think that they have the potential for misunderstanding and are thus worth exploring.

Intelligence analysis is about finding an answer to a problem via a sequential process

Some existing models of the intelligence analysis view it as an answer-finding process with a sequential flow, as noted in several models of the intelligence analysis process.[12,15,27] This perception presumes that the process is linear, sequential, and discrete by step. Pirolli and Card's[7] sensemaking model includes the notion of iterations and revisions between steps, but the fundamental assumption is that separate stages exist throughout the process and that analysts transition between stages.

However, this model was not the intelligence process we observed. Instead, the process appeared to be more parallel and organic, as one analyst described:

> Intelligence analysis is not about getting from point A to point B along the route, but it is better associated with basic research where you don't necessarily know where you are going to go. You're cutting a path through the jungle that's never been explored. That's what you're doing in most intelligence analysis projects. It's not a mechanical process in a sense that an assembly line is. It's

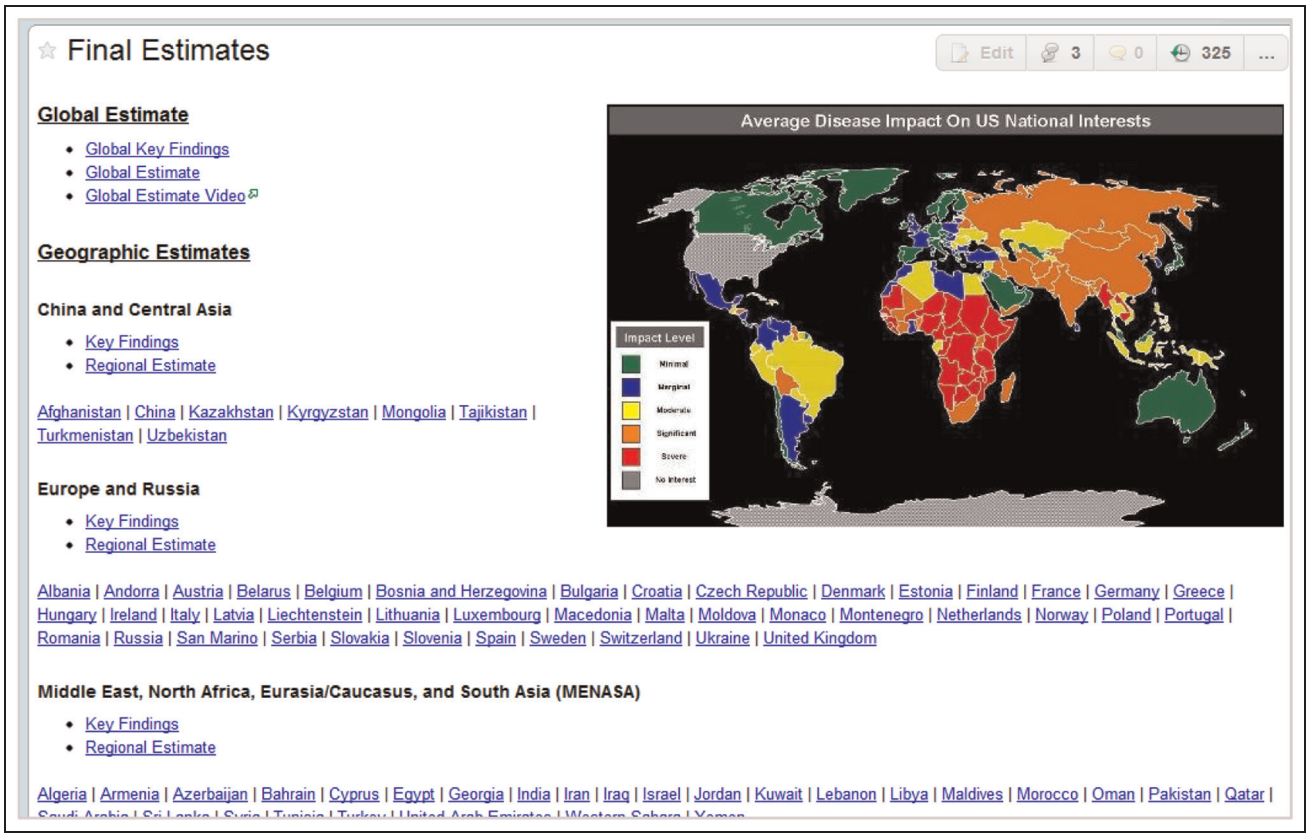


Figure 7. Final Estimates at the global, regional, and country level.

> a very exploratory activity by nature. You have to expect that some of the stuff you do, some of the things you think, you have to be willing to discard them. Because at the end, they will be rarely relevant, or you find something better that contradicts to it, and that's just the nature of it.

The key part of the intelligence process is the analysis of a specific set of data

Visual analytics systems often manipulate preprocessed data for the analysis. A primary misconception about intelligence analysis is that the data analysis process, in which investigators analyze a set of collected data, is the most difficult part and takes the most time. This belief assumes that the analysis occurs after investigators collect all data required for the analysis.

This view, however, needs to be changed. Although analysis is important, we observed that the process of "constructing a frame," as described in the data–frame theory,[28] is more important. In other words, intelligence is about determining how to answer a question, what to research, what to collect, and what criteria to use. This process becomes part of the analysis—analysis implicitly occurs during the process of the construction. Analysts also explore different sets of analytical techniques to address a problem. Deciding which method to use is important, but it often changes during analysis as the way that analysts think about the problem evolves, depending on information that constantly flows in.

Understanding that collection and analysis are integrated together in the process of building a frame is extremely important. Systems are not likely to be successful in supporting intelligence without acknowledging that fact. One analyst commented:

> Intell analysis is not like that you have a set of data in hand and run a program. It's like a conundrum from the very beginning. You have to learn how to learn, how to frame the question, and how to answer it through collecting and evaluating sources.

Analysts do not often collaborate

One common perception of intelligence views analysts as isolated individuals who prefer to work alone, struggling with pieces of information, rather than as

collaborative teams.[5] However, a faculty member at Mercyhurst countered this perception:

> Collaboration is almost all intelligence analysts have done in the context of the team. In the CIA or DIA, working as a team is pretty normal. While working on a particular topic within an agency is typical, also typical is working on an interagency team that consists of analysts from different agencies such as state department teams, DIA teams, and NSA teams.
>
> Analysts are normally organized by function or geographical region. These typically operate as loose teams. Strategic projects almost always involve a team as do crisis projects (for example I am sure there are multiple Libya teams that did not exist a month ago). In short, teamwork is the norm although the teams differ in the degree of formality and to the degree that there is a designated leader.

During our study, we also observed many collaborative elements of intelligence analysis. Collaboration is commonplace in intelligence analysis, and understanding how that occurs is important because it influences one's whole notion of the process. The IC itself has recognized the importance of improved collaboration since 9-11.[29] Although collaborative tools have been built and they are pushing users into tighter collaboration, it is still important to understand where tighter collaboration will be beneficial and where it may not help much.

We found that multiple layers of collaboration exist in intelligence analysis and that the degree of collaboration differs depending on the type of task and the group dynamics. We observed that analysts usually do not collaborate tightly on data and content—the actual collection and analysis. Although the teams had meetings frequently—twice or three times per week—the main purpose was to discuss their status, issues, and the next steps, as two analysts said:

> We come up with an agenda before meeting, a list of what we're supposed to talk about—what we did, what we want to do, what the questions we need to solve as a group. We didn't really plan that way, but it just happened. It's the way it is.
>
> There was no detailed, content-based work during the meeting although sometimes we had discussions on controversial issues. We basically do work in our own time, get preliminary ideas of it, and get together and discuss what issues we had.

The teams often worked together in the same laboratory, but it was rare that they worked on the same document or content. They worked on their own part, but they often talked to each other about issues and questions. Essentially, they took advantage of being in the same place at the same time. To better support analysts' work, we need to understand the collaborative aspects of the intelligence process and where technology can leverage collaboration. Some tasks are inherently done better by an individual.

We can help intelligence analysts by developing sophisticated analytical tools that assist their thinking process

Visual analytics researchers often seek to help intelligence analysts by developing technologically advanced analytical tools, thereby assisting their cognitive processes. The tools support specific types of analysis, specific analytical methods, and specific stages of the process. Such tools certainly can be helpful, especially to assist analysts to handle a flood of information.

However, our study revealed that analysts want something more than that. Currently, more than 50 analytical methods exist in the IC,[30] and analysts try many different kinds of techniques depending on the problem. Consequently, their dependency on a specific analytical technique is relatively low. Instead, the ability to manage the intelligence process effectively and employ various analytical methods and tools quickly is more important, as the instructor and an analyst said:

> Everything is fragmented. I've got Mindmeister here, Mindmeister doesn't interface with my search technology and Google reader. I've got to manually go out and figure out what all those bullets are. No help from a computer … But there isn't any set that ties all these, the pieces are there, Mindmeister, Zotero, RSS Google reader, MSWord, the wiki, are here, but nothing links all that in one seamless thing so I can go from the requirement to a product in a single package, in a single way.
>
> Doesn't need to be perfect at all. Needs to be able to jump back and forth and if somebody says to me, "Oh no, your product due tomorrow!" I've got to be able to take whatever system. I'm only in 2/3 of the process and I've got to be able to jump to the end of process and write the final report, get it done by tomorrow. That's the way intelligence works.

If the processes of collection and analysis are integrated in a single system, this helps analysts apply structured analytical methods such as ACH, social network analysis, geospatial mapping, and decision matrix. In our interviews, two teams mentioned that if they had more time, they would have tried other analytical techniques. Analysts always want to push their findings and triage, aggressively reshuffling their analysis. One of the most effective ways to do this is to employ multiple analytical methods and compare and contrast findings from each. The ability to try various techniques with the data can help analysts find effective ways for addressing questions and strengthening their analysis.

Rethinking the intelligence analysis process

Linear versus parallel

One might believe that the way intelligence analysts work is quite simple and straightforward. First, they specify requirements, build a conceptual model of what to research, then collect information, analyze data using various techniques, and finally write a report. This belief is a common misconception about intelligence as mentioned in the previous section. The reality is quite different. Rather than working linearly, analysts work on *everything* during almost the entire project. That is, analysts do not hold writing until enough information is collected; they keep revising analysis and writing as new information flows in. Analysts do not decide what to research and move on to collecting information; they start searching for information even when they are not sure what to research. Analysts do not produce final products after they are done with analysis; they already have an idea or a structure of final products in the very beginning, although it may be rough.

This "parallelism" is portrayed well in Wheaton's model of the intelligence process (Figure 8). In each phase, one of the core processes is emphasized most but all other functions operate in parallel. Wheaton[31] argues that "All four functions begin almost immediately, but through the course of the project, the amount of time spent focused on each function will change, with each function dominating the overall process at some point."

Although several distinct elements exist in the analysis process, all are very closely coupled and the connection is very organic. One can easily observe an analyst working on collecting new information while analyzing and checking the credibility of previously collected sources at the same time. In our study, we observed that a team's conceptual model changed drastically in the middle of the process, that a new information source was added 10 days before the deadline, and that a previous analysis report was discarded and new analysis began at a later stage. The matrices also kept changing as new information arrived. While the teams were working on the matrix, they were collecting information at the same time to make sure that they were familiar with the area. Several quotes better explain this:

> But it isn't as rigidly isolated as it's on that (traditional) cycle because you can't build a good conceptual model without knowing what's out there. So there's little bit of collection as you're building the model and we refined it.

> Our conceptual model is changing. It doesn't get set in phase 1 and we drive it, that's the difference between this process and an outline. An outline drives your production. But we are using it differently. As it changes, we're changing our analytic focus, we're making decisions about production, who's going to write something, who's going to do the analysis, based on how it's changing and that's being informed by new information that comes in.

> It's the most updated version. There's never final. We are constantly revising as we go along. You get to the point where it's ready to be in the final report, but then if you get new information that contradicts that, then that paper either has to be drastically edited or has to be thrown out entirely.

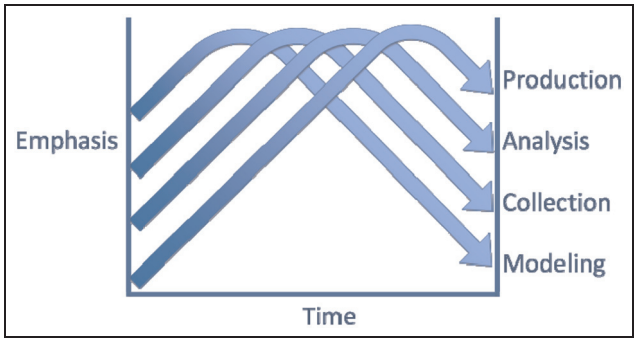


Figure 8. Wheaton's multiphasic model of the intelligence process.[28]

Pirolli and Card's sensemaking model

How does this new way of thinking about the intelligence process relate to Pirolli and Card's[7] sensemaking model? Because it is the most widely used model in the visual analytics domain, we were curious about how well their model explains real-world intelligence analysis processes.

Pirolli and Card's model provides new insights into the intelligence process, suggests leverage points for analysis tools, and has guided the design of many visual analytics systems. However, we argue that the model still implies sequential, discrete stages of the intelligence process although it acknowledges that analysts can move either top-down or bottom-up or jump to different stages. For example, the model does not explain why analysts so frequently jump from one state to another state that is not adjacent. Many visual analytics tools thus support specific states only (e.g. shoebox and evidence file, evidence marshaling, and foraging), and often they do not blend into the entire process of intelligence analysis.

More importantly, the model describes how information transforms and how data flow, rather than how analysts work and how they transition. It gives a very insightful illustration of how the form of information

evolves from raw data to reportable results. However, it does not quite fit analysts' mental model of their work process because they do not work as information is transformed. Rather, information is transformed by how analysts proceed. Similarly, all different states of the model can exist at any point during the process. Analysts may have polished reports on certain subtopics, drafts of analysis, structured matrices, and a collection of documents at a time.

The Pirolli–Card model identified various leverage points for visual analytics tools, but the linearity of the model could give researchers an inaccurate impression of the process. While models are inherently abstract and stage based, it is important to understand the context and the purpose of the model. We would characterize their model as more of an information-processing process rather than intelligence analysis process. Pirolli–Card explicitly state that the model was suggested as a starting point to investigate the domain. While it has contributed to visual analytics researchers understanding of the domain, now we need to change our assumptions to build systems that better help intelligence analysts with their work.

Where and how collaboration occurred

Collaboration throughout the process

Throughout the project, the teams worked tightly together although the degree to which they collaborated differed depending on the phase of analysis. Once the project started, the team set up weekly meetings. The first thing they had to decide was to specify requirements of the problem and then they collaboratively worked on building the conceptual model. Whether the team kept using this model or changed to a matrix, it played a role as a representation of their "group thinking," as an analyst described:

> You want to say that this is the way I'm thinking about this problem. These are some of things I need to think about. And what we've done by building the conceptual model is to have that sort of group interaction, which is not necessarily harmonious action. There can be disagreements about how we should be thinking about this. And if there's shifting, moving it around, that represents an evolution of the way of our thinking.

Once the team had an idea of the areas to explore, they divided up the work and assigned concepts to each analyst. While each one worked on different concepts, they collaborated in collecting information by using a group library. Although this seems to be loose collaboration, the benefit the team gained was invaluable because it could significantly save time and effort in collection. An analyst explained how they worked in collection using Zotero:

> Zotero is a good example of one way we collaborate. Each person creates a group library on the Zotero server. If I find a website that I think is useful, whether for my topic or someone else's topic, I add it to our group collection, and then other members can see it before they go searching the Internet for something. And if she doesn't find that in Zotero, then she might go out Google. So … try Zotero first, you might already have it.

While working on and analyzing their own topics, team members often met with each other to check status and discuss issues. When not all team members were available, they used typewith.me,[32] a web-based collaboration tool for writing. When most of the areas they had planned to explore were covered and analyzed, they collectively wrote the key findings—the crux of the analysis project. Very tight collaboration occurred in this work. They met together and spent significant time to synthesize findings from all the topics and write the key findings.

Sharing versus content versus function

We found that three different types of collaboration exist when analysts discuss the topic: sharing, content, and function.

Sharing is a way to collaborate by sharing information. In our study, analysts shared sources to better assist their search process and understanding of the topics. At a higher level, however, this can be the sharing of analytical products as well as information sources. When people mention the importance of collaboration in the IC, they primarily refer to the sharing of information sources and analytical results.[33] This type of collaboration can be significantly supported by technology.

Collaboration also occurs at the *content* level. This type of collaboration, in which analysts work together to create analytical products, can be seen more often in a small-sized team. Examples in this project include constructing a conceptual model together, dividing concepts and assigning to each analyst, commenting on each other's analysis, working on ACH together, and writing the key findings together. However, in our study, once work was divided, then each part was done individually. The degree of tightness in this type of collaboration may directly affect the quality of analysis. The more closely the team works together, the more that output is coherent and logical. However, in reality, it is difficult to collaborate on content because of efficiency. This type of collaboration is also difficult to facilitate via technology because so many subtle

issues—such as social dynamics, politics, teamwork, and motivations—are involved.

Functional collaboration is needed to execute practical tasks for completing the project, such as editing, creating a matrix structure, specialized analysis on a specific topic, and polishing deliverables. While analysts work on the same thing and divide up the analytical product in the content level, functional collaboration naturally emerges at the later stage of the process as the team begins to think about allocating multiple functions. In this type of collaboration, analysts reinforce their strength. For example, if one is a good editor and has a detailed eye, then that person would do the editing, as one analyst explained:

> There was a lot of collaboration. A spent a lot of time working on wiki stuff. B spent time doing the ACH stuff. C did a lot of technical stuff. So each of us spent extra amount of time doing something specific. Whichever parts of this you want to take on, those are the parts that get divided up. As different as I do this (analysis), I do this concept, I do this concept.

Olson et al.[34] similarly characterized collaborative activities of groups by analyzing design meetings from four software teams. Focusing on the time spent in meeting activities, they found similar patterns across design teams. In meetings, teams spent 40% of the time in direct discussions of design, 30% of meeting time was spent taking stock of their progress, and coordination activities consumed approximately 20% of the time. Clarification of ideas across these activities took one-third of the time, indicating that participants spent a large amount of time sharing and explaining expertise.

How wikis facilitated the analysis process

Previously, we described how analysts used wikis throughout the analysis. While wikis do not seem to provide analytical support per se, we examined how wikis assist the analysis process more broadly. The instructor commented that he had observed projects over several years and teams were more productive when they adopted the wiki as early as possible. What characteristics of wikis helped the process and offered productivity gains?

Low initial barriers, easy to use, and compatible

The primary reason that the analysts liked to use wikis is that they are easy to use and work well with other software. Analysts are often pressed for time. While a number of sophisticated systems exist, analysts do not have plenty of time to learn about and get familiar with a system. If a system has a high learning curve, the chances are rare it will be adopted by analysts. Analysis also occurs across several tools. Results and information snippets often need to be integrated into a single system, and wikis seem to do a reasonable job in embedding outputs from other applications such as YouTube videos or Google docs. In our study, the teams noted:

> We chose Wikispaces as our preferred wiki platform because it is relatively easy to use and it works well with various external applications. For example, it allows for the embedding of You Tube videos or Google Documents. It is easy for new users to learn rules that allow for the basic editing of pages. More advanced users can use the text editor which is part of the wiki, or the widget function called "Other Code" to change or add code in order to add additional sophisticated functionality to wiki pages.

Even though wikis have relatively low initial barriers, analysts who are not familiar with such applications would have adoption issues and stick to traditional approaches. Once they overcome the initial barrier, however, the productivity gains seem to be significant, as shown in the following quote:

> When they choose to continue more traditional style work flow, that has everything to do with the adoption problems—classic problems in wikis. Two factors that are important in that are "how easy is wiki to use" and "how reliable it is." Once analysts overcome the problems and adopt it early, the productivity gains are huge.

All-in-one-place nature, from requirements to production

The analyst teams used the wiki platform as a main tool throughout the entire project—from requirements to production. In the beginning, the wiki served as a structural, organizational space for the remaining analysis. Then it became a living document for work in progress, ultimately resulting in a final deliverable. Although the analysts employed other tools at some point when they needed a specific analytical functionality, the wiki stayed as the major place of their work.

This all-in-one-place nature allowed the analysts to shift their attention easily between the four functions (phases) of the process. By capturing most of the process up to date, the wiki serves as an effective platform for starting collection efforts, analyses, and even finished products.

The ability to support production seemed to be attractive to both analysts and decision makers. They

can export the wiki as a printable format such as pdf and use either an electronic or a printed version of the product.

Focus on the process as well as the result

Since the team members (and possibly decision makers) consistently reviewed and monitored the wiki pages throughout the project, they were able to focus on details in every stage of the process. That is, wikis reflect the reality of the intelligence process, allowing the teams to focus on the process as well as the result of the intelligence.[35] This is the particular advantage the wikis provided compared to other systems.

Transparency in process

Wikis provide transparency in the analysis process. For example, everyone can access the history of edits including edits on pages, files, and tags. Being able to see the history of edits increases accountability for the analysis process. The instructor commented on this feature:

> This team made about 2,000 edits so far. Go and find, and take a look at how many edits were made in the key findings section. I like to watch the history of important documents like key findings. You can look at the initial version, you can look at the middle version, you can look at the final version, and you can watch the evolution of changes, much as same as Wikipedia. Same kind of thing happens here. An important document gets seen or edited a lot; you can actually watch the team think. That's pretty cool.

Editors and readers are often interested in the history of edits and like to compare the changes made, particularly for important pages, similarly as done for Wikipedia.

Platform for asynchronous, calm collaboration

A single analyst working on a project without colleagues may see little benefit in using a wiki, other than the fact that it is accessible anywhere. However, as described in the previous sections, almost all analytical processes are collaborative among many people.

The benefit of wikis is maximized when the team size is relatively large. Most of the analyst teams we studied had about four or five members, which might have been manageable without wikis. Still, the teams told us that wikis were very effective for collaboration by making the collaboration status—who is working on what and how much progress has been made—visible.

Wikis also encouraged discussion and communication of ideas and questions among analysts. Since analysis should be objective and incorporate different perspectives, these collaborative features of wikis helped the analysts accommodate varying viewpoints. In addition, wikis allowed each individual to easily review and edit others' work. This peer-review process provided an opportunity for analysts to challenge assumptions, double-check sources, and identify areas that need further research. Ultimately, with repetitive peer reviews to the same article, the bias and inaccuracies could be minimized. One analyst commented about this collaborative support of wikis:

> Wikis don't make me faster, they make us faster.

Flexibility in structure: easy to reorient anything

It is hard to structure a site without knowing its contents and flow. Since analysis is an ongoing process and requirements and collection needs often change, analysts cannot predict and arbitrarily create pages in wikis. That is, the table of contents and organization of pages keep changing until the very end of the analysis.

Wikis have a very important characteristic that can help with this matter; they provide the analyst teams with enormous flexibility. Using traditional approaches, it is quite difficult and cumbersome to reorient everything, but wikis support this very efficiently.

Analysts in our study often changed the structure of the wiki pages, even at the end of the process, but they did not find it difficult. Because all the information and writings the team had created were stored as separate pages by topics, it was relatively easy for them to add/remove, change structures, and modify links. The flexibility inherent in the wiki seemed to allow for easy reorganization of the pages.

How wikis are used for collaboration—domain comparisons

While a handful of research publications have investigated wikis and their usage, an article by Fuchs-Kittowski and Köhler[35] provides a good explanation of why wikis are good at cooperative knowledge creation and exchange. He argues,

> With the help of a wiki, users can easily gather and integrate knowledge into the existing (wiki) knowledge base by the user. The particular advantage of the wiki approach compared to other cooperative knowledge generation and exchange systems is the focus on the *process* as well as the *result* of communication.

For example, content and document management systems tend to focus on the exchange of results of tasks

done by several people. Discussion boards, in contrast, focus mostly on the cooperation process, making the exchange of opinions possible. The results of the discussion are generally implicit in the individual postings and have to be abstracted afterward. Wikis, in contrast, allow users to discuss and work on the result simultaneously. That is, cooperative production of content becomes very efficient through "the realignment of the distinction between author and reader." This characteristic, also discussed in the previous section, significantly benefited the intelligence analysis process in our study. Intelligence analysis consists of knowledge-intensive tasks and work processes and is characterized by nonlinear sequences and dynamic social interaction. A wiki is an appropriate interface in that it supports the knowledge creation processes.

Many researchers have explored how wikis are used for collaboration, especially in corporate settings. These studies[36–40] reported that wiki technology was used to support a wide range of work activities within a corporation, including team collaboration, project management, information dissemination within communities of practice, idea generation, project planning, and e-learning.

The focus of previous research studies regarding wikis has been adoption and sustainability. For example, a survey of 168 wiki users conducted across diverse firms indicated that corporate wikis were sustainable "based on the length of wiki existence, the number of participants, the number of lurkers, and the frequency of accesses."[39] In the research, authors identified three types of benefits achieved through the participation and use of corporate wikis: "benefits to enhanced reputation, benefits to making work easier, and benefits to helping an organization to improve its processes." They argue that benefits are more likely perceived when work tasks require novel solutions (rather than routine tasks) and when other wiki contributors are believed to provide credible information. While our study was not conducted in a corporate setting, it seems that the characteristics of the tasks our analysts conducted and the tight group dynamics were quite suitable to wikis and helped analysts perform the work. Their study also identified three groups of wiki contributors—adders (who add pages and content), synthesizers (who integrate, reorganize, and rewrite whole paragraphs), and commenters (who comment and make small corrections). We found this classification interesting and somewhat congruent with our study results. In the previous section, we discussed that some members spent more time in practical tasks such as editing and reorganizing—synthesizers. However, we did not see a clear distinction between the three groups of contributors in our study since all members actively participated in adding pages and content and commenting on each other's work. We assume that this might be because of a small group size, where the level of contribution is even more transparent. In Arazy et al.'s[36] study, respondents also rated highly the direct benefit of wikis in supporting their work and enhancing their productivity, and the benefit seemed to be correlated with their level of proficiency in using wikis.

Phuwanartnurak[41] describes a field study of two interdisciplinary design teams, seeking to discover how wikis support information sharing in software development projects in a university information technology (IT) department. The study provides evidence that a wiki is simple and easy—all members from both software development teams liked the wiki because it is easy to use. Our study also proved that analysts favored low initial barriers and easy-to-use features of wikis. He also reports that wikis encouraged more information sharing among team members during the software development process and made the work more visible to other disciplinary groups, which were apparent in our study. His study had interdisciplinary teams, and he observed that roles and tasks have a lot of influence on how design team members would use a project wiki. Although we did not examine that aspect specifically, this will be worth exploring for future research.

Egli and Sommerlad[42] report the experiences of a law firm with adopting wikis for knowledge management and collaboration over 2 years. They describe that wikis created a business advantage for the lawyers through better reuse of their know-how within the firm, and external wikis for clients created new revenue opportunities and higher client satisfaction. They divide their internal wikis into three groups—collaboration wikis (used for internal purposes only), information-display wikis (open to clients), and personal wikis (initial wikis that are not yet developed into a collaboration wiki). In our study, analysts used a wiki as a workplace for collaboration and also as a final deliverable for the clients, which means that the entire process of analysis was transparent to the clients. This difference seems to be due to the characteristics of the work and the degree of confidentiality. As the advantages of wikis, Egli and Sommerlad point out the easy establishment of new instances and the level of speed and flexibility and emphasize how wikis can be made useful as a collaboration tool, as shown in the quote:

> The real power of Wiki is fully revealed as a collaboration tool. By sharing your thoughts with others you enable them to develop their own ideas starting not from scratch but from a certain level. Their thoughts then again help you to have new ideas. The result of this mutual feedback process is more than the sum of all single ideas. It's something new that couldn't have been created without Wiki.

> As already mentioned this does not only work within a team but also when you use it for our very own purposes. An idea you had as a 35 year old person might be looked at from a very different angle at the age of 45.

The study also identifies the disadvantages of wikis, one of which being the lack of a task management tool and schedule. It is crucial when using the system as a workflow management tool, that is, a system administrating a legal file's documents as well as the related tasks and deadlines.

The use of wikis to support distributed collaborative writing has been also investigated.[43–45] Chi et al.[45] developed Dandelion, a tool that extends wikis to support coordinated, collaborative authoring. Through real-world pilot trials, they found that the system was especially useful in structured, collaborative authoring situations with designated coordinators, where the role of a coordinator is clear and the document outline is relatively known. For free-formed collaborative authoring, Dandelion added little on to a standard wiki.

Several studies focused on challenges in the adoption and use of wikis.[46,47] Holtzblatt et al.[46] explored factors that impacted the use of wikis within a corporate environment and discovered two major factors that contributed to staff's unwillingness to share information on a wiki—(1) a reluctance to share specific information due to a perceived extra cost, the nature of the information, the desire to share only finished content, and sensitivities to the openness of the sharing environment and (2) a heavy reliance on other non-wiki tools based on work practice, lack of guidelines, and cultural sensitivities. Grudin and Poole[47] explored where, how, and why people use wikis at work, focusing on the creation and use of wikis on corporate intranets in scientific and engineering organizations. They identified challenges in adoption and long-term sustainability of wikis, which include (1) mismatch between management visions and the benefits delivered by successful wikis, (2) limitations of current wiki-based tools that impact long-term use such as the difficulty of reorganizing information, (3) the disruption of bringing another technology for communication and information sharing into environments with established practices, and (4) unfamiliarity with the collaboration model inherent in wikis, including uncertainties about accountability. The authors conclude that wikis may be most successful in supporting newly established groups or short-term activities.

How visual analytics can help: design implications

How can visual analytics help intelligence analysis? Based on our study findings and reflections, we suggest several design implications for systems supporting intelligence analysis.

Externalize the thinking process—help analysts continuously build a conceptual model

> Good analytic practices encourage continuous improvement upon the conceptual model throughout research, which continues through the end of the project.

The analysts in our study told us that the process of making sense of a problem and building a conceptual structure is one of the most important parts of intelligence analysis as it decides the direction of analysis. In most cases, analysts encounter a situation in which they need to learn about new subject matter, but it takes time and effort until they become familiar with the domain. Because they cannot build a good mental model of the problem without knowing what information is available, they struggle to know more about the domain until the later stage of analysis.

Using the power of representation, visual analytics systems can help analysts build a conceptual model or a structure of the problem and domain. For example, the system can take the main question the analyst has and suggest a number of possibly related concepts and keywords based on online encyclopedias, table of contents of books, and tagging services. The system should allow the analyst to refine the concepts, so that it can repeat the search and suggest other relevant concepts. By connecting, grouping, and organizing concepts, analysts can continuously build up their conceptual model or structure of the area throughout the process. One analyst cited experience:

> Ok, I got to model something, I've got to do a report on Ghana, I don't know anything all about Ghana, where's the tool that if I hit the button, it gives me a picture of what the relationship is, the model how to think about Ghana? It gives me 60-70% of the solution. But it gives me the ability to input and tweak and change those. Because I want to have a role in that, I can't allow the computers to do all my thinking, you know.

Support for this externalization should occur throughout the analysis process because as analysts learn more about the domain, they alter their way of thinking and refine their visual model. Externalizing the thinking process can also assist analysts when they review their analysis after the project terminates. Supporting this activity would be especially useful because it will inform how the analysts could have done better and the areas that need to be examined if they did a similar project, as the instructor said:

> The other thing this model helps you do is at the end of the project you can look back and go, "What did we not have time to do? And how does that impact our company, our estimates?" Because whatever reason we didn't get to it, this was important, we thought this define the space … We can sit back and go, ok, how confident are we on our estimates, knowing that our analysis is always at some level incomplete? And it's always incomplete, but how does it impact our confidence in our product? That's another way to use this representation.

Support source management—enable managing both pushed and pulled information and organizing sources meaningfully

One prominent characteristic of how analysts think about sources is that they have to be always vigilant of new sources. They often search for the same keywords again to see if any new materials have been added regarding the topic (pulled sources). They also receive news articles through RSS feeds every day and check if they have received interesting information (pushed sources).

This process of searching sources takes more time than one may think, and systems should allow analysts to manage both pushed and pulled information associated with concepts they have identified. For example, a system could populate several concepts chosen by the analyst and store all the pulled sources in a database such as Zotero. Based on sources already found, the system also could recommend push resources such as blogs and news articles. For each source collected, the analyst could express if it is a useful source or not. Analysts commented on this functionality:

> Sources are what we have to get, but where is the tool where I can integrate them? My RSS feeds dump into me every morning. But then I do searches as well. Where's the tool that allows me to integrate all data, the information that is useful for me?

> If that kind of system exists, I have the ability to go back and find all my sources. Automatically, this (keywords, phrases) gets populated. And every point, I have the ability to say no or yes, no or yes to a source. But the actual extraction or the pulling, and the organization of that are automatic from that.

Then the list of sources can be organized in a meaningful way—for example, by keyword queries, by tags the analyst annotated, or by date the source was added. The system could also provide several ways of representing source results such as summary and tag clouds. Further support for analysis or visualization of collected sources as a group would be extremely beneficial.

All these technical capabilities currently exist in visual analytics systems. Now, it is important that they be integrated together appropriately.

Support analysis with constantly changing information—integrate collection and analysis in a single system and help analysts use structured methods during collection

As described in the previous section, collection and analysis are not separate but highly integrated processes. Analysts do not wait until all the data are gathered; rather, they start analysis even when they have only a few pieces of information. Through the repeated process of collection and analysis, they revise a frame and use the collected data as supporting evidence for the frame.

One of the reasons why wikis were successfully used in our study was because of its flexibility in structure—it is relatively easy to reorient anything in wikis. While requirements and collection kept changing throughout the process, Wikis provided the analyst teams with enormous flexibility and allowed them to easily restructure the content of the pages.

Currently, many systems provide analytical support assuming that processed data are available. If a system does not support a seamless transition between collection and analysis, it is likely to be less successful in assisting the analysis. Analysts collect during analysis, and they analyze during collection. This differs from the statistical analysis, in which a structure or a frame about how to analyze the data is clearly defined and analysis is done with clean dataset. An analyst mentioned:

> If they had more reliable, structured data, I'd use statistical analysis. But intelligence data is unstructured and dirty. You don't know what the best way to analyze it is until the middle of the process, or even the end of the process.

Multiple visual analytics systems provide analytical capabilities. By supporting more flexible data manipulation, so that analysts can easily import and remove data from the analysis pool, these systems will be more usable, with better integration into the analysis process.

If the processes of collection and analysis are integrated in a single system, this helps analysts apply structured analytical methods such as ACH, social network analysis, geospatial mapping, and decision matrix. In our interviews, two teams mentioned that if they had more time, they would have tried other analytical techniques. Analysts always want to push their findings and triage, aggressively reshuffling their

analysis. One of the most effective ways to do this is to employ multiple analytical methods and compare and contrast findings from each. The ability to try various techniques with the data can help analysts find effective ways for addressing questions and strengthening their analysis.

> We had this time crunch. We pretty much got rid of the process of re-evaluating our hypothesis, finding what's the most important to make it perfect, and hitting on that, and going back to the stuff that we didn't deem as important. If we had time, we would fill that in.

Help analysts create convincing production—support insight provenance and sanity checks of analytical products

Production is what differentiates intelligence analysis from general sensemaking, which does not necessarily entail external representation. Even when analysts finish their analysis, they need to convert the results into a concise format, so that decision makers can understand their findings. This can be a tedious and time-consuming part of the intelligence process.

When asked about the most difficult part of their project, two teams mentioned production. Interestingly, this difficulty comes from sanity checking and insight provenance, not simply from formatting and writing issues. The sanity check, or qualitative double-check, takes time because data and findings are derived from many sources and analysts have meshed them through the process of collection and analysis. Analysts need to return to original sources and provide a rationale by which their statements are made. They also have to add references to their statements, for which they have to revisit original sources. The following quote from an analyst illustrates those difficulties:

> Most difficult part … basically going back through all the sources we used to grade these technologies, people, and companies, then taking basic pieces from those and making a narrative out of it. So explaining why we thought they are the keys and then relating it to the rest of the other findings.

A system that promotes simple insight provenance during analysis could help analysts save their time in production.

Support asynchronous collaboration rather than synchronous collaboration for exploratory analysis

We discussed three different layers of collaboration in the intelligence process and that the degree to which technology can contribute varies. In particular, visual analytics systems seem to have the potential to help collaboration in "sharing" and "content."

From our study, we found that these types of collaboration tend to occur asynchronously, rather than synchronously. When meeting face-to-face, analysts did not work on actual tasks but spent time checking their status, coordinating next steps, and discussing issues. Even when they worked in the same laboratory for several hours, team members took their own computer and worked individually. Although they often talked to each other, it was for simple coordination issues or specific questions about the content. One analyst stated about his perception on collaboration:

> We discussed how each of us interprets the data. We're very group-oriented when it comes to discussing to a consensus. Other than that, we prefer to work individually especially for the actual analysis. Of course we collaborate even when we work on our own parts, but there's no one who really knows about those concepts or entities like you do.

In our study, wikis were effectively used for asynchronous, calm collaboration—by making collaboration status visible, encouraging discussion and communication of ideas, and allowing individuals easily to review and edit others' work.

In a nutshell, analysts collaborate cognitively. Rather than trying to build a system that allows analysts to work at the same time in the same workspace, providing a system that not only promotes individual workspaces but also provides asynchronous collaborative features—such as the ability to share sources and data, view and comment on others' work, and merge individual work together—would appear to be more beneficial.

Note that our findings are based on strategic intelligence. In other types of intelligence such as tactical and operational intelligence, which form the basis for immediate action, real-time collaboration is also important because such intelligence must be shared and used quickly.

Unifying the pieces

Because their typical processes of requirements gathering, collection, analysis, and production are so intertwined, and it takes considerable time to coordinate between different software systems, it appeared to us that analysts want an all-in-one system that can streamline the analysis process and save their time. In our study, one of the most important reasons for using wikis in the intelligence analysis was because the all-in-one-place nature of wikis allowed the analysts to

shift their attention easily between the four functions (phases) of the process. By capturing most of the process up to date, the wiki serves as an effective platform for starting collection efforts, analyses, and even finished products. When asked about their "dream" system, a few analysts answered:

> If I had to go back to the beginning and start all the way over, I should be able to jump back and forth seamlessly between all of these processes. We need a tool that compensates for that.
>
> It should be one program. We spend more time to make it work together. Nothing's compatible with others. We want a program that syncs all the documents. Help us do our visualization with the documents. A program that is compatible with Excel spreadsheet. Don't want to open 20 different programs.

Thus, a hypothetical tool that simplifies the intelligence analysis process would function as follows:

- The analyst enters requirements into the system.
- The system suggests various concepts associated with key terms, phrases, and ideas in the requirements.
- The system not only automatically draws connections between concepts but also allows the analyst to draw connections, group, and organize them.
- The system takes the concepts and starts populating them, collecting information sources using the concepts as keywords (pull sources).
- The system uses sources the analyst identified and suggests new articles relevant to the sources (push sources).
- All these pulled and pushed sources are integrated into a source repository.
- For documents in the database, the analyst can highlight important facts and annotate his or her thoughts. On demand, the system extracts entities requested by the analyst.
- For intuitive analysis, the analyst can write reports in a preferred format, walking through each document.
- For structured analysis, the system helps the analyst try a variety of structured methods. It takes all the information identified by the analyst and integrates it directly into the methods.
- At the end of the process, when the analyst produces final output, the system automatically links each statement to relevant sources and the process by which the statement was derived.

Thus, analysts could flexibly move between conceptual model, collection, analysis, and production. The system accompanies the analyst from requirements to product in a single platform, speeding up the process, as expressed in one analyst's comment:

> If I had something like that, I'd be blazingly fast. I mean I would be able to do this 10-week project in three weeks.

Interestingly, our suggestions reiterate the findings of other researchers who identified the importance of unifying disparate tools in a different domain. In an observational study of the scientific data analysis process, Springmeyer et al.[48] concluded that "an effective data analysis environment should provide an integrated set of tools which supports not only visualization, but some of the additional functionality" such as capturing the context of analysis and linking materials from different stages of analysis.

Conclusion

In this article, we described an empirical study to understand intelligence analysts and their processes. We observed three teams of student analysts working on typical intelligence problems. Our contributions include documentation of the processes and methods they followed, clarification of issues regarding the intelligence process, and design implications for visual analytics systems for intelligence analysis.

The study has several limitations. We followed only three teams (14 analysts). Also, the analysts were not working professional analysts but were student analysts in training. The analytical questions studied were from strategic intelligence, one type of analysis. Possible future work includes the study of more cases, particularly with professional analysts working on similar or other types of intelligence problems. Of course, the design implications can serve as motivation for new visual analytics systems, ideally created through participatory design with analysts.

Funding

This work was supported by the National Science Foundation under award IIS-0915788 and the VACCINE Center, a Department of Homeland Security's Center of Excellence in Command, Control and Interoperability.

Acknowledgements

We thank Professor Kristan Wheaton, Department of Intelligence Studies, Mercyhurst College, for his support and input on this article. He helped us contact and study the student analysts and provided valuable comments that increased our understanding of their work process. We also thank the three teams of

analysts for sharing their work and opinions during the study.

References

1. Bier E, Card S and Bodnar J. Entity-based collaboration tools for intelligence analysis. In: *IEEE conference on visual analytics science and technology (VAST)*, Columbus, OH, 19–24 October 2008, pp. 99–106. Piscataway, NJ: IEEE Press.
2. i2—Analyst's Notebook. http://www.i2group.com/us (accessed 15 March 2011).
3. Stasko J, Gorg C and Liu Z. Jigsaw: supporting investigative analysis through interactive visualization. *Inf Vis* 2008; 7(2): 118–132.
4. Wright W, Schroh D, Proulx P, et al. The sandbox for analysis: concepts and methods. In: *ACM SIGCHI conference on human factors in computing systems*, Montreal, QC, Canada, 22–27 April 2006, pp. 801–810. New York, NY: ACM Press.
5. Chin G, Kuchar OA and Wolf KE. Exploring the analytical processes of intelligence analysts. In: *ACM SIGCHI conference on human factors in computing systems*, Boston, MA, 4–9 April 2009, pp. 11–20. New York, NY: ACM Press.
6. Johnston R. Analytic culture in the United States intelligence community: an ethnographic study. http://oai.dtic.mil/oai/oai?verb=getRecord&;metadataPrefix=html&identifier=ADA507369 (2005, accessed 12 October 2012).
7. Pirolli P and Card S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: *International conference on intelligence analysis*, MacLean, VA, May 2–4, 2005.
8. Robinson A. Collaborative synthesis of visual analytic results. In: *IEEE conference on visual analytics science and technology (VAST)*, Columbus, OH, 21–23 October 2008, pp. 67–74. Piscataway, NJ: IEEE Press.
9. Munzner T. A nested model for visualization design and validation. *IEEE T Vis Comput Gr* 2009; 15(6): 921–928.
10. Kang Y and Stasko J. Characterizing the intelligence analysis process: informing visual analytics design through a longitudinal field study. In: *IEEE conference on visual analytics science and technology (VAST)*, Providence, RI, October 23–28, 2011, pp. 21–30. Piscataway, NJ: IEEE Press.
11. Gotz D, Zhou MX and Wen Z. A study of information gathering and result processing in intelligence analysis. In: *Workshop on intelligent user interface (IUI) for intelligence analysts*, Sydney, Australia, January 29 – February 1, 2006.
12. Krizan L. *Intelligence essentials for everyone*. Joint Military Intelligence College, Occasional Paper no. 6, June 1999, Washington, DC: Joint Military Intelligence College.
13. Clark R. *Intelligence analysis: a target-centric approach*. Washington, DC: CQ Press, 2004.
14. Heuer R. *Psychology of intelligence analysis*. Center for the Study of Intelligence, Washington, DC: Central Intelligence Agency, 1999.
15. Central Intelligence Agency. *A consumer's guide to intelligence*. Washington, DC: Central Intelligence Agency, 1993.
16. Dearth DH. National intelligence: profession and process. In: Dearth DH and Goodden RT (eds) *Strategic intelligence: theory and application*. 2nd ed. Washington, DC: Joint Military Intelligence Training Center, 1995, pp. 15–31.
17. Department of Intelligence Studies at Mercyhurst College. http://intel.mercyhurst.edu/ (accessed 15 March 2011).
18. National intelligence estimates. http://www.cfr.org/iraq/national-intelligence-estimates/p7758#p3 (accessed 7 June 2011).
19. Heidenrich JG. The state of strategic intelligence. *Stud Intell* 2008; 51(2).
20. Charmaz K. Qualitative interviewing and grounded theory analysis. In: Gubrium JF and Holstein JA (eds) *Handbook of interview research: context and method*. London, UK: SAGE publications, pp. 675–694.
21. Millen DR Rapid ethnography: time deepening strategies for HCI field research. In: *ACM conference on designing interactive systems*, New York, 17–19 August 2000, pp. 280–286. New York, NY: ACM Press.
22. Strauss A and Corbin J. *Basics of qualitative research: grounded theory procedures and techniques*. Newbury Park, CA: SAGE, 1990.
23. Mindmeister. http://www.mindmeister.com/ (accessed 15 March 2011).
24. Zotero. http://zotero.org (accessed 15 March 2011).
25. Norman DR. Websites you can trust. *Am Libr* 2006; 37(7): 36–37.
26. Peterson JJ. *Appropriate factors to consider when assessing analytic confidence in intelligence analysis*. Master Thesis, Mercyhurst College, Erie, PA, 2008.
27. Treverton GF. *Reshaping national intelligence in an age of information*. Cambridge: Cambridge University Press, 2001.
28. Klein G, Moon B and Hoffman RR. Making sense of sensemaking 2: a macro cognitive model. *IEEE Intell Syst* 2006; 21(5): 88–92.
29. Eli AB and Hutchins J. Intelligence after Intellipedia: improving the push pull balance with a social networking utility. Research Report from Information Science and Technology Directoratet, Defense Technical Information Center, Fort Belvoir, VA, February 2010.
30. Heuer RJ and Pherson RH. *Structured analytic techniques for intelligence analysis*. CQ Press College, 2010, Washington, DC: CQ Press.
31. Wheaton K. Wikis in intelligence. Unpublished manuscript, 2011.
32. Typewith.me. http://www.youtube.com/watch?v=lOybsa4CXII (accessed 15 March 2011).
33. Andrus DC. The wiki and the blog: toward a complex adaptive intelligence community. *Stud Intell* 2005: 49(3).

34. Olson GM, Olson JS, Carter M, et al. Small group design meetings: an analysis of collaboration. *Hum Comput Interact* 1992; 7: 347–374.
35. Fuchs-Kittowski F and Köhler A. Wiki communities in the context of work processes. In: *International symposium on wikis and open collaboration*, San Diego, CA, October 16–18, 2005, New York, NY: ACM Press, pp. 33–39.
36. Arazy O, Gellatly I, Jang S, et al. Wiki deployment in corporate settings. *IEEE Technol Soc Mag* 2009; 28(2): 57–64.
37. Danis C and Singer DA. Wiki instance in the enterprise: opportunities, concerns and reality. In: *ACM conference on computer supported cooperative work (CSCW)*, 2008, pp. 495–504. San Diego, CA, November 8–11, 2008, New York, NY: ACM Press.
38. Hasan HM and Pfaff CC. Emergent conversational technologies that are democratising information systems in organisations: the case of the corporate wiki. In: *Proceedings of the international systems foundations*, Canberra, Australia, September 27–28, 2006, pp. 197–210.
39. Majchrzak A, Wagner C and Yates D. Corporate wiki users: results of a survey. In: *Proceedings of the international symposium on wikis*, Odense, Denmark, August 21–23, 2006, New York, NY: ACM Press, pp. 99–104.
40. White KF and Lutters WG. Midweight collaborative remembering: wikis in the workplace. In: *Symposium on computer human interaction for the management of information technology*, Cambridge, MA, March 30–31, 2007, New York, NY: ACM Press.
41. Phuwanartnurak AJ. Did you put it on the wiki? Information sharing through wikis in interdisciplinary design collaboration. In: *ACM international conference on design of communication (SIGDOC)*, Bloomington, IN, October 5–7, 2009, New York, NY: ACM Press, pp. 273–280.
42. Egli U and Sommerlad P. Experience report—wiki for law firms. In: *International symposium on wikis and open collaboration*, Orlando, FL, October 25–27, 2009, New York, NY: ACM Press, pp. 19–23.
43. Kasemvilas S and Olfman L. Design alternatives for a media wiki to support collaborative writing in higher education classes. *Issues Inf Sci Inf Technol* 2009; 6: 45–64.
44. Liccardi I, Davis HC and White S. CAWS: visualizing awareness to improve the effectiveness of co-authoring activities. *IEEE Distrib Syst Online* 2008; 9: 6–12.
45. Chi C, Zhou MX, Yang M, et al. Dandelion: supporting coordinated collaborative authoring in wikis. In: *ACM conference on human factors in computing systems (CHI)*, Atlanta, GA, April 10–15, 2010, New York, NY: ACM Press, pp. 1199–1202.
46. Holtzblatt LJ, Damianos LE and Weiss D. Factors impeding wiki use in the enterprise: a case study. In: *ACM conference on human factors in computing systems (CHI)*, Atlanta, GA, April 10–15, 2010, New York, NY: ACM Press, pp. 4661–4676.
47. Grudin J and Poole ES. Wikis at work: success factors and challenges for sustainability of enterprise wikis. In: *International symposium on wikis and open collaboration*, Gdańsk, Poland, July 7–9, 2010, New York, NY: ACM Press, pp. 1–5.
48. Springmeyer RR, Blattner MM and Max NL. A characterization of the scientific data analysis process. In: *IEEE visualization conference*, Boston, MA, October 19–23, 1992, Los Alamitos, CA: IEEE Computer Society Press, pp. 235–242.

Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw

Carsten Görg, *Member, IEEE* Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, and John Stasko, *Senior Member, IEEE*

Abstract—Investigators across many disciplines and organizations must sift through large collections of text documents to understand and piece together information. Whether they are fighting crime, curing diseases, deciding what car to buy, or researching a new field, inevitably investigators will encounter text documents. Taking a visual analytics approach, we integrate multiple text analysis algorithms with a suite of interactive visualizations in order to provide a flexible and powerful environment that allows analysts to explore collections of documents while sensemaking. Our particular focus is on the process of integrating automated analyses with interactive visualizations in a smooth and fluid manner. We illustrate this integration through two example scenarios: an academic researcher examining InfoVis and VAST conference papers and a consumer exploring car reviews while pondering a purchase decision. Finally, we provide lessons learned toward the design and implementation of visual analytics systems for document exploration and understanding.

Index Terms—Visual analytics, information visualization, sensemaking, exploratory search, information seeking, document analysis.

1 Introduction

Everyday, analysts and investigators confront large collections of data as they make decisions, solve problems, or simply seek to understand a situation better. Frequently, the data collections include text documents or documents with key text components. While numerical or structured data is more amenable to statistical and computational analysis, text data is conversely often messy and noisy, requiring a very sequential, slow processing (reading documents one-at-a-time, in order).

Investigators working with such document collections gather bits of information as they explore the data, hoping to form new insights about the issues at hand. Large, unstructured document collections make this task more difficult; the investigator may not know where to begin, what is important, or how concepts/events are related. The following situations are examples of these kinds of tasks:

- An academic researcher moves into a new area and seeks to understand the key ideas, topics, and trends of the area, as well as the set of top researchers, their interests, and collaborations.
- A consumer wants to buy a new car but encounters a large variety of possible models to choose from, each of which has ten to twenty "professional" reviews and a web forum with hundreds of postings.
- A family learns that their child may have a rare disease and scours the web for documents and information about the condition, easily encountering many articles.
- A police investigator has a collection of hundreds of case reports, evidence reports, and interview transcripts and seeks to "put the pieces together" to identify the culprits behind a crime.

Such processes, sometimes called Sensemaking [39,50,54], Information Seeking Support [44], or Exploratory Search [43, 66], go beyond the initial retrieval of data or the simple return of the "right" document. Instead, they involve analysts browsing, exploring, investigating, discovering, and learning about the topics, themes, concepts, and entities within the documents, as well as understanding connections and relationships among the entities.

One approach to this problem is the computational analysis of document text, including text mining [3, 22]. However, as many researchers have noted [37, 58], simply performing computational analysis of the documents may not be sufficient for adequate understanding of a document collection—the investigator inevitably will think of some question or perspective about the documents that is either not addressed by the computational analysis or not represented accurately enough to draw a conclusion.

Another approach leverages information visualization to show information about document contents [40,47,59]. However, interactive visualization itself may not be sufficient for sensemaking either—as the size of the document collection grows, interactively exploring the individual characteristics of each document may simply take too much time.

• *C. Görg is with the Computational Bioscience Program, University of Colorado, Aurora, CO, 80045; Z. Liu is with the Department of Computer Science, Stanford University, Stanford, CA, 94305; J. Kihm is with Cornell University, Ithaca, NY, 14850; J. Choo, H. Park, and J. Stasko are with the School of Interactive Computing & School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: Carsten.Goerg@ucdenver.edu, zcliu@cs.stanford.edu, jk2443@cornell.edu, {joyfull,hpark,stasko}@cc.gatech.edu*

Our approach to the problem combines these two analytics methods: (1) automated computational analysis of the text documents and (2) interactive visualization of the documents and of the analysis results. Such a combination is described as a *visual analytics* approach [36, 58], and it leverages the strengths of both the human and the computer. Humans excel at the interactive dialog and discourse of exploration and discovery. They develop new questions and hypotheses as more and more information is uncovered. They reason about the importance of new facts that are discovered. The computer excels at complex analyses to calculate metrics, correlations, connections, and statistics about the document collection. It can rapidly analyze large collections of documents in ways that would be prohibitively time-consuming for people to do.

Relatively few systems to date have deeply and smoothly incorporated both automated computational analysis and interactive visualization while providing a tight coupling between the two. Systems (as discussed in the related work) usually focus on one of the two approaches and provide a few elements from the other. For instance, computational analysis tools sometimes provide rudimentary visualizations to depict analysis results. Alternatively, interactive visualization systems may provide a few simple analysis techniques such as filtering or statistical analysis of the data.

Elaborating on this notion, Keim et al. [36] state:

> Visual analytics is more than just visualization. It can rather be seen as an integral approach to decision-making, combining visualization, human factors and data analysis. The challenge is to identify the best automated algorithm for the analysis task at hand, identify its limits which can not be further automated, and then develop a tightly integrated solution which adequately integrates the best automated analysis algorithms with appropriate visualization and interaction techniques.

In this article we explore this coupling through Jigsaw [55], a system for helping analysts explore document collections. Jigsaw is a relatively mature system, and has garnered trial use in the field by analysts in law enforcement, investigative reporting, fraud detection, and academic research, among other areas. An initial user study of the system showed its potential in helping investigators work with documents and in supporting different analysis strategies [34].

Earlier versions of Jigsaw emphasized multiple, coordinated visualizations but provided relatively little computational analysis of documents' text. The system primarily visualized connections between entities across documents to help investigators follow trails of information. More recently, we have added a variety of automated text analyses to the system including analysis of document similarity, document sentiment, document clusters by content, and document summarization through a few words or sentences. These new analyses aid investigators in determining the documents to examine first, the documents to focus on or discard, and the documents that may be related to different investigative angles.

Our focus is not on developing novel innovative algorithms for computational text analysis. Instead, we explore ways to smoothly integrate existing computational analyses into an interactive visual interface in a seamless manner that will provide a natural and fluid user experience. Furthermore, new computational analysis algorithms frequently are developed for well-defined tasks or problems with carefully constructed inputs and data. Real-world visual analytics systems, conversely, encounter messy, noisy data and must support open-ended analytical reasoning and sensemaking. Thus, our research also examines how computational analysis techniques can be used throughout visual exploration on challenging real-world data.

The contributions of this research include (1) methods for fluidly integrating computational text analysis and visualization, (2) illustration of the utility of such an approach through two example usage scenarios, and (3) lessons learned toward the design and construction of visual analytics systems for document exploration and understanding. Additionally, we provide implementation advice and experience on the integration of text analysis algorithms as a broader benefit for other researchers.

2 Related Work

Computationally-aided analysis and visualization of text and documents to assist human investigators with sensemaking has been a topic of intense research interest recently. Furthermore, different subdisciplines of computer science each bring their own focus to the problem. Thus, a comprehensive examination of related work likely would take a complete paper itself. Here, we highlight some of the existing research most strongly related to our work to provide the reader with greater context and familiarity of the varied approaches others have taken.

Systems in this area typically focus on some aspect of a document or document collection to present. Broadly, they visualize (1) metadata about the documents; (2) the document source text (words); (3) computed features and attributes of the documents including entities; and/or (4) general concepts, themes, and models across the documents. Visualization techniques have been developed for single documents or large collections of documents, though the techniques for individual documents often can be generalized to collections.

Systems with a specific focus on helping people understand various attributes of an academic paper collection are a good example of presenting **metadata about a set of documents**. PaperLens [40] employs a variety of bar chart, list, graph, and text-based visualizations to show author, topic, and citation data of past CHI papers. The system uses a clustering analysis to help group papers by topic as well. Selecting an author, paper, or concept in one visual representation loads related items into the other visualizations. A follow-on system, NetLens [33], focuses on visualizing content-actor data within document collections such as scientific publications and authors. NetLens uses bar charts, histograms, and lists to represent the data and help analysts understand statistics and trends from conference papers and their citations.

A number of innovative visualization techniques have been developed to represent the **words and source text of documents**. The SeeSoft system [21] represents a line of a text document by a row of pixels on the screen, with the length of the text line (number of characters) mapped to the length of the row of pixels. The goal of the technique is to visually depict

documents that are larger than what can normally be shown on one screen. Other well-known source text visualization techniques such as TextArc [47], Word Clouds [61], Word Trees [64], and Phrase Nets [59] actually still show text, unlike SeeSoft. They also show frequency and relationships of particular words or terms within documents.

Many systems, in fact, inhabit a conceptual space that transitions from visualizing document source to visualizing **computed metrics or features of a document or documents**. For example, Themail [60] analyzes collections of email messages using a variant of the term-frequency inverse document-frequency (TF-IDF) algorithm that focuses on each sender. The system's visualization is temporally-based and shows lists of keywords from the emails to characterize the main topics of messages during each month and over entire years.

Other techniques such as Arc Diagrams [63], DocuBurst [12], and Parallel Tag Clouds [13] compute metrics about a set of documents and visualize the computed metrics in unique ways. The PaperVis system [10] combines a relevance-determination algorithm with visualization to show relationships among academic articles. PaperVis performs citation and keyword analysis and presents the results through bullseye and radial space filling visualizations. The size of a node (document) and its distance to other documents denote its importance and relevance, respectively.

Keim and Oelke [38] perform numerous text analysis measures not seen in other document analysis systems including measures such as average word length, sentence length, number of syllables per word, and other measures such as Simpson's index and Hapax Legomena and Dislegomena (number of words occurring once and twice). The visualization of the analysis results for each of these measures uses a heatmap style display. Together with colleagues they subsequently added sentiment analysis to their measures [45] and added node-link network visualization to communicate relationships among the documents' sentiments [46].

One particular computed attribute sometimes visualized by systems is an entity within a document or documents. Identifying entities may be as simple as looking for particular strings or expressions within a document's text or it may involve complex computations to determine unique entities and their types. Different systems then choose to visualize the results of the computation in unique ways.

FeatureLens [19] uses text mining to identify words or expressions that recur across a set of documents. The system presents lists of the frequently occurring words and expressions, small overview rectangles representing each document with term positions identified by small marks, graphs of appearance count across documents, and textual views with terms highlighted. Primary users of the system may be literary scholars or journalists reviewing books or speeches. A follow-on system, POSVis [62], performs word-based part-of-speech analysis on documents and then displays the results using pixel-based overviews, word clouds, and network diagrams.

Entity Workspace [4] focuses on entity-based analysis and provides a "snap-together" visualization of entities and their attributes. Its analysis capabilities include spreading activation techniques to calculate degree-of-interest for the entities.

The IVEA system [57] uses entities of interest from a document collection to support faceted navigation and browsing across the collection. The system employs a matrix-style visualization with semantic zooming to represent the facets within documents.

Another set of systems move beyond the calculation of specific features, entities, or linguistic metrics of documents. These systems employ sophisticated text mining techniques to compute document **models and abstractions, often including concepts or themes across the documents**. Models and abstractions become especially useful as the size of the document collection grows.

The ThemeRiver technique [28] uses a river metaphor to represent temporal themes across a document collection. The river visualization extends from left-to-right to show the chronological progress of documents, and individual currents (colored bands) within the river represent different concepts. The vertical width of a current portrays its strength at a certain point in time.

Document topic modeling through latent Dirichlet allocation (LDA) [6] has become a popular technique for driving visualizations of document collections. TIARA [41] performs LDA analysis to identify themes throughout documents, and it portrays the results using a ThemeRiver-style visualization that has been augmented with word clouds. The system thus shows how topics grow and decline in focus over time. The system also supports user interaction to drill down and provide more detail on concept regions and to see the actual documents (emails) generating the concepts. TIARA can be used in many domains such as consumer reviews, email, and news. TextFlow [14] extends TIARA, showing how topics emerge and merge over time, how keywords relate to topics, and critical events within topics.

Parallel Topics [20] also employs LDA to model topics across a document collection and uses a ThemeRiver style visualization to present the results, coupled with a Topic Cloud to show important terms within topics, and a parallel coordinates visualization to show how individual documents contribute to the different topics. Other systems use LDA but provide different visualizations of the identified topics including word and topic lists [9], word clouds and sentences [23], force-directed networks [27], or custom-designed 2-D projections [11].

The FacetAtlas system [8] helps an analyst understand relationships between entities and facets within collections of documents sharing traits similar to academic articles. FacetAtlas uses a density map-style visualization with bundled edge connections between facets and entities along with rich interactive operations to present complex relationships between concepts in a document collection. Users can either search for specific concepts or interactively explore through the visualization interface.

The IN-SPIRE [26,30] system takes a different approach to visualizing document themes. It utilizes powerful automated analysis, clustering, and projection capabilities, primarily operating at the document level. IN-SPIRE computes high-dimensional similarities of documents and then visualizes these relationships through galaxy or themescape style pro-

jected representations that show the documents grouped into multiple clusters.

Finally, some visual analytics systems focus not on unique visualizations of text and documents but on creating environments where an analyst can analyze and reason about the documents. Often these systems use visual representations to help analysts explore the documents and develop hypotheses, and their target domain is frequently intelligence analysis. The systems' main goal typically is to give an investigator a faster, better understanding of a large collection of documents to help understand plots, themes, or stories across the collection.

nSPACE/TRIST/Sandbox [32, 67] provide sophisticated document analysis including entity identification and relations, trend analysis, clustering, and other automated operations. The systems present the documents through views of the documents' text or via groups of documents as small icons, but they augment this representation with sophisticated user interface flexibility for analysts to reason and develop stories about the data.

Commercial tools such as i2's Analyst Notebook [31] help intelligence, law enforcement, and fraud investigators work with document collections, among other types of data. Analyst's Notebook primary visualization is a node-link graph that shows connections between key entities in an investigation. Typically, however, the human investigator establishes these connections and constructs linkages.

As we will show in the following sections, *our contribution beyond this vast body of related work centers around the breadth of computational analysis techniques paired with a suite of rich interactive visualizations and integrating the two in a fluid, consistent manner*. Jigsaw provides multiple, varied perspectives to portray analysis results that allow the investigator to rapidly explore documents in a flexible manner. The particular emphasis on communicating entity connections across documents within concept-, temporal-, and sentiment-based perspectives also distinguishes it from existing systems.

3 Computational Text Analyses

An earlier version of Jigsaw, described in [55], focused on interactive visualization support rather than on computational modeling and analysis of documents' text. In an evaluation study [34] we found that the system was overall useful and supported a variety of strategies participants used to conduct their investigations on a document collection. However, we also found a number of situations in which the participants might have benefitted from additional information provided by computational text analysis, especially to get started with their investigation.

Some participants first read many of the documents to gain familiarity with the collection. Automated text summarization could have helped them to speed up the initial reading by reducing the amount of text to examine; document metrics, such as documents' date or length, could have provided order and structure to make the initial familiarization more efficient. Other participants focused early in their investigation on certain entities and tried to learn everything about them. Document similarity measures or features for recommending related documents could have supported this task by highlighting related information in other documents; showing documents clustered by content also could have helped them to step back and see the topics already examined or overlooked. Another group of study participants first randomly selected a few documents for acquiring evidence on which to start their investigation. Clustering documents by content could have been beneficial to help them to choose documents from different clusters for broader initial evidence.

We made similar observations on the potential benefit of computational analyses from our own use of Jigsaw, especially through our participation in the VAST Contest and Challenges [24, 42], as well as from other researchers' use of the system [49, 53]. In addition, we noticed that sentiment analysis would be another useful computational technique since product reviews are a natural document set for an investigation.

Computational text analyses are not without their own set of issues and concerns, however. As Chuang et al. note, text mining algorithms generate *models* of a document collection to be visualized, as opposed to source data about the documents [11]. When models are presented to the analyst, *interpretation* and *trust* arise as important concerns. In Jigsaw, we use an extensive suite of interactive visualizations to provide multiple perspectives on analysis results, thus enabling the analyst to review and explore the derived models and determine which are most helpful.

We now describe the suite of computational analyses added to the system and, most importantly, we focus on how the analyses integrate with different visualizations. First, we explain each analysis measure and how Jigsaw presents its results. Subsequently, we provide two example usage scenarios that illustrate how an analyst explores a document collection with the system (Section 4) and we present the implementation details of the analysis algorithms (Section 5). Our main focus has been on developing techniques for smoothly combining the computational analyses with interactive visualizations. We have emphasized an integrated user experience throughout, one that provides information where and when it is most helpful and that ideally feels natural and coherent to the analyst using it for an investigation.

Document Summarization

Jigsaw provides three different techniques to summarize a document or a set of documents: one sentence summaries, word clouds, and keyword summaries. A one sentence summary—a determination of the most significant sentence—of a single document helps analysts first to decide whether to read the full text of the document and subsequently to recall the content of a document read earlier. Jigsaw presents a one sentence summary above the full text of each document in its Document View (Figure 4). Additionally, the one sentence summary appears via tooltip wherever a document is presented through icon or name. Word clouds, the second type of document summary, help analysts to quickly understand themes and concepts within sets of documents by presenting the most frequent words across the selected documents. Jigsaw presents

word clouds of selected documents in its Document View and flexibly allows a fewer or greater number of words to be shown. The final type of summary, keyword summaries of document sets, labels sets of grouped documents in the Document Cluster View (Figure 5) and Document Grid View (Figure 6) in order to help an analyst know what the group is about. Keyword summaries are based on different metrics: word frequency in each set, word uniqueness across sets, or a combination of both. Summaries based on word frequency help to understand the content of each set, word summaries based on uniqueness help to analyze differences among sets. Jigsaw allows the analyst to interactively change the metric chosen. Overall, document summarization helps an analyst to quickly decide whether a document (or set of documents) is relevant for a specific task or question at hand and whether it should be investigated further.

Document Similarity

The similarity of two documents is measured in two different ways in Jigsaw: relative to the text within the documents or to the entities connected to the documents. The latter similarity measure is of particular interest for semi-structured document collections such as publications in which metadata-related entities (e.g., authors, years, and conferences) are not mentioned in the actual document text. Document similarity measures help an analyst to determine if a document is unique (an outlier in the collection) or if there exist related documents that should be examined as well. We implemented a new view in Jigsaw (the Document Grid View) to present, analyze, and compare document similarity measures. The view organizes the documents in a grid and provides an overview of all the documents' similarity to a selected document via the order and color of the documents in the grid representation. In all other views showing documents, an analyst can retrieve and display the five most similar documents to any document through a simple menu command.

Document Clustering

Clustering of similar and related documents also is based on either document text or on the entities connected to a document. Clusterings can be either computed fully automatically (using default values for the parameters of the clustering algorithm), or the analyst can specify the number of clusters and themes within clusters by selecting seed documents. Additionally, the analyst can interactively change clusters and define new clusters based on identified entities or keyword searches across the document collection. Document clustering partitions the documents into related groups to help an analyst explore the collection more systematically. Jigsaw presents clusterings in its Document Cluster View. The Document Grid View also provides an option to organize the documents by cluster when showing document metrics.

Document Sentiment Analysis and Other Metrics

Jigsaw computes a document's sentiment, subjectivity, and polarity, as well as other attributes such as a document's length and its number of connected entities. These metrics help an analyst seeking documents that are particularly high or low in key attributes. Jigsaw integrates and presents these metrics in its new Document Grid View. One metric can be used to determine the order of the documents within the grid, and a second metric (or the first metric again) can be mapped to the documents' color. The combined representation of any two of these metrics (by the documents' order and color) provides a flexible and powerful analytical view.

Identifying Entities in the Documents

The initial version of Jigsaw used a statistical entity identification approach from the GATE [15] package. We have added additional packages for automated entity identification and Jigsaw now provides three different approaches for automatically identifying entities of interest in text documents: (1) statistical entity identification, (2) rule-based entity identification, and (3) dictionary-based entity identification. It uses statistical approaches from GATE, Lingpipe,[1] the OpenCalais webservice,[2] and the Illinois Named Entity Tagger [52] to identify a variety of entity types, including person, place, organization, date, and money. For the rule-based approach we define regular expressions that match dates, phone numbers, zip codes, as well as email, web, and IP addresses. The dictionary-based approach allows analysts to provide dictionaries for domain-specific entity types that are identified in the documents using basic string matching.

The automatic identification of entities is still error-prone, especially in noisy, real-world data. Therefore, Jigsaw also provides functionality to correct errors in the set of identified entities. Within different visualizations, an analyst is able to add entities that were missed (false negatives), remove entities that were wrongly identified (false positives), change the type of entities, and define two or more entities as aliases.

Recommending Related Entities

To find embedded connections among entities (that might be connected via a long chain of other entities and documents) Jigsaw recommends related entities for further examination. The recommended entities are computed by searching for connecting paths between two or more entities in the document-entity network. The chain(s) of connected entities and documents are presented in the Graph View.

4 INVESTIGATIVE SCENARIOS

To better understand how these computational analysis techniques operate within Jigsaw and aid an investigation, we present two example use scenarios: a researcher exploring academic publications to learn about a research area and a consumer exploring product reviews to help make a purchase. The two scenarios involve relatively small document collections (in the hundreds) in order to make the presentation here more concise. We have used Jigsaw on larger collections numbering in the thousands of documents, however, and have found the new computational analysis capabilities to be even more useful at this larger scale. Because the static descriptions in this article cannot adequately convey the dynamic nature of the investigator's interaction with the system, we refer the reader to the accompanying videos for further illustration and elaboration of similar scenarios.

1. http://alias-i.com/lingpipe
2. http://www.opencalais.com

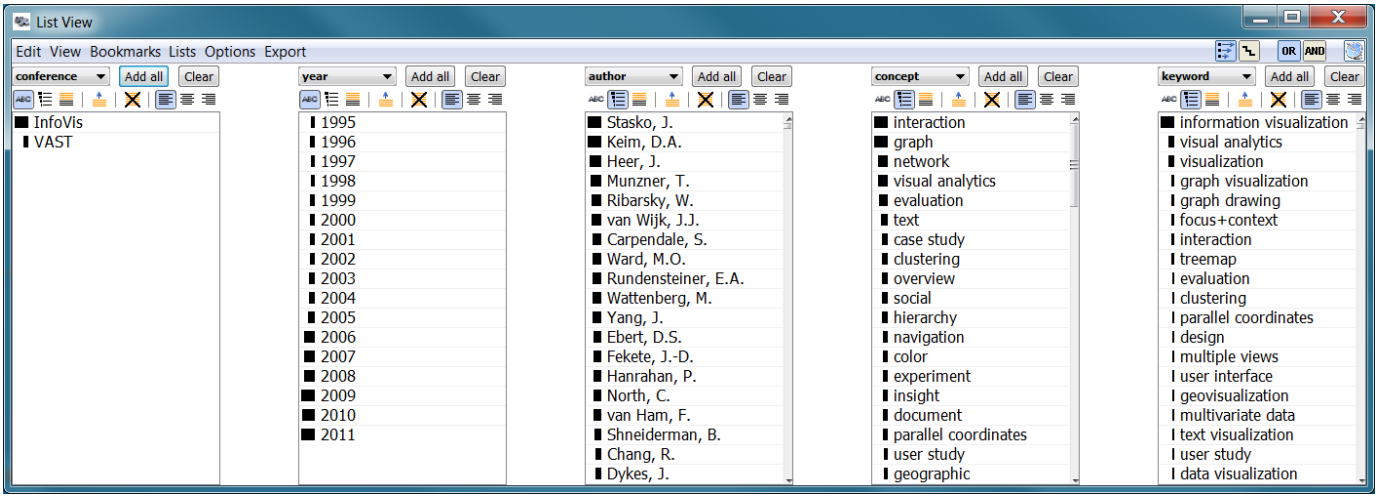


Fig. 1. List View showing conference, year, author, concept, and keyword, with the last three sorted by frequency.

4.1 Investigative Scenario: InfoVis and VAST Papers

In this scenario we illustrate an investigation of a dataset involving all of the IEEE InfoVis and VAST conference papers from 1995 to 2011; the InfoVis conference has run from 1995 to 2011 and VAST from 2006 to 2011. The dataset contains 578 documents, one for each paper presented at either of the conferences; each document includes the title and abstract of the article it represents; its entities are the paper's authors, index terms, keywords, conference, journal, and year. Additionally, we added an entity type "concept" including 80 domain-relevant terms such as *interaction*, *brushing*, *network*, and *evaluation* to be found within the articles' titles and abstracts.

To generate this dataset, we gathered information about the papers from the IEEE Digital Library. Throughout the data gathering process we performed a few cleaning steps and we resolved aliases for authors. We unified each unique author to one specific name because it was not uncommon to find initialized names or inconsistent inclusion of middle names. For keywords, we unified terms effectively meaning the same thing to one common string identifier. For example, the terms "Treemap", "tree-map", "treemaps", all were changed to the string "treemap". Jigsaw's List View (Figure 1) was very useful in this data cleaning phase as we could enumerate all the instances of any entity type in alphabetical order and easily check for similar strings. Additionally, we identified a set of documents to serve as seeds for clustering the documents.

Clearly, our domain knowledge helped in this initial data cleaning and entity resolution. Such transformations are typically necessary in any analysis of semi-structured text document information [2]. Jigsaw allows the results of such a process to be saved as an XML data file for sharing with others. In fact, we have made this conference paper dataset available on the web.[3]

For the purpose of this scenario, we introduce a hypothetical academic researcher, Bill, who works in the database area. Bill has developed a new technique for representing database schemata as graphs or networks and he has worked with a student to build a visualization of it. Bill knows a little about visualization research but not much detail about the IEEE InfoVis and VAST Conferences. He would like to learn whether one of these conferences would be a good fit for his paper, and if so, which one. Questions such as the following naturally arise in such an investigation:

- What are the key topics and themes of the two research areas?
- Have these topics changed over the history of the conferences?
- Who are the notable researchers in the different areas?
- Which researchers specialize in which topics?
- Are particular topics relating to his work present?
- Are there specific papers that are especially relevant?

Bill starts the investigation by examining statistics about the dataset to gain an overview of the conferences and areas. Jigsaw's Control Panel (not shown here) indicates that 1139 different researchers have contributed papers. These authors self-identified 1197 keywords and IEEE designated 1915 index terms for the papers. 78 of the 80 concepts (we generated) appeared in at least one title or abstract.

After gaining a general overview, Bill wants to learn more specifics about the key topics and authors so he opens Jigsaw's List View (Figure 1). He displays conference, year, author, concept, and keyword, then changes the list ordering from alphabetic to frequency-of-occurrence on the final three entity types to see the top-occurring entities. The small bar to the left of each entity denotes the number of documents in which it occurs. The general terms *information visualization* (101 occurrences), *visual analytics* (42), and *visualization* (40) are unsurprisingly the most frequent author-identified keywords. More interesting are the next most-common terms: *graph visualization* (18), *graph drawing* (17), *focus+context* (16), *interaction* (16), *treemap* (16), *evaluation* (14), *clustering* (13), and *parallel coordinates* (13). The term *interaction* (96) was the most frequent concept found in titles and abstracts, followed by *graph* (91), *network* (63), *visual analytics* (63), *evaluation* (55), and *text* (43). While these notions are likely familiar to someone within the field, they help a relative outsider such as Bill to understand some of the most important ideas in the research area.

3. http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html

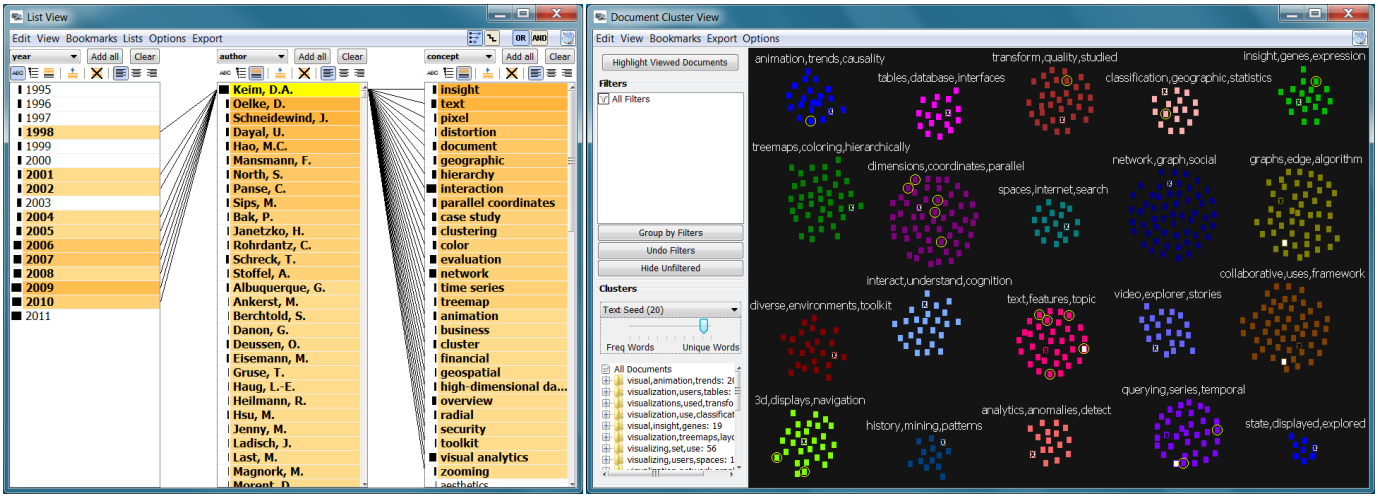


Fig. 2. List View (left) showing years, co-authors, and concepts connected to *Keim*. Document Cluster View (right) showing different clusters of related papers (small rectangles in different colors). Papers authored by Keim are selected (surrounded by a yellow circle).

Examining the author list, Bill notes that his old friend from database research, *Daniel Keim*, is one of the very top authors at the conferences. Bill is curious about Keim's papers at the conferences and decides to explore this further. He selects *Keim* in the List View and reorders the author and concept lists by strength of connection to that selection in order to see the entities most common with him (Figure 2, left). Connections in Jigsaw are defined by document co-occurrence, either of identified entities in the document text, such as concepts, or of meta entities of the document, such as authors. Connection strength is defined by the number of document co-occurrences: more co-occurrences signify stronger connection. (Further details of Jigsaw's connection model are described in [55].) The List View highlights entities connected to the selection via an orange background, with darker shades indicating stronger (more frequent) connections. Entities with white backgrounds are not directly connected. The terms *insight*, *text*, *pixel*, *distortion*, *document*, and *geographic* are the most connected concepts. Keim's most frequent co-authors are *Oelke*, *Schneidewind*, *Dayal*, *Hao* and *Mansmann*; he has published frequently from 1998 to 2010.

Bill now wants to explore ideas related to his own research. He notes that the concepts *graph* and *network* are the second and third most frequent, suggesting his work might be a good fit for these conferences. He selects the concept *graph* to learn which authors work on the topic. Jigsaw shows the most connected authors *van Ham*, *Abello*, *Hanrahan*, *Munzner*, and *Wong* and illustrates (dark shade of orange for recent years) that this has been a strong topic recently (Figure 3). Selecting *network* shows the most connected authors *Brandes*, *Ebert*, *Fekete*, *Hanrahan*, *Heer* and *Henry Riche* and that the topic also has been important recently. Surprisingly, the two author lists have many different names, which puzzles Bill since the two topics seem to be closely related.

To investigate further and gain a better understanding of the different topics within the conferences based on the articles' titles and abstracts, Bill switches to the Document Cluster View that displays each document in the collection as a small

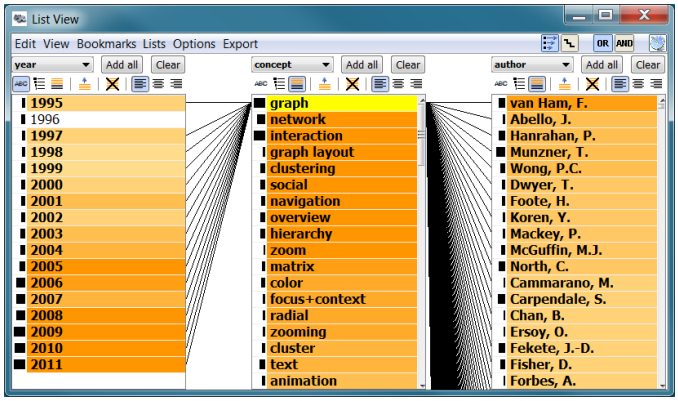


Fig. 3. List View with the concept *graph* selected, showing strongly connected years, concepts, and authors.

rectangle. Upon starting Jigsaw, Bill ran Jigsaw's automated computational analyses that calculated the similarities of all documents and a set of clusters based on these similarities.

The Document Cluster View (Figure 2, right) shows the 578 papers divided into 20 clusters resulting from the cluster analysis. The groups are each assigned a different color and are labeled with three descriptive keywords commonly occurring in the titles and abstracts in each cluster. If the summary terms are selected based solely on their frequency, common terms such as "data" and "visualization" represent many clusters which likely is not useful. The Cluster View provides a word frequency slider (left, lower center) for the investigator to interactively modify to show either more common or more unique terms affiliated with each cluster. Bill moves the frequency slider to the right, thus labeling clusters with terms more unique to that cluster. The resulting cluster labels represent important topics in these areas including toolkits, treemaps, text, animation, parallel coordinates, social networks, 3d, and databases (Figure 2, right).

Bill is curious which clusters his friend Daniel Keim's papers fall into. He applies cross-view selection and filtering [65], one key capability of Jigsaw. It can, for example, show the topics (clusters) in which an author publishes simply

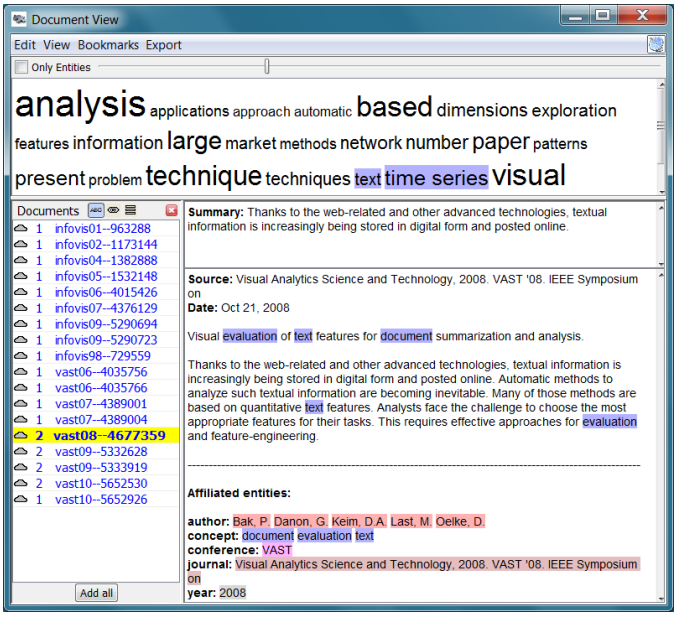


Fig. 4. Document View showing all the papers authored by *Keim*. Above the selected document's text (right) is a one sentence summary and below are the affiliated entities. The word cloud (top) summarizes all documents loaded in the view.

by selecting that author in any other view. Selecting *Keim* in the List View (Figure 2, left) immediately updates the Cluster View (Figure 2, right) and highlights (yellow circles around document rectangles) the papers *Keim* has authored. As shown in the figure, his work is relatively focused with five papers each in the "dimensions, coordinates, parallel" and "text, features, topic" clusters, and eight other papers scattered among six other clusters. Knowing Keim's research, Bill is quite surprised to see none of his papers in the cluster with "database" as a descriptive word. He decides to load all of Keim's papers into a Document View to examine them more closely.

The Document View (Figure 4) presents a list of documents (left) with the selected (yellow highlight) document's text and related information presented to the right. Below the text are the associated entities and above the text is the one sentence summary of the document computed by Jigsaw's summary analysis (described in Section 5.2). The word cloud at the top shows the most common words (with highlighted keywords and concepts) in the abstracts of these loaded papers. Bill reviewed all the papers quickly and noticed that indeed none were about database research. He grows a little concerned about whether these conferences would be a good fit for his paper.

Next, Bill wants to understand the evolution of topics in the conferences over time to learn which have waned and which have been growing in importance recently. To do so, he selects the first four years (1995 to 1998, all InfoVis) in the List View and notices strong connections to the "internet", "toolkit", and "3d" clusters in the Cluster View; additionally the List View shows strong connections to the concepts *interaction*, *case study*, *navigation*, and *animation*, with the concepts *network* and *graph* as the sixth and seventh most frequent. Selecting

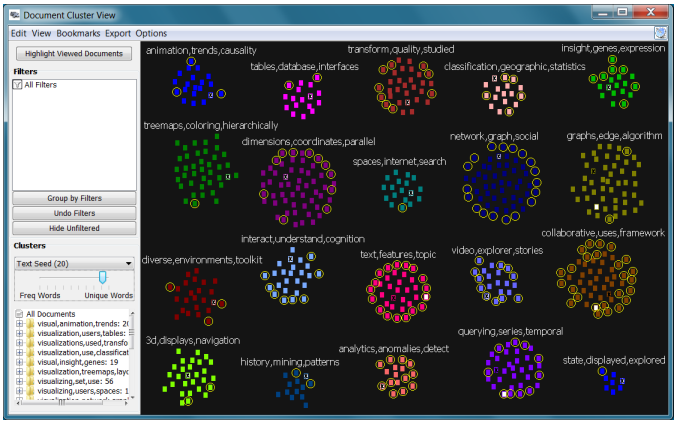


Fig. 5. Document Cluster View with the VAST Conference papers highlighted. Note the clusters where they provide a strong presence.

the most recent four years (2008 to 2011, both InfoVis and VAST) illuminates strong connections to multiple clusters but only connections to one document in the "3d" cluster and to two documents in the "internet" cluster. These topics clearly have waned over time. The terms *graph* and *network* are each in the top five connected concepts; thus, Bill sees how they have remained strong notions throughout the history of the conferences.

Bill next wants to better understand how the two conferences differ, so he explores the key concepts and ideas in each. He selects each conference, one at a time, in the List View and observes the connections. Among the ten most common concepts for each conference, five terms appear in both: *interaction*, *network*, *evaluation*, *graph*, and *case study*; the five other unique terms for InfoVis are *overview*, *hierarchy*, *color*, *navigation*, and *experiment*, and for VAST are *visual analytics*, *text*, *collaboration*, *clustering*, and *insight*. As shown in Figure 5, VAST papers (far fewer in number) occupied more than half of the "analytics, anomalies, detect", "video, explorer, stories", and "collaborative, uses, framework" clusters. These simple interactions help Bill begin to understand the subtle differences in the two conferences. His work still appears to fit well into either, however.

To learn more about the papers potentially related to his own work, Bill uses cross-view filtering in an opposite manner as he did earlier. He selects an entire cluster in the Document Cluster View and observes the resulting connections in the List View. For example, selecting the potentially related "network, graph, social" cluster shows that *Shneiderman*, *Fekete*, *Henry Riche*, *McGuffin*, *Perer*, and *van Wijk* are highly connected authors to its papers. Another potentially related cluster to Bill's work, "graphs, edge, algorithm" has top authors *Koren*, *Munzner*, *Abello*, *Ma*, and *van Ham*, all different than those in the previous cluster.

Bill decides to explore the papers in the "graphs, edge, algorithm" cluster. Since there are many, he moves his mouse pointer over the small rectangles in that cluster to quickly read a one sentence summary (tooltip) of each document. This document summary tooltip is available in other views such as the Document Grid View (Figure 6) and the Graph View where small iconic representations of documents are shown. None of

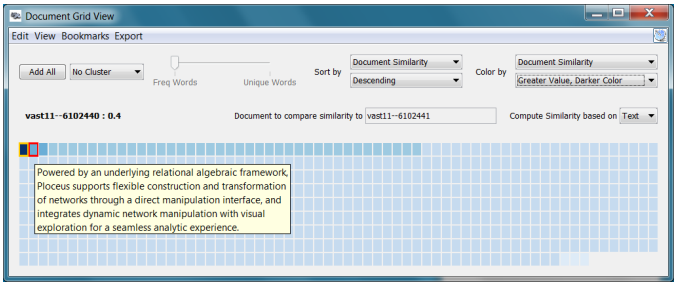


Fig. 6. Document Grid View with the document (small rectangle) order and shading set to correspond to the document's similarity to the selected Orion paper.

the papers in this cluster seem to be relevant to his research; they are not about the general representation of structure and relationships in networks but about specific details of layout techniques and their mathematical optimizations. Therefore, he moves on to the "network, graph, social" cluster. Here, he discovers a paper whose summary sparks his interest: "Despite numerous advances in algorithms and visualization techniques for understanding such social networks, the process of constructing network models and performing exploratory analysis remains difficult and time-consuming." Bill decides to load all the papers from this cluster into a Document View and selects this paper's icon in the Document Cluster View, thus also displaying it in the Document View. He reads the abstract of the VAST '11 paper by Heer and Perer about their Orion system and notices that it is definitely related to his work.

Bill now wants to know if papers similar to the Orion one have been published at the conferences. To find out, he uses Jigsaw's Document Grid View. The Document Grid View displays all the documents and is able to sort them by various text metrics, one being similarity to a base document. The Document Grid View in Figure 6 shows the similarity of papers compared to the Orion paper.

Bill decides to examine the most similar papers more closely, so he selects the eight most similar ones and displays them in a Document View (Figure 7). He observes that four of the eight papers are from InfoVis and four are from VAST. However, the paper most similar to the Orion paper is also from VAST '11 and is titled "Network-based visual analysis of tabular data." Upon reading the abstract, Bill learns that his work is quite similar to that done in this paper. Thus, he has both found some very relevant related work to explore further and he has determined that his new paper likely would fit in either conference, but VAST may be a slightly better match.

Through this abbreviated scenario, we illustrated how Jigsaw's analysis and visualization capabilities help analysts to gain quick insight on places to start an investigation, to learn about the key entities and topics in certain areas, and to explore connections and relationships in more depth. We also showed how it helps identify leaders, rapidly summarize sets of documents, compare and contrast information, find similarities and differences, and determine what should be investigated in more depth at a later point.

As shown in this scenario, investigative analyses of textual documents are often open ended and explorative in nature:

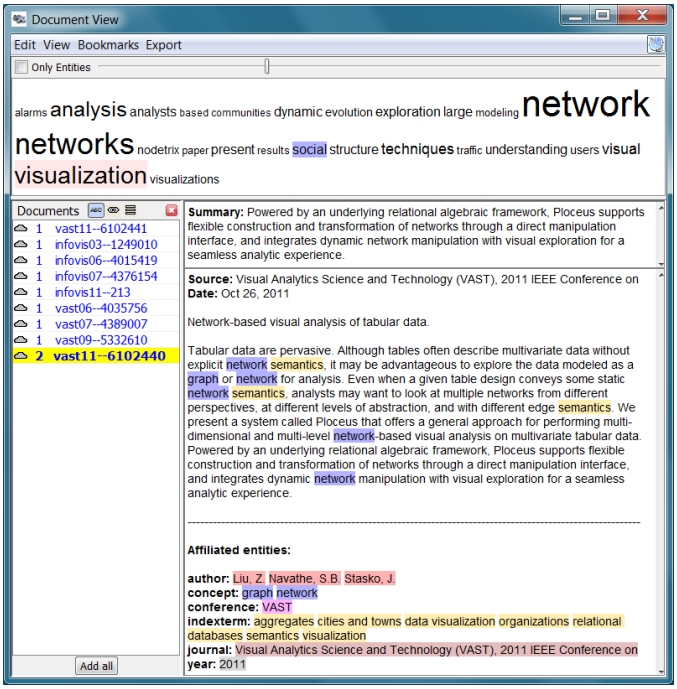


Fig. 7. Document View with the eight most similar papers to the Orion paper loaded. Selected here is the most similar document.

detailed questions or precise hypotheses may not be known at the beginning of an investigation but rather arise and evolve as the investigation unfolds. Analysts often switch back and forth between analyzing general trends, such as examining key topics, their relationships, and how they change over time, and more focused explorations about specific entities. Formulating new questions and finding supportive as well as contradictive evidence are fundamental tasks throughout these types of investigations.

4.2 Investigative Scenario: Car Reviews

The next scenario illustrates a different kind of investigation using documents—a consumer, Mary, who is shopping for a car. A colleague is selling his 2009 Hyundai Genesis, so to learn more about this particular model Mary examines a document collection consisting of 231 reviews of the car from the edmunds.com website. Mary wants to gain a general sense of consumers' views of the car and determine whether she should buy it. Specific concerns and goals that have arisen in her mind include:

- Identify and understand the important topics being discussed throughout the reviews,
- Learn the strong and weak points of the car,
- Determine whether perceptions of the car have improved or weakened over time,
- Identify the key competitive makes/models of cars,
- Judge whether particular attributes of the car such as its gas mileage, power, sound system, and reliability are good.

Mary could, of course, examine these 231 reviews one-by-one from the website just as anyone could do when exploring

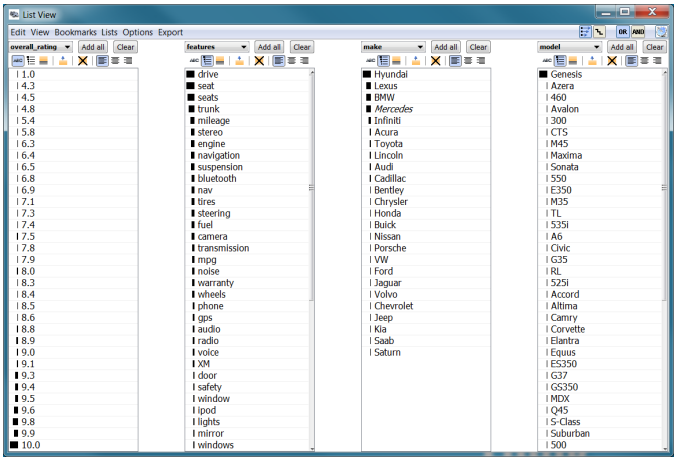


Fig. 8. List View showing the overall rating, feature, make, and model entity types and their values from the reviews. The last three are sorted by frequency.

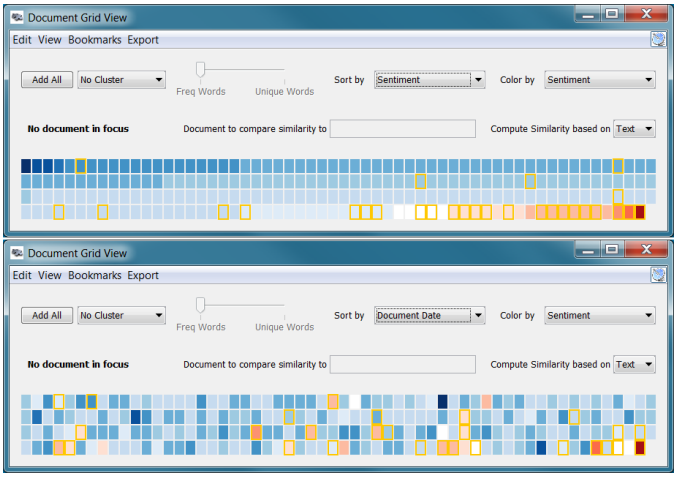


Fig. 9. Document Grid View showing all the reviews colored by sentiment: blue indicates positive, white neutral, and red negative. The top view displays the documents sorted by sentiment as well, while the bottom view shows them ordered by date ranging from the top-left (oldest) to bottom right (newest).

a collection of consumer reviews or webpages retrieved from a search engine. However, this process is tedious and may not illuminate well the key themes and connections across the reviews.

For illustrating Mary's use of Jigsaw in this scenario, we scraped reviews of the 2009 Genesis from the edmunds.com website and imported them into Jigsaw. Each review, including its title and main narrative written by a consumer, is represented as a document. The document's entities include various rating scores (e.g., exterior design, fuel economy, and reliability) that the review author explicitly designated. We also calculated an overall rating that is the average of all the individual ratings. We added three other entity types to be found within the document text (title and review narrative): car make (e.g., *Audi*, *Ford*, *Lexus*), car model (e.g., *525i*, *Avalon*, *ES350*), and car "feature", for which we defined 57 general terms about cars such as *seat*, *trunk*, *transmission*, and *engine*.

To get an overview of the reviews, Mary begins her investigation by invoking Jigsaw's List View (Figure 8). She displays the overall ratings from consumers, as well as the features, makes, and models discussed in the reviews, each sorted by frequency. Mary notices that the review ratings are generally high (indicated by longer frequency bars near the bottom of the first list); *drive*, *seat(s)*, *trunk*, and *mileage* are the most mentioned features; *Lexus*, *BMW*, *Mercedes*, and *Infiniti* are the most mentioned makes (excluding *Hyundai* itself); and *Azera*, *460*, *Avalon*, *300*, and *CTS* are the most mentioned models (excluding *Genesis* itself). This is useful information to know about the key competitive cars and most commented-upon features of the Genesis.

Although the ratings are generally good for the car, Mary wants to know more details about reviewers' thoughts. An analysis of the sentiment [26] of the reviews is useful here. To calculate sentiment, Jigsaw uses a dictionary-based approach, searching for positive or negative words throughout the document text. Here, Mary uses Jigsaw's capability to augment the dictionary by domain-specific words. For example, terms such as "quiet" and "sweet" are positive car sentiment words, while "lemon" and "clunk" indicate negative sentiment. Mary opens the Document Grid View and orders and colors the reviews by sentiment (Figure 9, top). Positive reviews are colored blue and shown first, neutral reviews are colored white and appear next, and negative reviews are colored red and shown last. Darker shades of blue and red indicate stronger positive and negative sentiment, respectively. At first glance, the reviews for the Genesis appear to be positive overall, roughly mirroring the overall rating scores shown in the List View.

Mary once had a car that developed a number of problems after a year of driving it, so she is curious what the most recent reviews of the car express. Thus, she changes the order of the reviews in the Document Grid View to be sorted by date, as shown in Figure 9, bottom. The oldest review from 06/26/2008 is placed in the top-left position in the grid and the most recent review from 07/24/2011 is in the bottom-right position. The view indicates that the earlier reviews were generally positive (shaded blue) but the more recent reviews begin to show more negative (red) perceptions. The most recent review is, in fact, the most negative, which is a concern. This trend might indicate that some issues with the car were not apparent when it first appeared but were revealed over time as the car matured.

To learn more about the car's potential weaknesses, Mary sorts the feature entities in the List View by their strength of connection to these negative reviews with overall rating below 8 (Figure 10). The terms *seat*, *tires*, *transmission*, *steering*, and *suspension* appear as the features most connected to the negative reviews, and Mary wants to investigate perceptions of these particular car features further.

For this task, document clustering by concept in Jigsaw is useful. Mary switches back to the Document Grid View and sorts the reviews into ten clusters where document similarity is calculated by Jigsaw based on the set of entities connected to each review. The clusters are labeled with descriptive keywords and the documents within each cluster are ordered and colored by their sentiment (Figure 11, left). The majority of the negative reviews aggregate into clusters 1 and 8 described

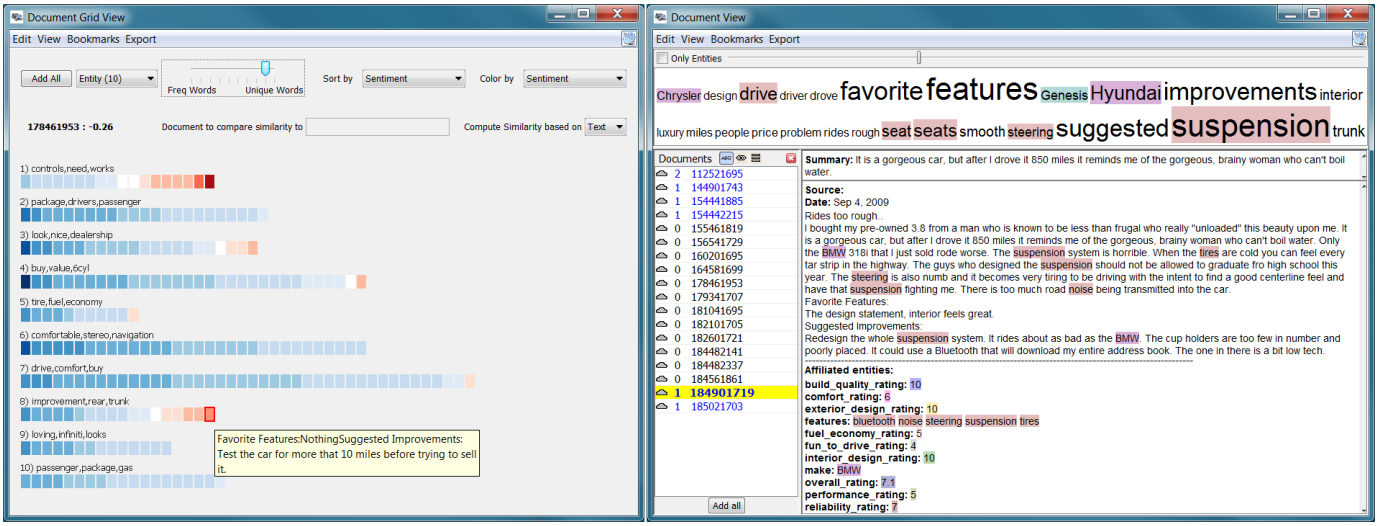


Fig. 11. Document Grid View (left) with the reviews grouped by similarity and ordered and colored by sentiment. Clusters 1 and 8 have the most negative sentiment. Document View (right) with the reviews from cluster 8 loaded. The word "suspension" is noteworthy within the word cloud at the top. The selected document illustrates an example of the views from reviews in this cluster.

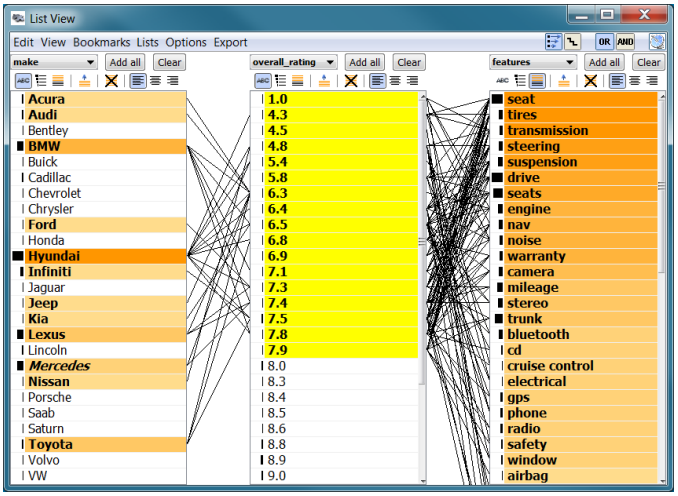


Fig. 10. List View showing the make, overall rating, and feature entity types. Low-rated reviews from 1.0 to 7.9 are selected and the feature list is sorted by connection strength to these selections.

by the terms "controls, needs, works" and "improvement, rear, trunk", respectively. It is not clear what each of these clusters is describing, so Mary loads the documents from each into a separate Document View to learn more.

The word clouds from each view highlight the most common words found in each review. The terms "suggested improvements" and "favorite features" are found in every review, so they are expectedly large. Similarly, the words "Hyundai" and "Genesis" also are common. However, the first cluster's word cloud also shows the word "transmission" in a large size, as does the second cloud for the word "suspension" (see Figure 11, right). This observation and the earlier similar finding from the List View suggests these may be key problems with the car. Mary decides to investigate further and reads all the reviews in cluster 8. She finds that the suspension is often described in a negative context, as shown in the review in Figure 11, right. She concludes that the suspension may indeed be a weak point of the 2009 Hyundai Genesis. Even though Jigsaw only performs document level sentiment analysis, Mary was able to also determine a type of feature-level sentiment analysis by combining the results of multiple computational analyses and coordinating their results across different visual representations of the document collection.

Mary now recalls that far more reviews were positive than negative, so she decides to examine the good aspects of the car. She selects all of the reviews giving the car a perfect overall rating of 10.0 in the List View (48 reviews in total, shown in Figure 12). The features *drive*, *seat(s)*, *stereo*, *fuel*, and *navigation* show up as being most connected. The terms *drive* and *seat(s)* occur in many documents overall as indicated by the long bar in front of the terms in the List View, so they may not be as useful. Mary now loads the documents mentioning *stereo*, the next highest term, into a Document View and reads these reviews. She learns that the Genesis' sound system is a 17-speaker Lexicon system and the reviewers typically rave about it, a definite plus to her.

Mary also wants to learn what are the other top, competitive brands of cars to consider as alternatives. She is curious about reviews mentioning other makes of cars. Thus, she sorts the car make entity by frequency in the List View and selects the top four other mentioned makes (all luxury cars), *Lexus*, *BMW*, *Mercedes*, and *Infiniti*, one by one. She notices that, overall, the connected reviews for each receive high ratings, suggesting that the Genesis is being compared favorably with these other makes. The reviews mentioning *BMW* exhibit slightly lower overall ratings, however. Perhaps prior BMW owners are not quite as favorably impressed as owners of the three other car brands. She reads the reviews also mentioning BMW and confirms that this is true.

To learn more about the ride quality of the car, an important feature to her, Mary displays Jigsaw's Word Tree View for "ride" (Figure 13). A Word Tree [64] shows all occurrences

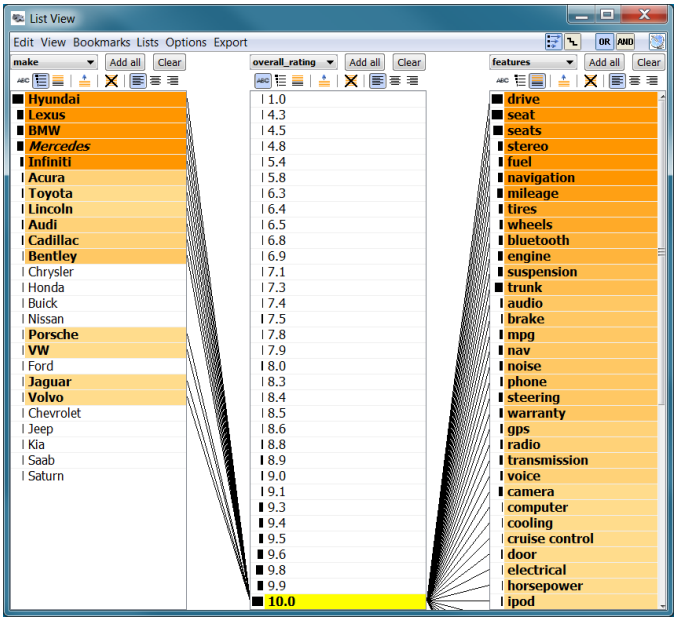


Fig. 12. List View showing the make, overall rating, and feature entity types. All the reviews with an overall rating of 10.0 are selected. The make list is sorted by overall frequency within the document collection and the feature list is sorted by connection strength to the 10.0 overall ratings.

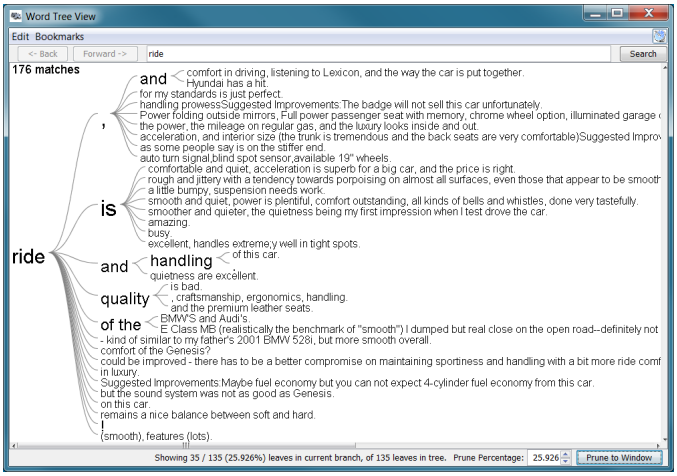


Fig. 13. Word Tree View showing occurrences of the word "ride" and the most common phrases that follow the word in sentences within the review collection.

of a word or phrase from the reviews in the context of the words that follow it, each of which can be explored further by a click. The Word Tree View shows that reviewers have different opinions about the quality of the ride, ranging from "a little bumpy" and "rough and jittery" to "comfortable and quiet" and "excellent".

Mary's investigations of the Genesis' reviews have helped her understand overall perceptions of the car and what the most recent impressions are. The computational analyses, the sentiment analysis and document clustering in particular, facilitated the identification of the car features perceived most favorably and unfavorably by the reviewers, and Mary learned more about other competitive makes and models of cars. An important part of such an exploration is reading the individual reviews of note, which we have not emphasized here for obvious reasons of brevity. However, we must stress that this activity is a key aspect of any document corpus investigation like this. The newly integrated computational analyses in Jigsaw help to more rapidly identify the documents of note for any of a variety of attributes or dimensions.

5 Computational Analysis Algorithms

In this section we provide a brief discussion of the text analysis algorithms we implement in Jigsaw, primarily for the reader interested in more detail. We integrate well-known algorithms for the different computational analyses, practical algorithms that can be readily implemented in Java (Jigsaw's implementation language) and that run in a "reasonable" time on computers that real clients would have. These descriptions and our experiences in designing and implementing the capabilities may be beneficial for other researchers who wish to integrate enhanced automated computational analysis in their visual analytics systems.

5.1 Preprocessing

To apply computational analyses, text documents are typically converted to a certain form of numerical vector representation. We use the standard "bag of words" encoding scheme where each dimension corresponds to a unique term, and the value represents the term count in the document. In Jigsaw, the vocabulary that constitutes the entire set of dimensions can be based on either all the terms occurring in the document corpus or only the entities that are identified within the documents. Thus, we obtain either a term-document or an entity-document matrix.

Then, we follow standard preprocessing procedures for text data such as stemming and stop word removal. For stemming, we use the Porter Stemmer [51] implementation in the Lingpipe library. Additionally, we exclude the terms and entities that appear less than three times throughout the entire document set. (The terms and entities are only excluded from the computational analyses; they are not removed from the dataset.) Based on empirical experiments we determined that these terms do not affect the results of the computational modules significantly while the vocabulary size is reduced drastically, often up to 40%, which improves both the computation time and memory usage.

After building the term-document matrix, we apply TF-IDF weighting and normalization [1]. TF-IDF weighting penalizes the terms that broadly appear in many documents since they would not contribute to the differentiation of one document from another. Normalization transforms each document vector to a unit norm to overcome the dependency on the document length.

Based on this numerical encoding of textual documents, we integrate three text analytical modules into Jigsaw: document summarization, document similarity, and document clustering. Document sentiment analysis, our fourth module, operates directly on the original document text.

5.2 Document Summarization

This module summarizes documents by extracting significant sentences. It first computes the importance scores (described below) for all the terms and for all the sentences within a single document, and then ranks the sentences with respect to the scores. The sentence with the highest importance score is determined to be the most representative sentence in the document and chosen as a summary sentence. The scored and ranked terms are used to summarize multiple documents with keywords (described in Section 5.4). To determine the sentences and the terms in a document we use a sentence splitter and a tokenizer from the Lingpipe library.

To implement the summarization algorithm, we apply the mutual reinforcement learning method [68]. This method first decomposes each document into a set of all the terms $T = \{t_1, \ldots, t_m\}$ and a set of all the sentences $S = \{s_1, \ldots, s_n\}$. A weighted bipartite graph between T and S is built with a weight matrix $W = \{w_{ij}\} \in \mathbb{R}^{m \times n}$ where w_{ij} is the frequency of the term t_i in the sentence s_j. Then we randomly initialize two vectors, $u \in \mathbb{R}^{m \times 1}$ and $v \in \mathbb{R}^{n \times 1}$, of the importance scores of terms and sentences, and perform a power iteration, i.e., $u = Wv$ and then $v = W^T u$, normalizing after every step. This iteration continuously passes the importance scores between terms and sentences until they converge.

5.3 Document Similarity

This module computes all the pairwise similarity scores for the documents in the corpus. The computation of similarity between two documents can be based on various measures. Although the most widely used measure is the Euclidean distance, semantically, cosine similarity can be a better choice for textual data [56] and therefore we use it in our implementation.

To obtain semantically better results, we do not compute the similarity based on the original document vector. Instead, we first reduce its dimension by applying the latent semantic analysis (LSA) technique [17] and then compute the similarity in the resulting reduced dimensional space. By grouping semantically similar terms, LSA improves similarity scores against polysemy and synonymy problems. After experimenting with different values, we chose to set the number of reduced dimensions to 20% of the number of dimensions after the preprocessing step of removing terms that occur in less than three documents. LSA requires the computation of the singular value decomposition (SVD) of the term-document matrix. We use the JAMA library[4] for matrix computations such as SVD.

Using the term-document or the entity-document matrix we can compute document similarity based on either the entire document text or on the entities identified in the documents.

5.4 Document Clustering

This module groups the documents into a given number of clusters, where similar documents fall into the same cluster. The similarity can be based on the document text or on the entities identified in the documents. We adopt the spherical k-means clustering algorithm [18], which uses cosine similarity as a distance measure.

The clustering algorithm requires the number of clusters as an input parameter. Theoretically, it is crucial to choose the "right" number of clusters to get optimal clustering results. Although there exist methods to quantitatively evaluate clusters [29], semantically it is difficult to determine the right number of clusters and achieve satisfactory results on noisy real world data. Thus, by default, we choose 20 as the number of clusters. Our reasoning behind this choice is that, on the one hand, if the document set has fewer clusters, then our result would show a few similar clusters that can be merged into a single true cluster by humans' further analysis. On the other hand, if the document set has significantly more clusters, e.g., 50 clusters, analysts might have difficulties in understanding their structure due to an unmanageable number of clusters, even if they represent the correct clustering result. However, we also provide a user option to specify the number of clusters in case an analyst is familiar with a specific document set and has some knowledge about its structure.

In addition, the algorithm requires a list of initial seed documents for the clusters as an input parameter. Although the algorithm is not sensitive to this parameter if the document set has a clearly clustered structure, we observed that the results can vary significantly depending on the initial seeds for most real-world document sets that do not have well-defined clusters. Thus, we carefully choose initial seed documents using a heuristic in which seed documents are recursively selected such that each seed document is the least similar document to the previously selected seed document. Due to space limitation we do not discuss details, such as optimizations and exceptions of this heuristic. We also provide a user option to choose initial seed documents in case an analyst is interested in specific topics in a document set and wants to steer the cluster analysis by choosing the seed documents according to the topics of interest.

To enhance the usability of the clustering results, we summarize the content of each cluster as a list of the most representative terms within its documents. We use the algorithm described in Section 5.2 after aggregating all documents in a cluster into one single document. However, instead of the most representative sentence, we use three high-ranking terms as the summary of the cluster. We compute a number of alternative term summaries for each cluster. One summary is based only on the term frequency within a cluster, whereas another summary also takes term uniqueness across clusters into account and eliminates any terms that would occur in multiple summaries; additionally we compute summaries that are gradually more strict on the uniqueness of summary terms (i.e., eliminating any terms that occur in 10%, or 20%, ..., or 90% of the cluster summaries). As described in Section 3 the analyst can interactively switch between the different cluster summaries to gain different perspectives on the clustering result, either examining the content of individual clusters or understanding differences among the clusters.

4. http://math.nist.gov/javanumerics/jama

5.5 Document Sentiment Analysis

This module provides two different implementations to characterize the text in a document on a positive-to-negative scale. It does not apply the preprocessing steps discussed in Section 5.1 but operates directly on the original document text.

One implementation is based on the classifier provided in the Lingpipe library. It applies the hierarchical classification technique described in Pang and Lee [48] and requires two classifiers: one for subjective/objective sentence classification and one for polarity classification. The technique involves running the subjectivity classifier on the document text to extract the subjective sentences first and then running the polarity classifier on the result to classify the document as positive or negative. We trained the subjectivity classifier with the data provided in [48]. To train the polarity classifier we used 2,000 product reviews (1,000 positive and 1,000 negative) extracted from amazon.com. We considered all reviews with a rating of 4 or 5 as positive and those with a rating of 1 or 2 as negative. We did not use reviews with a rating of 3.

An alternative implementation computes a quantitative sentiment score for each document on a scale from +1 (positive) to -1 (negative) via a dictionary-based approach, identifying "positive" and "negative" words in documents. We developed the list of words by creating two initial sets of negative and positive words and then iterating and checking results against known positive and negative documents. The results have been surprisingly good, particularly when characterizing documents strong in expected sentiment such as product reviews. We also allow the user to provide domain-specific dictionaries of positive and negative words to classify the documents. This feature was very useful for a collaborative analysis in which we examined wine reviews with a wine expert; we developed dictionaries of words that describe "good" and "bad" wines to classify the reviews.

5.6 Computation Time

The runtime of the computational analyses depends on the characteristics of the document collection (number of documents, average document length, and number of entities per document) and the available computational power (processor speed and memory size). The InfoVis and VAST papers dataset in our case study has 578 documents, the average length of a paper's title and abstract is 1,104 characters (min: 159; max: 2,650), and the average number of entities per paper is 17 (min: 4; max: 58). On a desktop computer with 8 GB of memory and two 2.4 GHz Quad-Core processors, computing the summary sentences and the sentiment analysis each took 2 seconds, the text-based similarity computation took 47 seconds, the entity-based similarity computation took 10 seconds, the text-based cluster computation took 20 seconds, and the entity-based cluster computation took 15 seconds, resulting in a total computation time of less than 2 minutes. The car reviews dataset is much smaller (231 documents) and all analyses finished in 12 seconds.

We also ran the analyses on other datasets with different characteristics. From a practical point of view, the computation time can be divided into three categories. For small datasets (about 500 documents) the computation time is a few minutes (coffee break), for medium-sized datasets (about 1,500 documents) the computation time is less than one hour (lunch break), and for larger datasets (about 5,000 documents) the computation time is several hours (over-night).

6 Discussion

Investigations on document collections proceed with the analyst gathering nuggets of information while forming new insights and deeper understanding of the document contents. Especially when the documents are unfamiliar, an investigator may not know where to start, what is related, or how to dive more deeply into analysis. We believe that fluid integration of computational analyses with interactive visualization provides a flexible environment that scaffolds the investigator's exploratory process.

In exploration and sensemaking, investigators likely want to ask a broad set of questions and also develop new questions throughout the investigation process. Interactive visualization supports this dynamic conversation or dialog between the investigator and the data, and it makes the results of powerful computational analyses more easily accessible and contextually relevant. Our efforts to integrate enhanced computational analysis support into Jigsaw have taught us a number of lessons about this process (resulting both from an implementation perspective and from working with users of the system [7, 35]), but five in particular stand out:

1. Make different computational analysis results available throughout the system in a variety of different contexts and views, not in just one canonical representation. Chuang et al. [11] identify *interpretation* and *trust* as two key issues to the success of visual analytics systems. With respect to the results of computational text mining, trust seems to be a primary concern. We have found that portraying the results of mining algorithms under different perspectives better allows the analyst to inspect and interpret the algorithm's results. In particular, multiple analyses within Jigsaw appear in several different views and can be examined under different perspectives. For example, the single sentence document summaries are shown above the corresponding full document text in the Document View as one might expect, but they also are available as tooltips anywhere a document is represented iconically or by name. Clusterings are shown (naturally) in the Document Cluster View but also in the Document Grid View that simultaneously can show similarity, sentiment, and summary analysis results. Furthermore, clusters are easy to select and thus inspect the member documents under other analysis perspectives and views. Given any set of documents resulting from a text analysis, one simple command allows those documents to be loaded into a Document View for further manual exploration.

2. Flexibly allow analysis output also to be used as input. Investigators using Jigsaw can select individual documents from any analysis view and can then request to see that document's text or see related documents. Jigsaw presents the results of similarity, clustering, and sentiment analyses visually

(output), but such results can be clicked on or selected by the analyst (input) to drive further exploration. This capability is pervasive throughout the system—any document or entity can be acted upon to drive further investigation. We believe this design, which helps to facilitate the core, iterative sensemaking cycle of visual analytics [36], enables smoother, more flexible interaction with the system, ultimately leading to deeper inquiry, exploration, and increased knowledge.

3. Integrate different, independent computational analysis measures through interactive visualization in order to extend functionality and power. A deep integration of automated analysis with interactive visualization results in capabilities beyond each of the two components ("the whole is greater than the sum of the parts"). For example, Jigsaw provides document-level sentiment analysis, but does not analytically provide sentiment with respect to specific terms, concepts, or features within a document. However, as illustrated in the car review scenario, by first performing content-based clustering that divides the car reviews into sets of documents discussing different car features, and then visualizing the sentiment on the resulting clusters, one achieves a type of feature-based sentiment. The scenario showed how the reviewers felt negatively about the car's suspension and transmission.

4. Provide computational support for both analysis directions: narrowing down as well as widening the scope of an investigation. Many investigations take the form of an hourglass: an analyst first confronts a large amount of data (top of the hourglass), iteratively filters and searches the data to discover a small number of interesting leads (middle of the hourglass), and then expands the data under investigation again by following connections from those identified leads (bottom of the hourglass). These new data points then represent the top of another hourglass and the analyst repeats the process. Cutting et al. [16] describe this narrowing-widening, iterative process as the Scatter/Gather method. To smoothly move through the different stages of an hourglass investigation, a visual analytics system should provide support for narrowing down as well as widening the scope of analysis. Jigsaw provides a variety of analysis support for both tasks. Document clustering and sentiment analysis help narrow down the scope by limiting it to one (or a few) clusters or taking only positive or negative documents into account; document similarity and recommending related entities help widen the scope by suggesting additional relevant documents; and identified entities help with both directions: they can be used to determine a germane subset of a document set (containing one or more identified entities) or to suggest other related documents (containing the entity of interest).

5. Expose algorithm parameters in an interactive user-accessible way. The effectiveness of many computational analyses depends on the choices of their parameters. Whenever possible, visual analytics tools should provide users intuitive access to the parameter space of the underlying analyses. In Jigsaw, we expose parameters in a number of different ways. For the k-means clustering, we expose the corresponding parameters directly since they are quite intuitive. Users can either choose default values or define the number of clusters, specify whether the clustering should be based on the document text or only the entities connected to a document, and provide initial seed documents for the clusters. They then can display different clusterings from different parameter choices in multiple Document Cluster Views to compare and contrast them. We take a different approach for the cluster summarization algorithm. Instead of exposing the summarization parameters directly, we precompute a set of summarizations and let users explore the parameter space by selecting cluster summaries via an interactive slider (based on uniqueness vs. frequency of the summary words). Users have preferred this approach more than exposing the (not so intuitive) parameters of the summarization algorithm directly. For the dictionary-based entity identification and the sentiment analysis, users can provide their own domain-specific dictionaries. This flexibility has proven to be very useful in various domain-specific investigations that we and others have conducted with Jigsaw (e.g., investigating wine reviews, car reviews, scientific papers, and Java code). User requests for exposing additional algorithm parameters, such as regular expressions for the rule-based entity identification approach, confirm the importance of this lesson.

7 Conclusion

Helping investigators to explore a document collection is more than just retrieving the "right" set of documents. In fact, all the documents retrieved or examined may be important, and so the challenge becomes how to give the analyst fast and yet deep understanding of the contents of those documents.

In this article we have illustrated methods for integrating automated computational analysis with interactive visualization for text- and document-based investigative analysis. We implemented a suite of analysis operations into the Jigsaw system, demonstrating how to combine analysis results with interactive visualizations to provide a fluid, powerful exploration environment. Further, we provided two example sensemaking scenarios that show both the methodologies and the utility of these new capabilities. We included brief descriptions of the computational analysis algorithms we chose to help readers seeking to implement similar operations in their systems. Finally, we described our experiences in building the new system and the lessons we learned in doing so.

The contributions of the work are thus:

- Techniques for integrating computational analysis capabilities fluidly with different interactive visualizations, and realization of those techniques in the Jigsaw system.
- Illustrations of the benefits of this approach via two example sensemaking scenarios. These scenarios provide sample questions and tasks, methods to resolve them, and the analysis and insights that result.
- Guidance for HCI/visualization researchers about the implementation of practical, text-focused computational analysis algorithms.
- Design principles for the construction of future document analysis and visual analytics systems.

A particular strength of Jigsaw is its generality for analyses on different types of documents. Many other systems have been tailored to a specific style of document or content domain and thus provide sophisticated capabilities only in that area. Jigsaw has been applied in the domains here (academic research and consumer product reviews) and in other diverse areas such as aviation documents [49], understanding source code files for software analysis and engineering [53], genomics research based on PubMed articles [25], and investigations in fraud, law enforcement, and intelligence analysis [35]. The system is available for download.[5]

Many avenues remain for future research. We admittedly have not conducted formal evaluations or user studies of these new capabilities within Jigsaw. Determining the best methods to evaluate systems like this is a research challenge unto itself. Our earlier user study involving Jigsaw [34] identified the potential benefits of the system, so we believe that the addition of the new computational analysis capabilities will provide even further value. In particular, the new capabilities address analysis needs identified in the user study and determined through earlier trial use of the system by clients.

We also plan to explore newer, more powerful methods and algorithms for calculating analysis metrics. The areas of computational linguistics, dimensionality reduction, and text mining are ripe with analysis methods such as topic modeling [6] and multi-word expressions [5] that could be integrated into Jigsaw. Furthermore, allowing user-driven interactive feedback to modify and evolve the computational analyses would provide an even more flexible exploration environment.

Finally, we made a claim that to achieve its fullest potential within visual analytics, a system must deeply and seamlessly combine automated computational analysis with interactive visualization. Actually, according to the definition of visual analytics introduced in *Illuminating the Path* [58], we omitted the third key piece of the equation: integrated support for analytical reasoning. Systems such as Jigsaw seeking to provide comprehensive analytic value also should include facilities for supporting human investigators' analytic reasoning processes and goals.

We are encouraged that the vision of visual analytics is beginning to be realized. The system and experiences described in this paper illustrate the potential of such an approach: fluidly integrating computational data analysis algorithms with flexible, interactive visualizations provide investigators with powerful data exploration capabilities and systems.

ACKNOWLEDGMENTS

This research is based upon work supported in part by the National Science Foundation via Awards IIS-0915788 and CCF-0808863, and by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

5. http://www.cc.gatech.edu/gvu/ii/jigsaw

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*. ACM Press, 2011.

[2] C. Bartneck and J. Hu, "Scientometric analysis of the CHI proceedings," in *ACM CHI*, 2009, pp. 699–708.

[3] M. Berry and M. Castellanos, *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer, 2008, vol. XVI.

[4] E. A. Bier, S. K. Card, and J. W. Bodnar, "Principles and tools for collaborative entity-based intelligence analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 178–191, 2010.

[5] D. Blei and J. Lafferty, "Visualizing topics with multi-word expressions," arXiv:0907.1013v1, Tech. Rep., 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[7] E. Braunstein, C. Görg, Z. Liu, and J. Stasko, "Jigsaw to Save Vastopolis - VAST 2011 Mini Challenge 3 Award: "Good Use of the Analytic Process"," in *IEEE VAST*, Oct 2011, pp. 323–324.

[8] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted visualization for rich text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1172–1181, 2010.

[9] A. J. B. Chaney and D. M. Blei, "Visualizing topic models," in *AAAI ICWSM*, 2012, pp. 419–422.

[10] J.-K. Chou and C.-K. Yang, "PaperVis: Literature review made easy," *Computer Graphics Forum*, vol. 30, no. 3, pp. 721–730, 2011.

[11] J. Chuang, D. Ramage, C. D. Manning, and J. Heer, "Interpretation and trust: designing model-driven visualizations for text analysis," in *ACM CHI*, 2012, pp. 443–452.

[12] C. Collins, S. Carpendale, and G. Penn, "DocuBurst: Visualizing document content using language structure," *Computer Graphics Forum*, vol. 28, no. 3, pp. 1039–1046, 2008.

[13] C. Collins, F. B. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *IEEE VAST*, Oct. 2009, pp. 91–98.

[14] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, X. Tong, and H. Qu, "TextFlow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.

[15] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*, 2011.

[16] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: a cluster-based approach to browsing large document collections," in *ACM SIGIR*, 1992, pp. 318–329.

[17] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. of the Society for information Science*, vol. 41, pp. 391–407, 1990.

[18] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1/2, pp. 143–175, 2001.

[19] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering interesting usage patterns in text collections: integrating text mining with visualization," in *ACM CIKM*, 2007, pp. 213–222.

[20] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Parallel topics: A probabilistic approach to exploring document collections," in *IEEE VAST*, Oct 2011, pp. 229–238.

[21] S. G. Eick, "Graphically displaying text," *Journal of Computational and Graphical Statistics*, vol. 3, no. 2, pp. 127–142, 1994.

[22] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[23] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi, "The topic browser: An interactive tool for browsing topic models," in *NIPS Workshop on Challenges of Data Visualization*, 2010.

[24] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Jigsaw meets Blue Iguanodon – The VAST 2007 Contest," in *IEEE VAST*, Oct. 2007, pp. 235–236.

[25] C. Görg, H. Tipney, K. Verspoor, W. Baumgartner, K. Cohen, J. Stasko, and L. Hunter, "Visualization and language processing for supporting analysis across the biomedical literature," in *Knowledge-Based and Intelligent Information and Engineering Systems*. LNCS Springer, 2010, vol. 6279, pp. 420–429.

[26] M. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner, "User-directed sentiment analysis: Visualizing the affective content of documents," in *Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 23–30.
[27] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. U. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 2, pp. 23:1–23:26, 2012.
[28] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: Visualizing theme changes over time," in *IEEE InfoVis*, Oct 2000, pp. 115–123.
[29] J. He, A.-H. Tan, C. L. Tan, and S. Y. Sung, "On quantitative evaluation of clustering systems," in *Clustering and Information Retrieval*, 2003, pp. 105–134.
[30] E. Hetzler and A. Turner, "Analysis experiences using information visualization," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp. 22–26, 2004.
[31] "i2 - Analyst's Notebook," http://www.i2inc.com/.
[32] D. Jonker, W. Wright, D. Schroh, P. Proulx, and B. Cort, "Information triage with TRIST," in *International Conference on Intelligence Analysis*, May 2005.
[33] H. Kang, C. Plaisant, B. Lee, and B. B. Bederson, "NetLens: iterative exploration of content-actor network data," *Information Visualization*, vol. 6, no. 1, pp. 18–31, 2007.
[34] Y.-a. Kang, C. Görg, and J. Stasko, "How can visual analytics assist investigative analysis? Design implications from an evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 5, pp. 570–583, May 2011.
[35] Y.-a. Kang and J. Stasko, "Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2869–2878, 2012.
[36] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," *Information Visualization: Human-Centered Issues and Perspectives*, pp. 154–175, 2008.
[37] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, Eds., *Mastering the Information Age – Solving Problems with Visual Analytics*. Eurographics Assoication, 2010.
[38] D. A. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *IEEE VAST*, 2007, pp. 115–122.
[39] G. Klein, B. Moon, and R. Hoffman, "Making sense of sensemaking 1: alternative perspectives," *IEEE Intelligent Systems*, vol. 21, pp. 70–73, 2006.
[40] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson, "Understanding research trends in conferences using PaperLens," in *ACM CHI: Extended Abstracts*, 2005, pp. 1969–1972.
[41] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 2, pp. 25:1–25:28, Feb. 2012.
[42] Z. Liu, C. Görg, J. Kihm, H. Lee, J. Choo, H. Park, and J. Stasko, "Data ingestion and evidence marshalling in Jigsaw," in *IEEE VAST*, Oct. 2010, pp. 271–272.
[43] G. Marchionini, "Exploratory search: From finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006.
[44] G. Marchionini and R. W. White, "Information-seeking support systems," *IEEE Computer*, vol. 42, no. 3, pp. 30–32, Mar. 2009.
[45] D. Oelke, P. Bak, D. Keim, M. Last, and G. Danon, "Visual evaluation of text features for document summarization and analysis," in *IEEE VAST*, Oct. 2008, pp. 75–82.
[46] D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L.-E. Haug, and H. Janetzko, "Visual opinion analysis of customer feedback data," in *IEEE VAST*, Oct. 2009, pp. 187–194.
[47] W. B. Paley, "TextArc: Showing word frequency and distribution in text," in *IEEE INFOVIS (Poster)*, 2002.
[48] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of Association for Computational Linguistics*, 2004, pp. 271–278.
[49] O. J. Pinon, D. N. Mavris, and E. Garcia, "Harmonizing European and American aviation modernization efforts through visual analytics," *Journal of Aircraft*, vol. 48, pp. 1482–1494, Sept-Oct 2011.
[50] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *International Conference on Intelligence Analysis*, May 2005.
[51] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
[52] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *CoNLL*, 2009, pp. 147–155.
[53] H. Ruan, C. Anslow, S. Marshall, and J. Noble, "Exploring the inventor's paradox: applying Jigsaw to software visualization," in *ACM SOFTVIS*, Oct 2010, pp. 83–92.
[54] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in *ACM CHI*, 1993, pp. 269–276.
[55] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, no. 2, pp. 118–132, 2008.
[56] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on Artificial Intelligence for Web Search (AAAI)*, 2000, pp. 58–64.
[57] V. Thai, P.-Y. Rouille, and S. Handschuh, "Visual abstraction and ordering in faceted browsing of text collections," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 2, pp. 21:1–21:24, Feb. 2012.
[58] J. J. Thomas and K. A. Cook, *Illuminating the Path*. IEEE Computer Society, 2005.
[59] F. van Ham, M. Wattenberg, and F. B. Viégas, "Mapping text with phrase nets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1169–1176, 2009.
[60] F. B. Viégas, S. Golder, and J. Donath, "Visualizing email content: portraying relationships from conversational histories," in *ACM CHI*, 2006, pp. 979–988.
[61] F. B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137–1144, 2009.
[62] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, "What's being said near "Martha"? Exploring name entities in literary text collections," in *IEEE VAST*, Oct. 2009, pp. 107–114.
[63] M. Wattenberg, "Arc diagrams: Visualizing structure in strings," in *IEEE INFOVIS*, 2002, pp. 110–116.
[64] M. Wattenberg and F. B. Viégas, "The Word Tree, an Interactive Visual Concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1221–1228, 2008.
[65] C. Weaver, "Cross-filtered views for multidimensional visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 192–204, March 2010.
[66] R. W. White, B. Kules, S. M. Drucker, and M. C. Schraefel, "Supporting exploratory search," *Communications of the ACM*, vol. 49, no. 4, pp. 36–39, Apr. 2006.
[67] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort, "The Sandbox for analysis: Concepts and methods," in *ACM CHI*, April 2006, pp. 801–810.
[68] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *ACM SIGIR*, 2002, pp. 113–120.

Carsten Görg is an Instructor in the Computational Bioscience Program in the University of Colorado Medical School. He received a PhD in computer science from Saarland University, Germany in 2005. His research interests include visual analytics and information visualization with a focus on designing, developing, and evaluating visual analytics tools to support the analysis of biological and biomedical datasets. Dr. Görg is a member of the IEEE Computer Society.

Zhicheng Liu received the BS degree in computer science from the National University of Singapore in 2006 and the PhD degree in human-centered computing from the Georgia Institute of Technology in 2012. He is currently a postdoctoral scholar at Stanford University. His current research interests include visualizing big data and developing novel interaction mechanisms in visual analysis.

Jaeyeon Kihm is a PhD student in Information Science at Cornell University. He received the MS degree from the Georgia Institute of Technology in 2011 and the BS degree from the Illinois Institute of Technology in 2009. He is currently developing an energy-efficient user interface system for mobile information appliances.

Jaegul Choo received the BS degree in electrical engineering from Seoul National University, Seoul, Korea in 2001 and the MS degree in electrical engineering from the Georgia Institute of Technology in 2009, where he is currently a research scientist as well as a PhD candidate in computational science and engineering. His research interests include visualization, data mining, and machine learning with a particular focus on dimension reduction and clustering methods.

Haesun Park is currently a Professor in the School of Computational Science and Engineering and director of the NSF/DHS FODAVA-Lead (Foundations of Data and Visual Analytics) Center at the Georgia Institute of Technology. Dr. Park has published over 150 peer reviewed papers in the areas of numerical algorithms, data analysis, visual analytics, bioinformatics, and parallel computing. She has served on numerous editorial boards including IEEE Transactions on Pattern Analysis and Machine Intelligence, SIAM Journal on Matrix Analysis and Applications, SIAM Journal on Scientific Computing, and has served as a conference co-chair for the SIAM International Conference on Data Mining in 2008 and 2009.

John Stasko is a Professor and Associate Chair of the School of Interactive Computing at the Georgia Institute of Technology. He received a PhD in computer science from Brown University in 1989. His research interests are in human-computer interaction with a specific focus on information visualization and visual analytics. Dr. Stasko is a member of the IEEE and is on the Steering Committee for the IEEE Information Visualization Conference.

Info Vis

Article

Information Visualization
2014, Vol 13(1) 59–89

DOI: 10.1177/1473871613488591
ivi.sagepub.com
SAGE

Ploceus: Modeling, visualizing, and analyzing tabular data as networks

Zhicheng Liu[1], Shamkant B Navathe[2] and John T Stasko[2]

Abstract

Tabular data are pervasive. Although tables often describe multivariate data without explicit definitions of a network, it may be advantageous to explore the data by modeling it as a graph or network for analysis. Even when a given table design specifies a network structure, analysts may want to look at multiple networks from different perspectives, at different levels of abstraction, and with different edge semantics. We present a system called Ploceus that offers a general approach for performing multidimensional and multilevel network–based visual analysis on multivariate tabular data. Powered by an underlying relational algebraic framework, Ploceus supports flexible construction and transformation of networks through a direct manipulation interface and integrates dynamic network manipulation with visual exploration through immediate feedback mechanisms. We report our findings on the learnability and usability of Ploceus and propose a model of user actions in visualization construction using Ploceus.

Keywords

Data transformations, graph and network visualization, multivariate data, interaction design, exploratory data analysis

Introduction

Network visualizations, often in the form of node-link diagrams, are an effective means to understand the patterns of interaction between entities, to discover entities with interesting roles, and to identify inherent groups or clusters of entities. Many existing approaches to network visualization and analysis assume a given graph. During an analysis process, however, selecting, filtering, clustering, or computing metrics over a static network is not always enough. Analysts may want to construct new networks and transform existing ones to explore the data from different perspectives and at different levels of abstraction.

The goal of our research is to provide a general approach for performing multidimensional and multilevel network–based visual analysis. We choose tabular data as the input data model considering the dominance of spreadsheets and relational databases in current data management practices. As we discuss in the following, tabular data may or may not contain explicit specification of nodes and edges in a graph, and its multivariate nature implies the need for dynamic network modeling for greater analytic power.

Forms of tabular data

Tabular data come in many forms, each unique in its schematic and semantic structures depending on the technology used and the data owner's goal. The term "tabular data" is thus fairly broad and can be interpreted as either *multivariate data* or *attribute relationship graphs*. We give examples of different types of tabular data in this section and will base our discussion on these examples throughout the rest of the article.

[1]Stanford University, Stanford, CA, USA
[2]Georgia Institute of Technology, Atlanta, GA, USA

Corresponding author:
Zhicheng Liu, 465 Wilton Ave, Palo Alto, CA 94306, USA.
Email: zcliu@cs.stanford.edu

Table 1. A table of sample visitor information to the White House.

ID	LName	FName	Type	Date	Loc	Size	Visitee
1	Dodd	Chris	VA	25 Jun 09	WH	2018	POTUS
2	Smith	John	VA	26 Jun 09	WH	237	Office visitors
3	Smith	John	AL	26 Jun 09	OEOB	144	Amanda Kepko
4	Hirani	Amyn	VA	30 Jun 09	WH	184	Office visitors
5	Keehan	Carol	VA	30 Jun 09	WH	8	Kristin Sheehy
6	Keehan	Carol	VA	8 Jul 09	OEOB	26	Daniella Leger

OEOB: Old Executive Office Building; VA: Visitor Access; AL: Agency Liaison; WH: White House; OEOB: Old Executive Office Building; POTUS: President of the US.

Table 2. Two tables describing employees and the departments they work for.

(a) Employee

ID	FName	LName	Bdate	Dpt
1	John	Smith	10 Jan 65	2
2	Franklin	Wong	9 Apr 52	3
3	Jennifer	Wallace	23 Oct 70	3
4	Ahmad	Jabbar	2 Nov 45	1

(b) Department

ID	Name	City	State	Latitude	Longitude
1	Headquarters	Los Angeles	CA	34.05	−118.24
2	Administration	San Jose	CA	37.34	−121.89
3	Research	Houston	TX	29.76	−95.36

Single tables are represented as spreadsheets and comma-separated value (csv) files. For example, Table 1 shows visits to the White House. For each visit, it records the last and first name of the person arranging the visit (LName, FName), the type of visit (Type), the date (Date) and location (Loc) of visit, the size of the visiting group (Size), and the visitee's name (Visitee). Such tabular data are essentially *multivariate data* where rows represent entities or facts and columns represent entity attributes or reference to other entities. In multivariate data, explicit definition of a graph structure is typically absent.

Multiple linked tables are often stored in relational databases, although the same tables can also be described in spreadsheets. In a relational database, the entity–relationship (ER) model[1] typically underlies database design. Each row in a table represents a fact that corresponds to a real-world entity or relationship. For example, Table 2(a) represents facts about employees in a company, and Table 2(b) represents facts about departments in the same company. The two tables are linked by a *one-to-many* DEPARTMENT– EMPLOYEE relationship type. That is, one department can have multiple employees, but one employee can work for only one department. *One-to-many* relationships are typically captured by foreign keys in a relational database.[2] In this case, Dpt in the EMPLOYEE table is a foreign key, referencing the DEPARTMENT table.

Another type of relationship in the ER model is the *many-to-many* relationship, and it is captured by a separate relationship table.[2] For example, Table 3(a) represents selected facts about research grants awarded by the National Science Foundation (NSF) in the Information & Intelligent Systems (IIS) division, and Table 3(b) represents facts about researchers. The two tables are linked by Table 3(c), which represents a many-to-many "work-on" relationship. That is, one researcher can receive multiple grants, and one grant can also involve multiple researchers.

These tabular data in multiple linked tables are essentially *attributed graphs*. Table 2 describes connections between employee and department entities. Similarly, Table 3 is a graph specifying the connection between two types of entities, researcher and grant, each with its own attributes.

An online analytical processing (OLAP) database, unlike spreadsheets and relational databases, is not built for low-level transactional operations such as insertion and update, but for retrieval, querying, and analytical purposes. It uses data cubes for better performance in operations such as slice/dice and roll-up/drill-down. The analytical power of OLAP, however, is not

Table 3. Tables describing researchers and the grants they receive.

(a) Grant

GID	Title	Program	Program Manager	Amount
1	Data mining of digital behavior	Statistics	Sylvia Spengler	2,241,750
2	Real-time capture, management and reconstruction of spatiotemporal events	Information Technology Research	Maria Zemankova	430,000
3	Statistical data mining of time-dependent data with applications in geoscience and biology	ITR for National Priorities	Sylvia Spengler	566,644

(b) Person

PID	Name	Org
1	Padhraic Smyth	University of California Irvine
2	Sharad Mehrotra	University of California Irvine

(c) Work-on

Person	Grant	Role
1	1	PI
2	1	CoPI
2	2	PI
1	3	PI

ITR: Information Technology Research; PI: Principle Investigator; CoPI: co-Principle Investigator.

necessarily suitable for network-based analysis because it focuses only on inherent relationships between entity attributes and assumes different entities are mutually independent.[3] As a result, the OLAP framework is not directly relevant for our purpose, and in this article, we focus on spreadsheets and databases, which provide a basis for an alternative network-centric framework.

Analytical gap and semantic distance

For visualization designers and analysts, spreadsheets and databases naturally become the infrastructure upon which higher level visual analysis is accomplished. As discussed in the previous section, multivariate data in the form of single tables do not contain explicit network semantics; even when multiple tables are used to describe a graph, analysts' own notions of a meaningful network may render different graph structures. First, the concept of an entity is often multilevel and nested: an attribute of an entity may be treated as an entity in its own right. For example, in Table 3(a), each row represents a grant entity with its own attributes such as title and program manager. A program manager can be in turn treated as an entity. In fact, it is often difficult to determine whether something is an entity or an attribute in data schema design.[4] Second, the same two entities can be connected via different semantics. In Table 1, for example, two people can be connected if they visited the same location, have the same last name, or started their visits on the same day.

The multivariate nature of tabular datasets thus implies opportunities for asking interesting questions that can be answered with network visualizations, and it is worthwhile to examine the nature of such questions more closely. Given the dataset in Table 3, for example, a grant applicant may want to understand the hidden dynamics, if any, in the process of awarding grants to choose an appropriate application strategy. NSF officials will want to understand the impact of the IIS program on the awardee social networks and on the creation and diffusion of intellectual property to evaluate funding policy. Many questions can thus be asked, for instance,

- Q1: Is there a strong affiliation between program managers and research institutions? That is, do certain program managers tend to give awards to a few selected institutions only?
- Q2: From which organizations do researchers tend to have more cross-institution collaborations?

One possible way to answer Q1 is to construct a network visualization (Figure 1(a)) where an

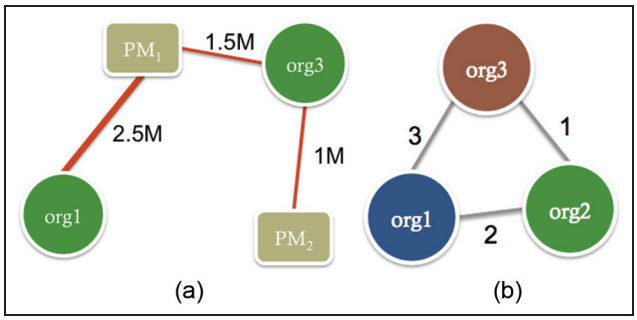


Figure 1. Visual models for answering questions on the NSF dataset. (a): a sample network where a program manager is connected to an organization if the manager has given grants to researchers from the organization; edge weight indicates amount of grants, (b): a sample network where an organization is connected to another organization if they have received grants together; edge weight indicates number of grants.
NSF: National Science Foundation.

organization and a program manager are linked if the manager has awarded at least one grant to researchers in that organization. We can define the edge weight to be the total grant amount as shown in Figure 1 or to be the number of grants awarded. Analysts can provide initial answers to Q1 by inspecting the overall connectivity of the network. If the network consists of multiple small subnetworks that are disconnected from each other, there is evidence that a strong affiliation does exist. It is also likely that there is no disconnection within the network, but certain organizations or managers occupy more central roles. Statistical measures will enhance visual inspection to provide a more precise assessment.

Similarly, to answer Q2, we can create a network visualization where two organizations are connected by an edge if there is at least one collaboration between any researchers from these two organizations. Figure 1(b) shows this network semantics, where the edge weight is based on the frequency of collaboration. Applying an appropriate layout algorithm to this network visualization and using statistical measures such as betweenness centrality will likely reveal important organizations that are "gatekeepers" connecting different subgraphs.

These questions have two major characteristics. First, they cannot be answered satisfactorily by simple "yes" or "no" or precise quantification. Analysts can define metrics to measure "affiliation strength," for example, in the case of Q1, but such metrics are only meaningful at the level of specific manager–institution pairs. Network visualizations are useful to show global structures in the network. Second, these questions are semantically rich and context dependent and cannot be described abstractly or captured a priori because they usually only emerge during the process of exploration.

Amar and Stasko[5] considered answering such questions as performing *high-level* analysis tasks, which can be contrasted with *low-level* tasks[6,7] that are usually topology based or attribute based. Topology-based tasks include finding neighbors, counting degree, finding shortest paths, and identifying clusters; attribute-based tasks include finding nodes with specific attributes or finding nodes connected by particular type of edges. Many of these low-level tasks are well-defined questions with clear-cut answers, and they can often be effectively answered using search or database queries without much visual representation.

Supporting only low-level tasks creates analytic gaps in addressing real analytic and sense-making goals. Many high-level tasks require analysts to go beyond manipulating a static network and to actively construct and simulate a model.[8] Illustrations of analysts' desired model based on their analytical questions are given in Figure 1(a) and (b). To effectively support model-based reasoning, analysts must be able to quickly choose the relevant entities and relationships for model construction.

The model will be subject to constant refinement and revision, where new variables and relationships are introduced and old ones transformed or discarded. Dynamic articulation of fluid network semantics is thus necessary, and the multivariate nature of many tabular datasets provides a fertile playground for performing this kind of model-based reasoning.

Objective and organization

With these considerations in mind, we present Ploceus (Ploceus is a kind of weaver bird that can build sophisticated nests), a system designed to support flexible network–based visual analysis of tabular data. Our focus is not on representation and interaction techniques for visually analyzing a given network; a number of commercial and research systems have been designed for this purpose.[9–15] Rather, we aim to address flexible and rapid construction and manipulation of networks from tabular data. The power of Ploceus is based upon a formal framework that systematically specifies operators for network construction and transformation and the implementation of these operators in relational algebra. A direct manipulation interface is coupled with the formalism to help analysts articulate the desired network semantics.

This article is an expanded and updated version of a paper presented at the VAST 2011 Conference.[16] In this version, we present research findings on additional issues related to network-based visual analysis in tabular data. More specifically, section "Visual encoding" explores automatic visual encoding of networks modeled from data tables. Sections "Extending to one-mode networks" and "Extending to one-mode graphs"

extend the scope of our formal framework and the system to support the construction of one-mode networks from reflexive relational tables. Section "User evaluation" reports our findings on the learnability and usability of Ploceus from a qualitative user study. We identify potential roadblocks in network modeling and propose a model of user interaction with Ploceus.

Related work

Visualizing multidimensional data

Systems such as DEVise,[17] Table Lens,[18] FOCUS,[19] InfoZoom,[20] Polaris,[21] and Tableau[22] visualize tabular data in the form of line charts, bar charts, scatterplots, or space-filling cells for analyzing distribution pattern and frequency aggregation. None of these systems pays special attention to the potential of imposing user-defined relationships between attribute values in the form of networks. Our motivation behind designing Ploceus does resonate with the approaches taken by Polaris and Tableau, which advocate the need for analysts to rapidly change the data they are viewing and how the data are visualized, as well as the need to integrate data transformation and visual abstraction in a seamless process.

Jigsaw[23] builds the semantics of relationships into the system design based on a simple assumption: Entities are identified purely lexically, and entities appearing in the same documents are connected. This approach, originally designed for unstructured text documents, can be extended to tabular data such as spreadsheets: one row in a table is equivalent to the notion of a document. The co-occurrence-based definition allows flexible explorations of entity relationships without having to explicitly specify the nodes and edges, but since the fundamental connection model is centered around documents/rows, the connections between table columns are less direct. Jigsaw also has limited data transformation support due to its indiscrimination between nominal and quantitative entities.

Network visualization and analysis

A number of systems in the form of toolkits[24,25] or executables[9–12,15,26,27] are available for analyzing a given graph. These systems vary in features and provide visualizations, computational metrics, or both. NodeTrix[14] explores how these different network representations can be integrated for the same underlying graph data. ManyNets[28] looks at visually exploring multiple networks. PivotGraph[29] provides attribute-based transformation of multivariate graphs. Creating and transforming network semantics from data tables are not the main focus of these systems.

NodeXL[13] integrates with Microsoft Excel to enable users to easily import, transform, visualize, and analyze network data. NodeXL stores network data in multiple sheets representing nodes and edges, and users likely will need to be Excel experts to be able to transform the data.

Attribute relationship graphs

Ploceus focuses on extracting trees and graphs from data tables, and variants of this idea have been explored in prior study. The need for retrieving and publishing selected information on the web leads to work that models databases as virtual graphs[30] and provides Extensible Markup Language (XML) document interfaces of relational data for web applications.[31] The Grammar of Graphics discusses an algebraic framework for mapping tables to directed trees.[32] Weaver proposed a data transformation pipeline for attribute relationship graphs[33] and is perhaps the closest to our algebraic framework presented in section "Computing connections."

A number of systems appear to be close to Ploceus in terms of design goal and functionality, including CineGraph demonstrating the attribute relationship graphs approach,[33] two Commercial systems TouchGraph Navigator[34] and Centrifuge,[35] and the Orion system.[36] Weaver[33] distinguishes between attributed graphs (where an object is connected to its attributes) and attribute relationship graphs (where attributes are connected based on occurrence). This notion of attribute relationship graph lays the foundation of our study. Weaver's Cinegraph system supports many network modeling operations such as deriving attributes and slicing and integrates the generated network with cross-filtered views.[37] TouchGraph Navigator[34] and Centrifuge[35] provide interfaces for creating attribute relationship graphs from data tables. The Orion system,[36] published concurrently with our VAST paper, also supports the construction and transformation of network data.

While Ploceus is not the first system that investigates the connection between data tables and graphs, we distinguish our study from related systems in two ways. First, we offer a comprehensive construction and transformation framework that integrates diverse operations in a flexible yet systematic manner (section "Operations" describes these operations in detail). Table 4 summarizes the operations provided by Ploceus and the related systems. The same operation may be named differently in different systems, and we provide the terms used by these systems whenever possible. Due to the inaccessibility of the commercial software TouchGraph Navigator,[34] we cannot do a comprehensive assessment of its features and thus

Table 4. A comparison between different systems in terms of the network modeling operations provided.

	Ploceus	Centrifuge	ARG	Orion
Create nodes	✓	✓	✓ (group)	✓ (promote)
Add attributes	✓	×	×	×
Create connections	✓	✓	✓ (clique)	✓ (link)
Assign weights	✓	×	✓ (weight)	✓ (weight)
Project	✓	×	×	*
Aggregate—pivoting	✓	×	×	✓ (roll-up)
Aggregate—binning	✓	✓	×	×
Aggregate—proximity grouping	✓	×	×	×
Slice 'n dice	✓	✓	✓ (slice)	✓ (split)
Filter by value	*	×	✓ (drill)	✓ (filter)

omit it from the comparison. As is evident from the table, all the systems provide support for basic operations such as creating nodes and connections. Adding node attribute, pivoting, projecting, and proximity grouping are absent in one or more of the other systems. Due to different interface designs, sometimes, there is no direct one-to-one mapping between the operations in our framework and those in other systems, and we indicate such situations as "*" in Table 4. For example, in Orion, there is no "project" operation, but users can create one-mode networks by "promoting" a column to a table and connect the column to itself.[36] Similarly, in our framework, there is no explicit "filter by value" operation, but analysts can identify specific node values through the search function provided at the interface level.

Second, we take a human-centered perspective in interface design. While systems such as Orion and CineGraph provide similar modeling operations, their interfaces are significantly different from those of Ploceus. In our design considerations, we aim to expose each network modeling operation as a conceptually meaningful unit to the users and map each operation to an action or an interface element. In particular, we design and implement a network schema view based on the notion of "visualization schemas"[38] to enable analysts to construct a network by combining the modeling operations without the need for programming.

Ploceus: overview

Ploceus provides a direct manipulation interface for fast construction and transformation of networks and shows immediate visual feedback on the network being created. Model construction and visual exploration are integrated. Ploceus contains three major views: a data management view on the top left, a network schema view on the bottom left, and a network view on the right (Figure 2). The data management view shows information about the columns in each table in a dataset; the network schema view is a sandbox-like environment where users can construct and manipulate networks at a conceptual level; the network view shows the corresponding network visualization and updates whenever the network schema is modified.

Operations

Ploceus currently supports the following types of operations. We describe these operations at a functional level in this section and discuss the precise mechanisms of accomplishing these operations in section "Computing connections."

- *Create nodes.* Transform the values in one or more columns into node labels. For example, we can construct a set of nodes representing the people visiting the White House from all the rows in Table 1 and can create the labels of the nodes from the LName and FName columns. This results in four nodes: "Dodd, Chris," "Smith, John," "Hirani, Amyn," and "Keehan, Carol."
- *Add attributes.* Transform the values in one or more columns as attributes of existing nodes. For example, we can add an attribute AccessType to the people nodes constructed from LName, FName earlier. The node "Dodd, Chris" will have the value "VA" for the AccessType attribute. Ploceus also supports adding columns as attributes from a different table. For example, we can add Role from Table 3(c) as an attribute for the Name nodes constructed from Table 3(b). Ploceus only allows a node to have one value for any particular attribute, so there will be two "Sharad Mehrotra" nodes in this case, one having a PI role and the other having a CoPI role.
- *Create connections.* Create edges between existing nodes. For example, we can connect LName, FName nodes and Loc nodes from Table 2 to see the visiting patterns by the visitors to the various locations. We can also connect nodes created from different tables, for example, ProgramManager nodes from Table 3(a)

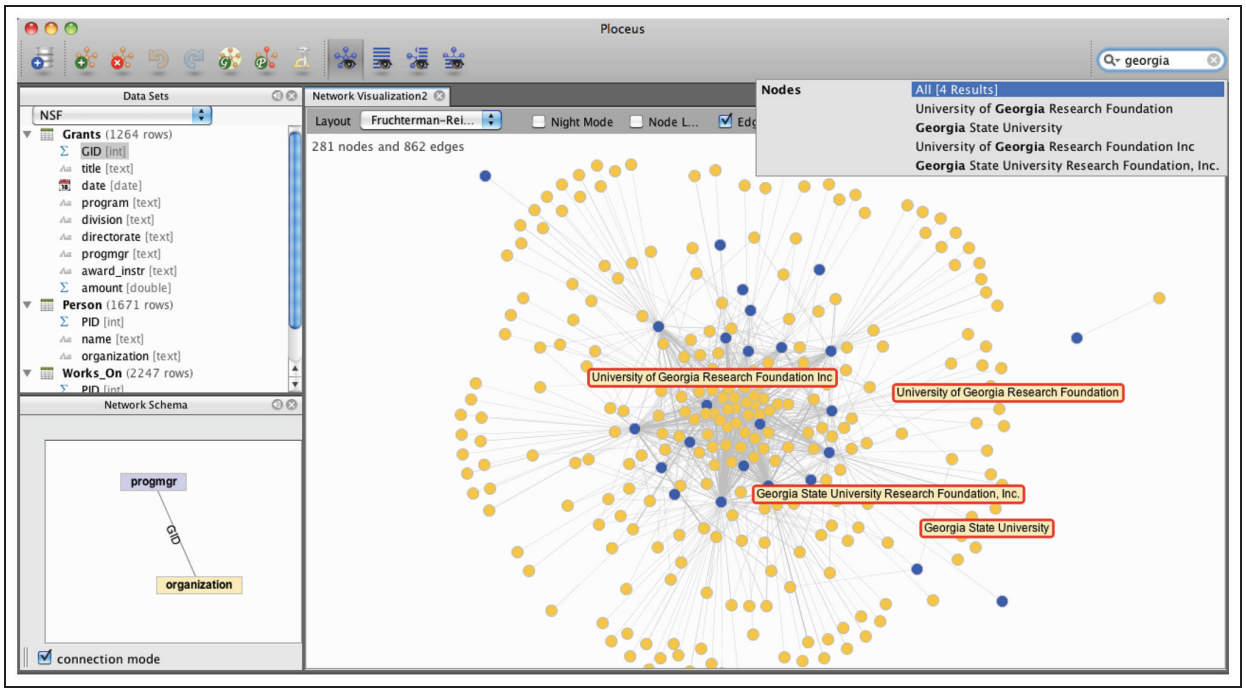


Figure 2. Ploceus system interface with a data management view on the top left, a network schema view on the bottom left, and a network visualization view on the right.

and Org nodes from Table 3(b). When multiple tables are involved, Ploceus tries to determine how the tables should be joined by analyzing the foreign key constraints between the tables using the Dijkstra shortest-path algorithm[39] (section "Higher-order graphs: transformation" provides technical details on this issue). In this case, the two tables are joined through Table 3(c). Ploceus computes whether there should be an edge between any two nodes as well as assigns a weight to that edge. When multiple ways of joining tables are possible, users can specify the join condition through a dialog.

- *Assign weights.* Assign numerical weights to edges. Ploceus by default assigns a weight to each edge created, indicating the frequency of co-occurrence between the nodes in the data (sections "Edge semantics and construction strategies" and "Computing connections" discuss edge weights in greater depth). For example, if we connect LName, FName nodes and Loc nodes from Table 1, by default, the edge between "Dodd, Chris" and WH has a weight of 1, indicating this person has visited the White House once in this dataset. We may instead want to represent the connection strength by the number of people he has brought on his visits and assign the column Size as the edge weight. The edge between "Dodd, Chris" and WH will have a weight of 2018. Only a single column can be assigned as edge weight, and that column must be quantitative.
- *Project.* Connect two nodes if they both are connected to the same node of a different type. Projection is a commonly used technique to reduce modalities of a network for analysis.[40] In a two-mode (i.e. there are two types of nodes) LName, FName—Loc network, for example, if "Dodd, Chris" is connected to "WH" (i.e. Chris Dodd visited the White House), and if "Keehan, Carol" is connected to "WH" also, after projecting LName, FName nodes on Loc nodes, "Dodd, Chris" and "Keehan, Carol" are connected. Figure 3(a) shows this process. The weight of edges after projection reflects the unique number of Loc nodes being projected.
- *Aggregate.* Group multiple nodes and treat them as one node. Ploceus automatically aggregates nodes with identical labels if no attributes are specified for these nodes and aggregates nodes with identical labels and values if attributes are specified for the nodes. As a result, we have four distinct LName, FName nodes from Table 1, while there are actually six rows in the table.

Other types of aggregation include, but are not limited to, the following:

- *Pivoting.* PivotGraph[29] terms this operation *roll-up*. Given LName, FName nodes with the attribute AccessType, we can aggregate people nodes

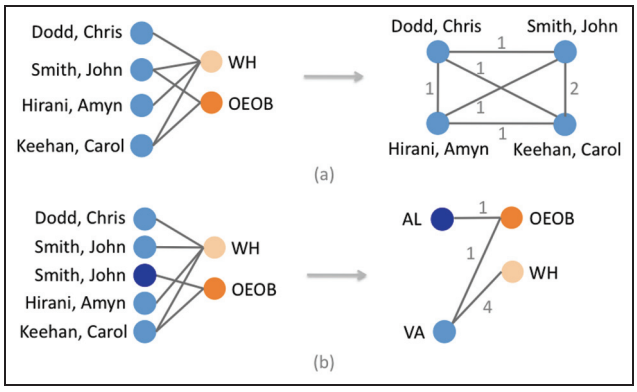


Figure 3. Project and pivot operations: (a) the visitor nodes do not have attributes and (b) the visitor nodes have attribute "Type" (with values "AL" and "VA") from the original data table.
OEOB: Old Executive Office Building.

when they share the same AccessType. The pivoting process is visualized in Figure 3(b). The resulting graph shows the locations that are typically visited for different types of visits.

- *Binning*. For nodes whose labels or attributes are derived from quantitative columns, value-based aggregation is possible. One type of value-based aggregation is *binning*: we divide the range from the minimum to the maximum attribute values into bins. For example, we can categorize Amount nodes created from Table 3(a) into three bins: "small" if Amount $\leqslant$ 500 K, "medium" if 500 K $<$ Amount $\leqslant$ 1200 K, and "large" if Amount $>$ 1200 K.
- *Proximity grouping*. Group nodes in a pairwise manner if they have values close to each other. For example, from Table 2(b), we can create City nodes with attributes Latitude and Longitude. We can then aggregate every pair of City nodes into one for which the distance between them, computed from the latitude and longitude information, is within 500 miles. This operation is combinatorial: if there are four cities, and everyone is within 500 miles of each of the other three, proximity grouping will produce $\sum_{k=1}^{(4-1)} k = 6$ nodes. Proximity grouping is useful when combined with projection, so that we can, for example, create a network of employees whose workplaces are within 500 miles to each other (to do this, connect employee names with cities, aggregate cities, and then project employees on cities).
- *Slice 'n dice*. Divide a network into subnetworks based on selected columns. For example, given that we have constructed an LName, FName—Visitee network from Table 1, we may want to see how the visiting pattern is related to the locations of visits by dividing the network using Loc slices. We will then have two subnetworks, one representing the visiting patterns at the White House ("WH") and the other at the Old Executive Office Building ("OEOB"). Slice 'n dice thus enables analysts to create and organize meaningful snapshots of a big network based on different perspectives. The values in columns used for slicing and dicing are either categorical or can be categorized. When hierarchical categories exist, analysts can slice and dice at multiple granularities, for example, for a *date* column: day → week → month → quarter → year.

We try to be comprehensive in choosing the relevant operations to be included based on three criteria: (1) The operation is indispensable for creating a basic network, (2) earlier related work shows the utility of the operation, and (3) the operation is considered useful based on our own experience in performing network-based visual analysis. Section "Expressive power" explores the issue of expressive power offered by these operations. In addition to these higher level operations for creating and transforming networks from data tables, Ploceus supports interaction with individual nodes such as selecting, filtering, moving, hiding, showing, and expanding (showing neighbors of a node); interaction with the visualization in the form of zooming, panning, adding new visualizations, and deleting existing visualizations; applying various network layout algorithms; and analytical measures such as node degree, shortest path, betweenness centrality, and closeness centrality. These features, though not the main focus of our research, are essential for integration with the above-mentioned operations for more complete user experience in performing data transformation and visual exploration.

Design of direct manipulation interface

In designing Ploceus, we wanted to make the interface accessible for users who do not necessarily possess programming skills. To achieve this goal, it is desirable to reduce articulatory distance, that is, assuming the analysts want to perform some operations, what is an intuitive way for them to communicate the intent to the system.

One possible design is to integrate a visual interface with a scripting interface as done in GUESS[12]—the manipulation of graphs is in the form of commands. The advantage of this approach is that script languages are precise, expressive, and concise; the disadvantage of this approach is that analysts must understand basic programming concepts.

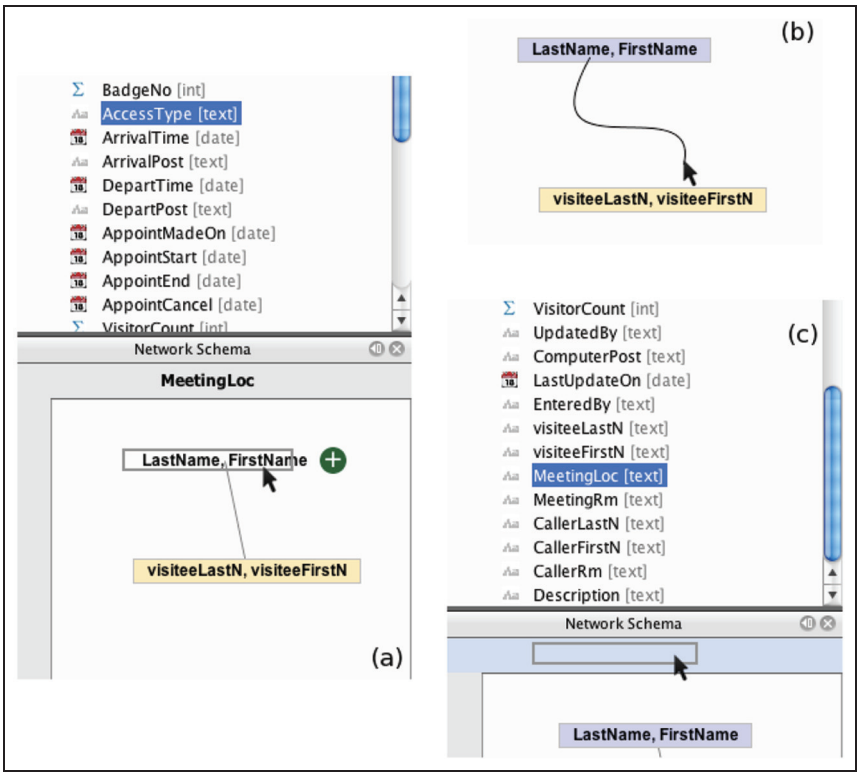


Figure 4. Direct manipulation interfaces for various operations: (a) add attributes, (b) create connections, and (c) slice 'n dice.

Another design alternative is programming by demonstration (PBD).[41] PBD typically uses a direct manipulation interface. For example, to create connections between LName, FName nodes and Loc nodes constructed from Table 1, assuming that we already have a visual representation of the existing nodes, we can use a click–drag–drop mouse gesture to connect a visitor node (e.g. "Dodd, Chris") and a location node (e.g. "WH"). After we have performed similar gestures two or three times, the system will figure out that our intention is to connect LName, FName nodes and Loc nodes and will perform the same operation automatically on the rest of the nodes.

PBD arguably shortens articulatory distance when it works on a direct manipulation interface. As shown in the create-connection example, users perform the exemplary operation at the level of individual data items, and the system generalizes from the user interaction to a high level by connecting different types of nodes. This bottom-up design approach has some shortcomings for network modeling. Analysts need to know if an edge indeed exists between two specific nodes, and thus, they need to access a low-level representation of the raw data and understand the mechanism of edge computation.

Our final design decision is to adopt a direct manipulation approach akin to that of Polaris,[21] Tableau,[22] and the visualization schemas approach.[38] Analysts directly interact with high-level conceptual representations of the relational data schema and indicate intention by manipulating these representations.

To create nodes, analysts drag and drop selected columns from the data management view to an empty area in the network schema view (Figure 2). Each drag-and-drop action creates a type of node, and the system assigns a color to that type. Dragging and dropping columns on top of an existing node type add those columns as an attribute to the node type (Figure 4(a)).

Given two types of nodes, analysts create connections between them by clicking on one type of nodes and dragging the mouse to the other type of nodes in the network schema view (Figure 4(b)). This action draws an edge between the two that takes effect when the mouse button is released. To designate a quantitative column as edge weights, analysts drag and drop the column over the edge representation in the network schema view. Ploceus supports slicing and dicing for up to two dimensions, designated as the horizontal and vertical axes in the visualization. Analysts specify the orientation of the slices (horizontal or vertical) by dropping columns to the appropriate shelf (Figure 4(c)).

Analysts specify aggregation and projection, two transformative operations on existing networks, using

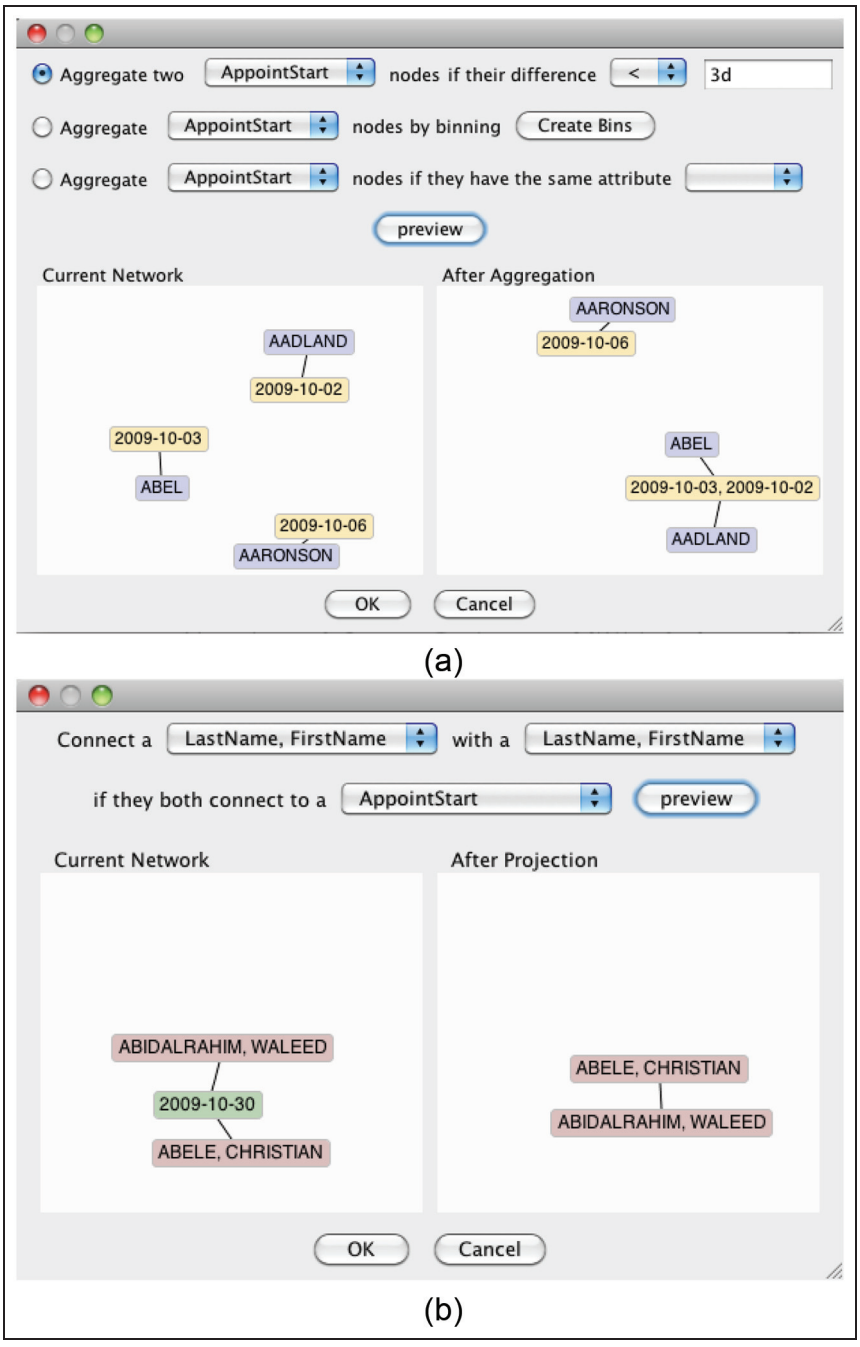


Figure 5. Dialogs for specifying (a) aggregation and (b) projection.

dialog interaction rather than drag and drop. We make this design choice considering the fact that it is difficult to articulate these two operations within the network schema view. Dialog-based interfaces provide text-based controls that are easy to understand. Our design uses combo boxes to let analysts specify the type of nodes they want to perform these operations on.

Currently, Ploceus supports three types of aggregation operations: proximity grouping, binning, and pivoting (Figure 5(a)). Analysts choose the type of aggregation through radio buttons. Depending on the properties of nodes selected, some operations may not be applicable. For example, when nodes have no attributes, pivoting does not make sense. To specify projection, analysts indicate through combo boxes the types of nodes to be projected (Figure 5(b)). Both dialogs offer previews of how the network will appear after the transformation, so that analysts can have a feel of the consequences of their actions.

Whenever analysts perform an operation, the network view provides immediate feedback in the form of a node-link visualization of the current network (Figure 2). Analysts can interactively add selected nodes and edges to the visualization through a search query field on the top right corner of the system toolbar (Figure 2). Analysts can also switch to a list-based view where different types of nodes are displayed in lists and the nodes are sorted by analytical metrics such as centrality. When the size of the network exceeds a threshold (currently defined as 450 nodes), to avoid screen clutter and low system performance, the node-link visualization will randomly sample and show a subpart of the network; the list-based visualization still shows the entire network.

When slice 'n dice dimensions are specified, Ploceus shows a grid containing multiple small networks in the form of node-link visualizations only with brushing support (e.g. Figure 14). If the dimension used contains [many distinct] categorical values, the large number of subnetworks can lead to usability and performance problems. In our current design, users can scroll to see subnetworks hidden from the current viewport. Systems such as ManyNets[28] will be useful to visualize summary statistics of these subnetworks for easy comparison.

Visual encoding

The direct manipulation interface supports the articulation of operations that determine the constituent data of desired networks. Subsequently, it is important to visualize these networks appropriately. Commonly used visual variables to encode data dimensions are color, size, and spatial positions. In particular, spatial position encoding, or graph layout, plays an important role in showing prominent visual structures such as clusters and outliers.

The node-link representation included in Ploceus supports five layout algorithms: Fruchterman–Reingold force-directed layout,[42] circular layout, spring layout, the Kamada–Kawai algorithm,[43] and Meyer's self-organizing layout.[44] The effectiveness of these layout algorithms often depends on the specific properties of the network being visualized, and analysts can experiment with different algorithms through a combo box.

In addition to providing mechanisms for spatial position encoding, Ploceus intelligently infers the appropriate visual encodings for the nodes based on the type of the underlying table dimensions. Studies have examined how people perform in perceptual tasks in terms of accuracy when different information types (quantitative, ordinal, and nominal) are represented using different visual variables (e.g. area, color, and density).[45,46] Researchers have explored the issue of automatic graphic encoding[46,47] by incorporating these established design principles into system logic for effective visualization.

Ploceus draws from these research findings to apply effective graphic presentations in the visualization of networks. As noted in the previous section, when analysts create a new type of node, Ploceus automatically assigns a new color to that type. In addition, when analysts designate table column(s) as attributes of nodes, Ploceus analyzes the type of the column(s) to choose a visual mapping. Currently, Ploceus supports four types of column types: integer, float, date, and text (string). If analysts assign a quantitative column (integer or float) as node attribute, Ploceus will sum up the quantitative values and represent it using node size. For example, after adding an attribute Size to the people nodes constructed from LName, FName in Table 1, the node "Smith, John" will have a value of 381 for the Size attribute. Figure 6 shows the resulting visualization based on a larger dataset of the White House visitor information. When analysts designate a date column as a node attribute, Ploceus represents dates using node size by converting date values into the number of milliseconds since 1 January 1970, 00:00:00 GMT.

Encoding a categorical node attribute is a design decision requiring more consideration. It is arguably best practice to use visually distinct colors to represent a categorical variable.[48] Alternative ways of encoding categories are to use texture or shape.[46] In any case, when there are many unique categorical values, it is difficult to define enough visually distinct representations.

In Ploceus, we use color to represent the type of node as discussed in section "Design of direct manipulation interface." This initial decision implies that we may have to use shapes to represent categorical node attribute values. The number of visually distinct shapes, however, is limited compared to the available choice of distinct colors.[49] Assuming that analysts usually will create relatively few node types, we experimented with using shape to encode node type and using color to encode the categorical attribute values instead. Figures 7 and 8 show visualizations generated with this approach. Figure 7 shows a network of the White House visitors (represented as circles) and visitees (represented by diamonds). Ploceus assigns a default color to all the nodes. Figure 8 shows the resulting network after we add an attribute denoting the meeting location to the visitor nodes, which is represented using color.

However, informal feedback gathered from visualization experts on this approach was not positive. They strongly preferred encoding node type as color instead

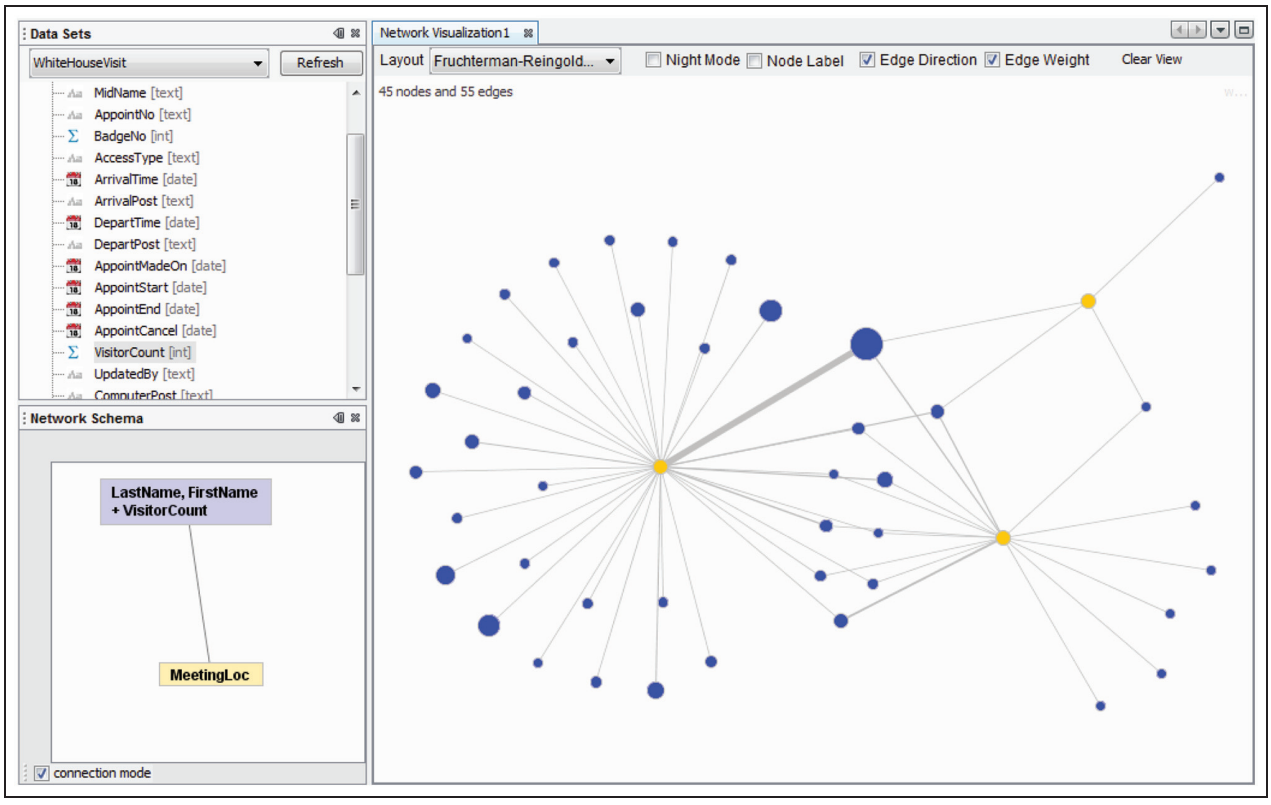


Figure 6. Ploceus visualizes the attribute VisitorCount of the visitor nodes constructed from LastName, FirstName as node radius (size).

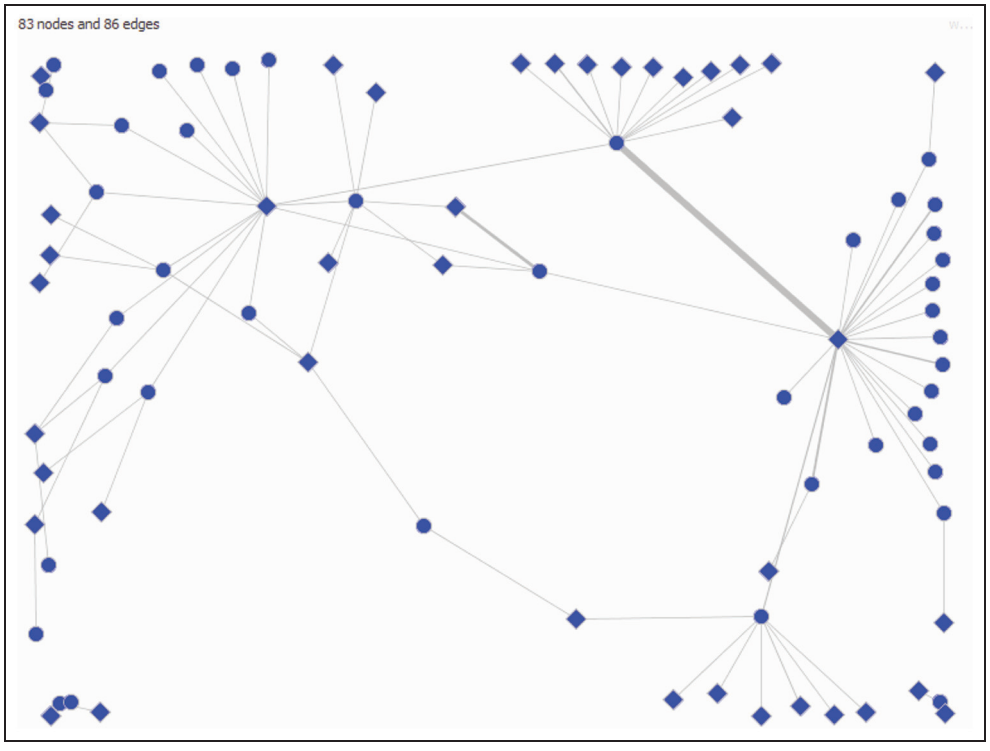


Figure 7. A network of the White House visitors, represented as circles, and visitees, represented by diamonds.

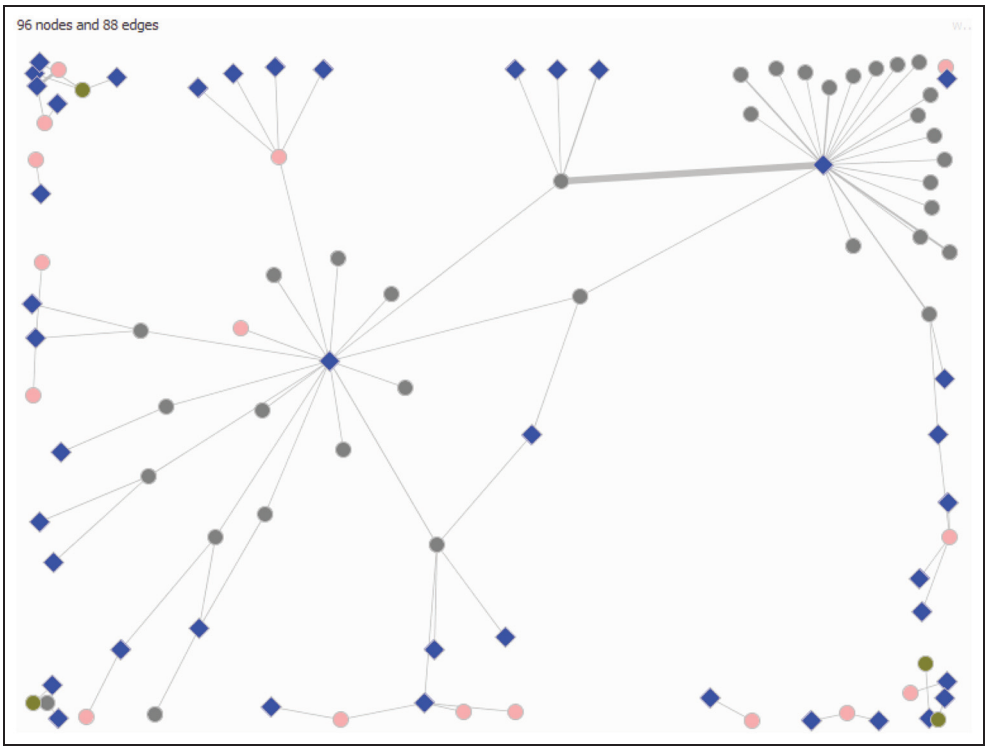


Figure 8. Adding "meeting location" as an attribute to the circular visitor nodes and representing the attribute using color.
Visitee nodes remain as blue diamonds.

of shape and suggested that it was potentially confusing to interpret (Figure 8). Since node type can also be considered as a node attribute and is automatically generated, it is more essential than an optional node attribute defined by users. We thus decided to treat node type as the default node attribute and to continue encoding it using color. Analysts can define new attributes by dragging and dropping columns onto the existing nodes, and if the new attribute is categorical, it will be color coded and replace the default color assigned to the node type. Users can change which node attribute to encode through a pop-up menu. Figure 9 shows the visitor–visitee network, where visitees are in yellow and the visitor nodes are colored by the locations of their visits. In our current implementation, Ploceus only supports one active attribute for a node type: adding a new attribute will replace any existing attribute assigned to the node type. This design decision was made to keep the current implementation tractable. In future versions of Ploceus, we plan to remove this constraint.

The study described in this section lays the groundwork for further investigation of a comprehensive graph visualization framework. The Polaris formalism[21] establishes an algebraic framework for table-based visualizations that provides effective mapping from data variables to visual variables. We envision that a similar framework is possible and is needed to describe the mappings between attribute relationship graphs and various graph visualizations. Such a framework will be useful for automated generation of graph visualizations and may suggest visualization techniques that have not been explored before.

Edge semantics and construction strategies

With such a set of diverse operations provided, it is important for analysts to correctly interpret the edge semantics in the networks created. When a network is created from a single table, the interpretation is usually straightforward. For example, connecting a visitor to a location indicates a visiting relationship, and the edge weight means frequency of visit. When these two types of nodes are from different tables, how the connections are constructed will affect the numerical weights assigned to the edges and how the edges are interpreted. For example, we can directly connect Program Manager nodes from Table 3(a) and Org nodes from Table 3(b), and the meaning of connection is that of managers granting awards to organizations. The exact meaning of the edge weight, however, is more subtle. Ploceus will determine that Table 3(c)

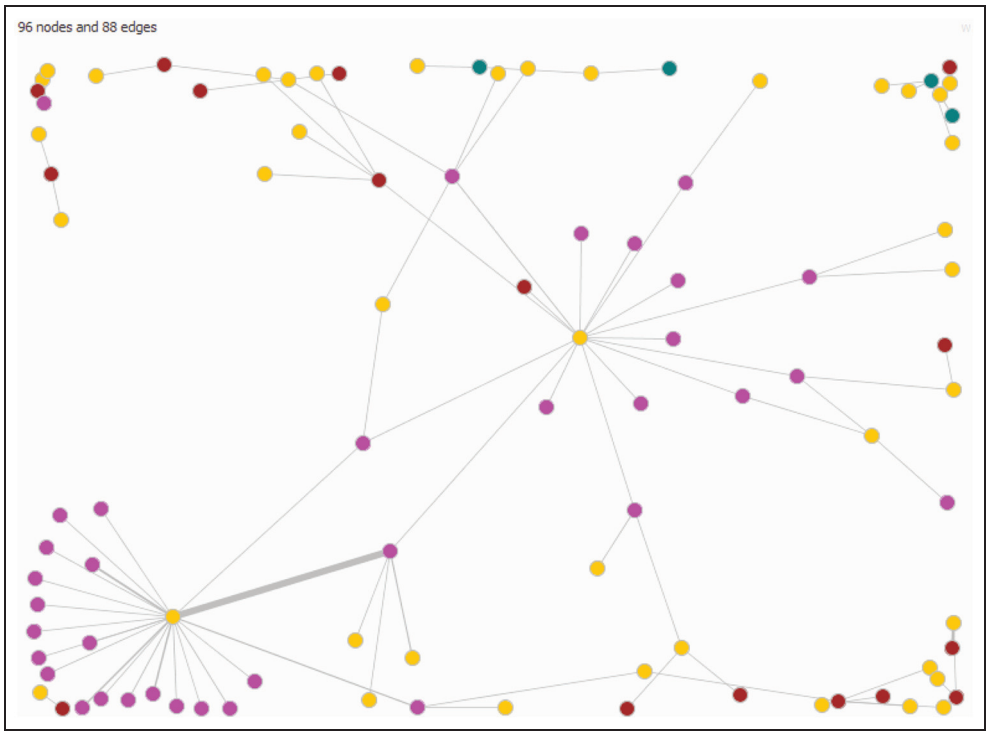


Figure 9. Ploceus visualizes the attribute MettingLoc of the visitor nodes constructed from LName, FName as node color. The visitee nodes are in yellow, and the visitor nodes are colored by the locations of their visits.

already defines an explicit network relationship of Researcher×Grant (or GID×PID). This relationship is used to create edges, and as a result, the edges between program managers and organizations will have the semantics of ProgramManager – GID×PID – Org. The edge between Sylvia Spengler and University of California Irvine, for example, will have a weight of 3, indicating that she has awarded grants to researchers from this organization three times (to Sharad Mehrotra once and to Padhraic Smyth twice). That is, both the number of researchers per grant and the number of grants will have an impact on the edge weight, and the edge weight is determined by the number of occurrences of the (GID, PID) pair.

However, this weight may not be at the right level of abstraction to the analyst, as Sharad Mehrotra and Padhraic Smyth have collaborated on a grant, and the program manager has in fact only awarded two grants to the organization. To let the weight reflect the number of unique grants awarded by the program manager to the organization only, we can connect Program Manager and GID explicitly first and then connect GID with Orgs. We then do a projection by connecting a Program Manager with an Org if they both connect to the same GID. The weight assigned to the edge between Sylvia Spengler and University of California Irvine will then be 2, indicating two grants.

These subtleties of edge construction reinforce that we can create connections between nodes with great flexibility and rich semantics. A program manager and an organization, for example, can be connected by the grants awarded by the manager to the organization, by the frequency of awards to researchers from this organization, or by the researchers from the organization who receive grants from the manager. This power comes with the requirement, however, of knowing the right operations to create the desired semantics. To help analysts keep track of what they are doing when connecting nodes from different tables, Ploceus labels the edge representation in the network schema view, indicating the semantics of the edges. Figure 10(a) shows the label for the first case and Figure 10(b) shows the label for the second case discussed in this section.

Visualization management and work flow

Another consequence of providing a variety of construction and transformation operations is that it is now easy to generate a large number of distinct networks. Managing the networks thus becomes an

important issue in the design of the user interface. In Ploceus, every network generated is associated with a tab. Analysts can generate new blank networks through the toolbar "New Network" button, and closing a tab deletes the network. Within each tab, analysts can switch between a node-link visualization and a list-based visualization; they can also tile these two visualizations side by side.

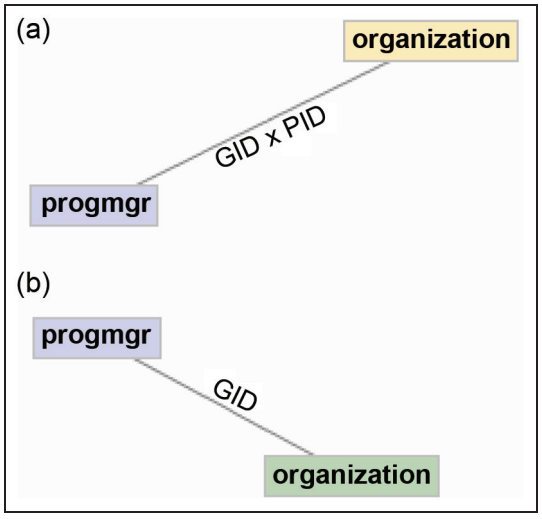


Figure 10. Edge semantics labels in the network schema view: (a) the edge weight between a program manager *p* and an organization *o* denotes the number of researchers from the organization *o* who have received grants awarded by *p* and (b) the edge weight represents the number of unique grants that a program manager has given to an organization.

In the case of slicing and dicing, analysts can right click on any of the subnetwork and choose "Analyze in detail" in the pop-up menu. Ploceus will display the chosen subnetwork in a new tab, where analysts can examine it more closely and change the representation to list-based visualization. In this newly created tab, Ploceus remembers the specific slice 'n dice dimension values associated with the subnetwork, so analysts can choose to delete the network while keeping the slice 'n dice values for further exploration of alternative networks from the same perspective. Whenever a new network is created or deleted, or an existing network is transformed, the network schema view will update accordingly to reflect the schema of the network in the currently active tab. Every network generated can be saved as a GraphML file and be reloaded into Ploceus. Figure 11 shows an overview of the work flow in using Ploceus.

Scenario: analyzing cross-institution research efforts

To illustrate how to use the direct manipulation interface in conjunction with the visualization and computational capabilities provided by Ploceus for fast analytical insights, we present an example analysis in this section. For a more interactive and complete view of the analytic process, we refer readers to the accompanying video.

In this scenario, we examine the research grants awarded by the NSF in the IIS division from 2000 to 2003. A subset of the data is presented in Table 3. It is

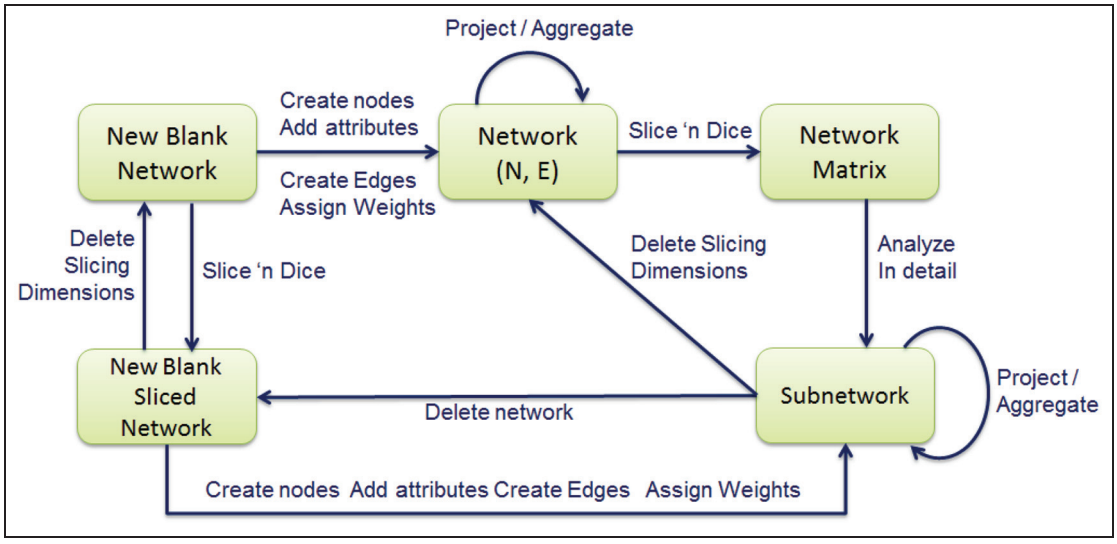


Figure 11. An overview of the work flow in using Ploceus.
Different states of the network are shown in rectangles, and the arrows represent the user interaction to transit between the states.

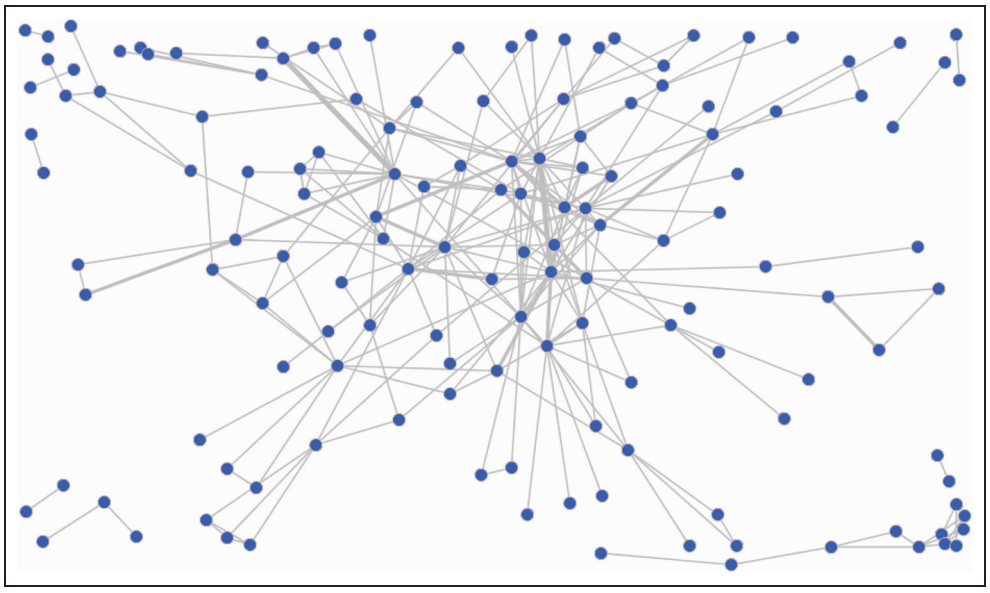


Figure 12. Collaboration between organizations on NSF IIS grants, 2000–2003.
NSF: National Science Foundation; IIS: Information & Intelligent Systems.

a long-standing policy of NSF to encourage interinstitution research collaborations, and it would be of interest to understand the structure of collaboration networks at an organizational level. In particular, researchers from which organizations tend to collaborate with colleagues from other institutions? What factors might have influenced the collaborations?

The dataset specifies an explicit 2-mode network at the actor level (PIs/co-PIs with grants). To construct a network at the organizational level, we drag and drop the organization column from the person table and the GID column from the Grant table to the network schema view and connect these two types of nodes. Immediately, we have a network showing the connections between organizations and the grants they have received. To establish a direct linkage between organizations, we perform a projection on the GID nodes. Since we are only interested in organizations that have collaborated with at least one other organization, we filter out the organization nodes whose degree is 0. The network shown in Figure 12 results. We can see that the network is fairly well connected, with a few very small clusters detached from the main network. This indicates that the collaboration over the years is not segregated in isolated clusters, which is a positive sign. Switching to a list-based view and ranking the organizations by degrees (Figure 13), we see that Stanford University, University of California Berkeley, University of Washington, Columbia University, and Georgia Tech are the top five cross-institution collaborators. It is also interesting to note that Georgia Tech is the only one in the top 5 that has not collaborated with the other four organizations in the top 5.

We can continue to explore the collaboration patterns of individual organizations, but to get a more systematic view of the structure of this network first, it may make sense to slice and dice it by both the year and the amount of the award. Assuming that we have defined how the amount dimension should be aggregated into categories, this gives us the network matrix in Figure 14. The visualization here seems to conspicuously refute our intuition about the relationships between grant size and collaboration. We would expect there would be less collaboration on small grants and more on larger grants. The visualization tells us instead that medium-sized grants seem to attract the least collaborations, and this observation is fairly consistent over the 4 years. Considering that there were 972 small grants awarded in this period compared with 159 medium grants and 133 large grants (shown in the shelf labels), however, the sheer number of small grants might just be the main reason that increases the chance of cross-institution collaborations. Upon closer examination, we can see that grant size does also play a part in shaping the structure of collaboration networks. For small grants, two-organization collaboration is typical, while for large grants, such collaboration patterns are much less common. In particular, there is a high level of collaboration occurring in large grants awarded in 2003.

To investigate further, we right click in the 2003- large grant cell and choose “Analyze in detail” to

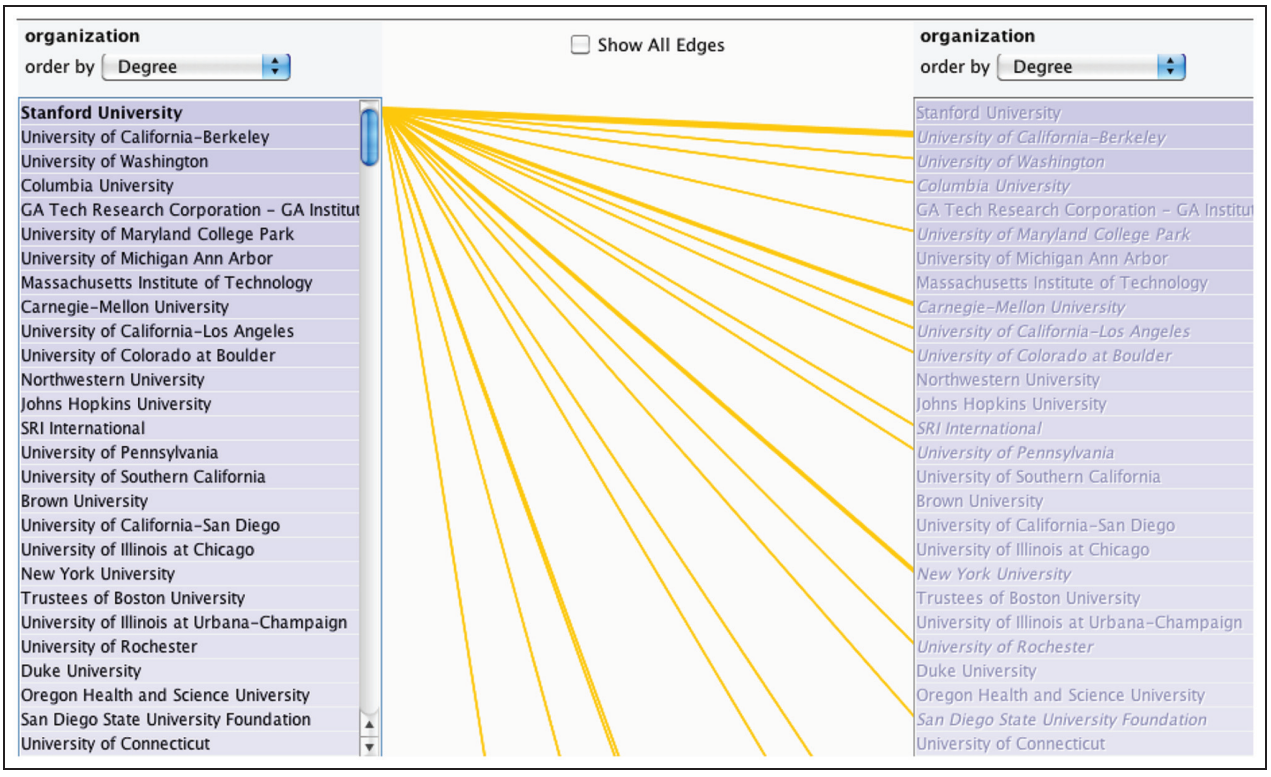


Figure 13. Collaboration between organizations on NSF IIS grants, 2000–2003, in a list representation.
NSF: National Science Foundation; IIS: Information & Intelligent Systems.

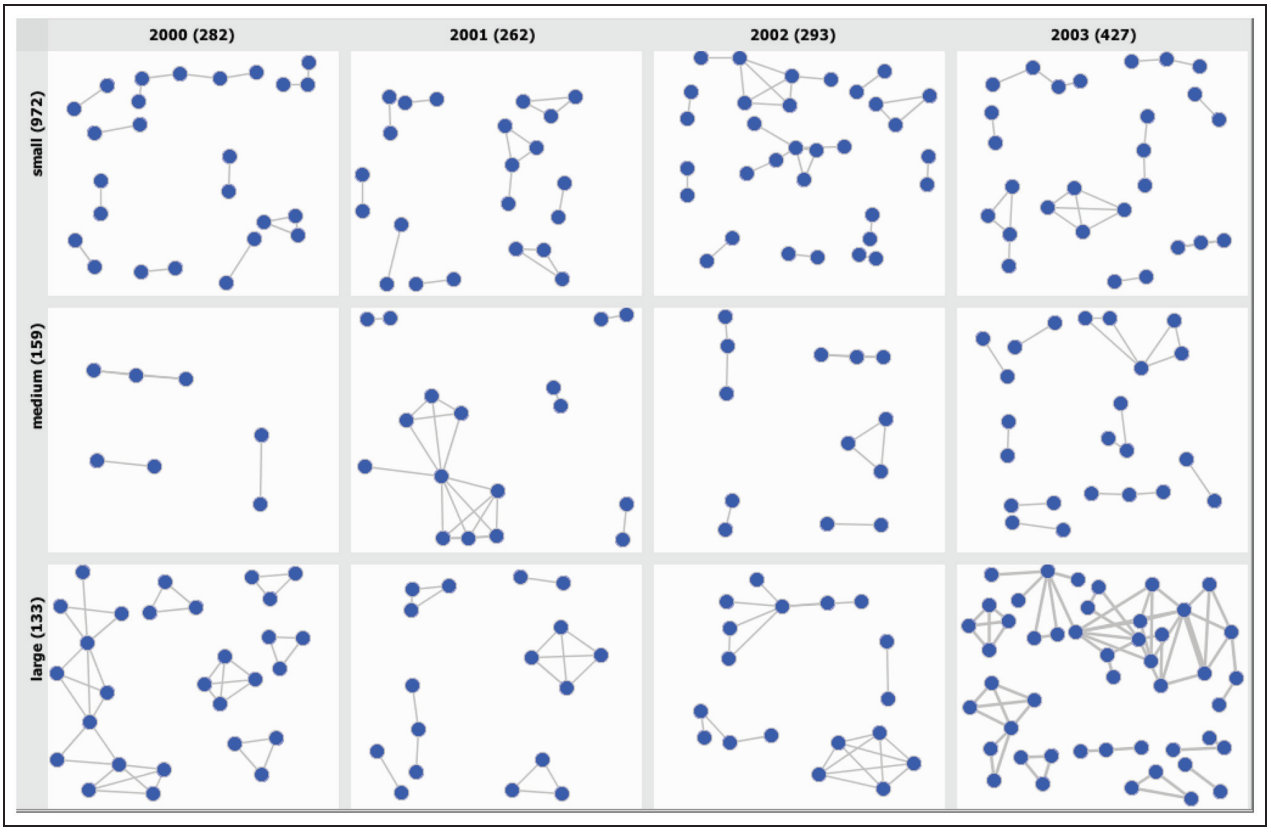


Figure 14. Collaboration between organizations on NSF IIS grants, broken down by year and amount.
NSF: National Science Foundation; IIS: Information & Intelligent Systems.

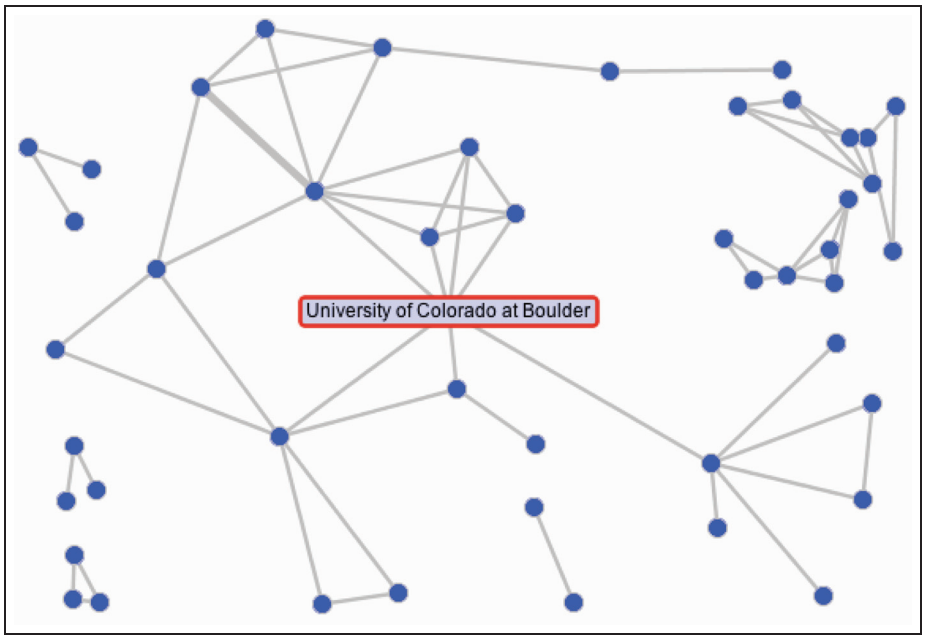


Figure 15. CU Boulder is an important actor in the 2003–large grant collaboration network
CU: University of Colorado.

open a new tab showing that subnetwork for closer analysis. We can see that University of Colorado at Boulder (CU Boulder for short) occupies an important position in this subnetwork where it connects multiple local clusters (Figure 15). This observation is confirmed after running the computational analysis, where CU Boulder has the highest betweenness centrality score, indicating that it is linking many organizations that are otherwise not linked. One reason for this is that CU Boulder has collaborated on quite a few different large grants with different organizations in 2003. To see the grants, it has received as well as the collaborating institutions for each grant, we clear the current subnetwork while keeping the 2003-large grant slice specification and construct an organization-name-title network, connecting organizations with the researchers who are connected with the grants they receive. We see the specific researchers from this school as well as the three large grants they have worked on: emotion in speech, tangible media, and semantic interpretation (Figure 16).

To look further at the role of program managers in the collaboration dynamics, we now go back to the previous tab and replace the date slices with program manager slices. Noting that William Bainbridge, Maria Zemankova, and Ephraim Glinert are the top 3 grant awarding managers, we find that a significant portion of their grants is small grants. After filtering out noncollaborating institutions, we find that grants awarded by them do not particularly show greater activities of collaboration (Figure 17). It is also obvious from the visualization that Ephraim Glinert has awarded a number of grants to groups of four institutions (visualized in the form of tetrahedra), and Stephen Griffin awarded one grant to a group of five collaborating institutions (in the form of a pentahedron). Such patterns, some of which are highlighted in Figure 17, are not seen in grants awarded by other program managers (including those hidden from the current view and have to be revealed by scrolling).

Extending to one-mode networks

One-mode networks as reflexive relational tables

The discussion and scenario so far focus on modeling and visualizing multimodal networks from tabular data. In these tabular datasets (see Tables 1–3), the relationship types among the entity types are *binary*, that is, the relationships are defined between two different classes of entities. Using the projection operator provided in Ploceus, we can create one-mode networks from multimodal networks. The system, however, did not initially provide direct support for modeling networks from tabular data that contain a *unary* relationship, defining references within one class of entities. In the ER data model, such a relationship is called a reflexive or recursive relationship.[2]

Table 5 shows a sample reflexive relational dataset of an egocentric social network of the user "jsmith" on Twitter. Table 5(a) records information about each

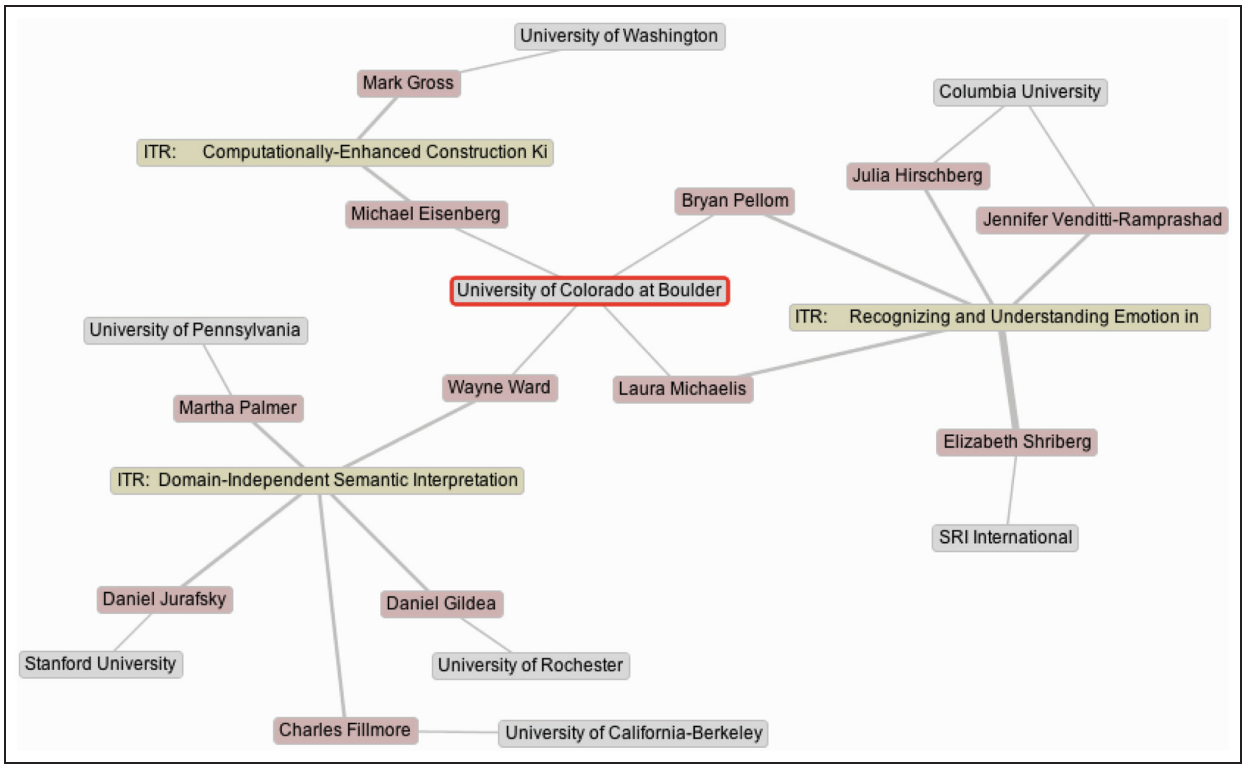


Figure 16. Large grants received by CU Boulder and other institutions in conjunction in 2003.

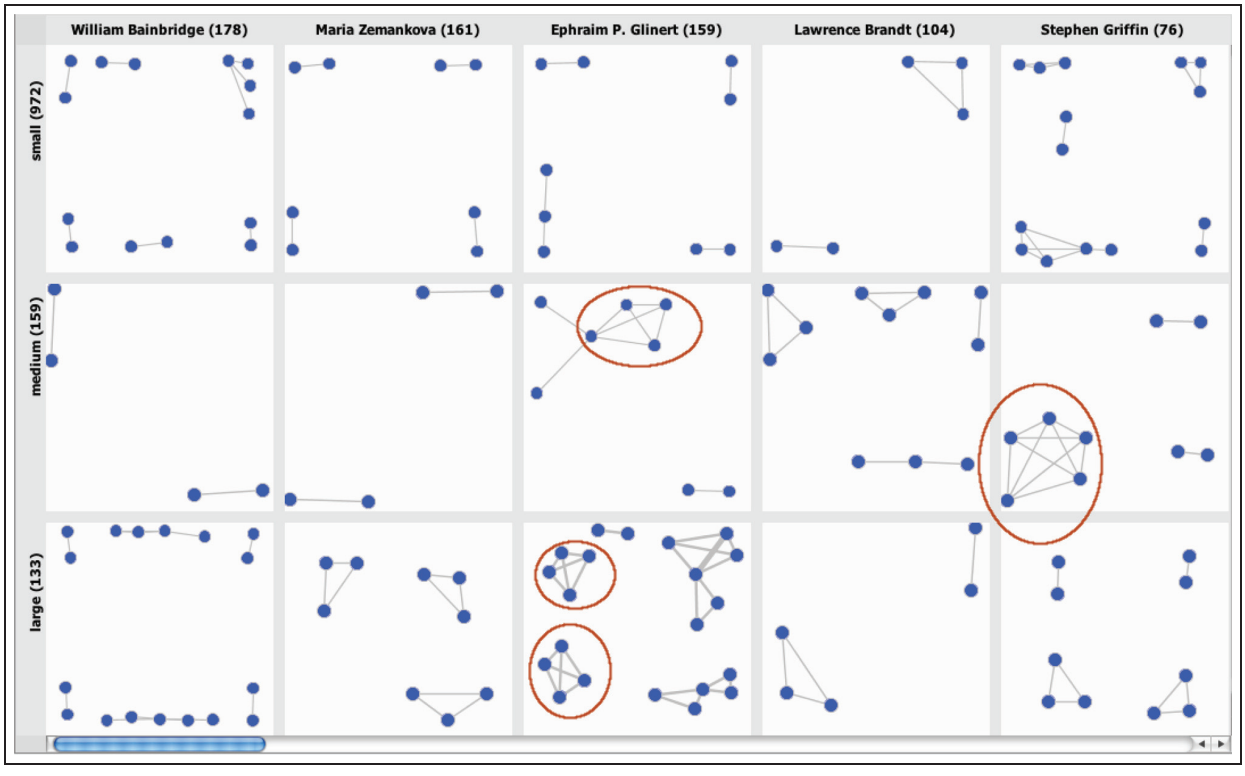


Figure 17. Collaboration between organizations on NSF IIS grants, broken down by program manager and amount. Cliques with more than three nodes are highlighted.

Table 5. Two tables describing relationships between individuals on Twitter.

(a) Person

ID	Join_Date	Favorites	Tweets	Location
jsmith	22 Feb 08	2	24	Silicon Valley
fwong	4 Apr 08	20	231	West Lafayette
jwallace	18 Nov 09	6	120	Finland
ajabbar	25 Jun 10	30	15	Paris, France
suzuki	28 May 09	9	567	San Francisco, mostly

(b) Relationship

Source	*Target*	*Relationship*	*Relationship_date*
jsmith	ajabbar	Following	11 Jan 11
fwong	jsmith	Followed	16 Jul 11
jwallace	jsmith	Mention	1 Nov 11
ajabbar	jsmith	Followed	2 Sep 10
jsmith	fwong	Mention	5 Feb 10

Twitter user including the account (ID), the date when the user joined Twitter (Join_Date), the number of tweets designated as favorites by the user (Favorites), the number of tweets by the user (Tweets), and the self-described location (Location). Table 5(b) records the relationships between the Twitter users. We can consider this dataset to be a *one-mode directed* network. Such data are pervasive given the proliferation of social network sites and social media, and Ploceus should provide reasonable means to incorporate these datasets.

Design considerations for modeling and visualization

Prior study such as PivotGraph[29] enables analysts to perform attribute-based node aggregation in a one-mode network through combo boxes. Since Ploceus provides more operations with greater flexibility, simple user interface controls may not suffice. We thus focus on extending the current interface design to support one-mode networks.

Incorporating one-mode networks into Ploceus' interface and interaction framework turns out to be relatively straightforward. Following the convention of representing one-mode networks as separate node and edge tables,[50] the data management view of Ploceus shows individual columns of the node and edge tables of a one-mode network, respectively (Figure 18). To provide a consistent experience in modeling both multimodal and one-mode networks, we continue utilizing the direct manipulation paradigm. To construct a Twitter network, for example, analysts follow similar steps to those outlined in section "Design of direct manipulation interface." They first drag and drop the ID column to the network schema view, and this operation adds all the Twitter users in the dataset to the network. In order to create connections, analysts must drag and drop the ID column again to the network schema view to create a dummy node. Ploceus recognizes that these two ID nodes in the schema view come from the same table column, thus treating them as the same type and assigning the same color. Finally, analysts click on one of the ID nodes in the schema view and then drag and release the mouse button on the other ID node to create edges. Since this is a directed network, Ploceus supports creating directed edges when analysts hold down the "ctrl" key while using the mouse to connect nodes. Ploceus infers the condition to join the node and edge tables and creates connections accordingly.

Potentially, there is an alternate way to model the same network. Instead of adding ID twice to the network schema view, analysts can drag and drop the Source and Target columns to the schema view, respectively, and connect these two columns. Since Source and Target are distinct columns, Ploceus will treat the nodes created from these two columns as having different types, which is counterintuitive. Furthermore, the source and target columns in edge tables are often numerical identifiers that refer to individual entities in node tables. Node labels created from these columns therefore are not intelligible to analysts. Based on these considerations, we decide not to pursue this approach.

Considering these factors, we make the following two design decisions. First, if analysts provide the one-mode data by importing files formatted for graph data (e.g. GraphML[51] and Pajek[10]), Ploceus parses the data and populates the node and edge tables. During this importing process, Ploceus marks the Source and Target columns in the edge table as hidden, and these two columns are absent in the data management view to prevent analysts from this kind of modeling strategy. Second, if analysts provide the one-mode data by

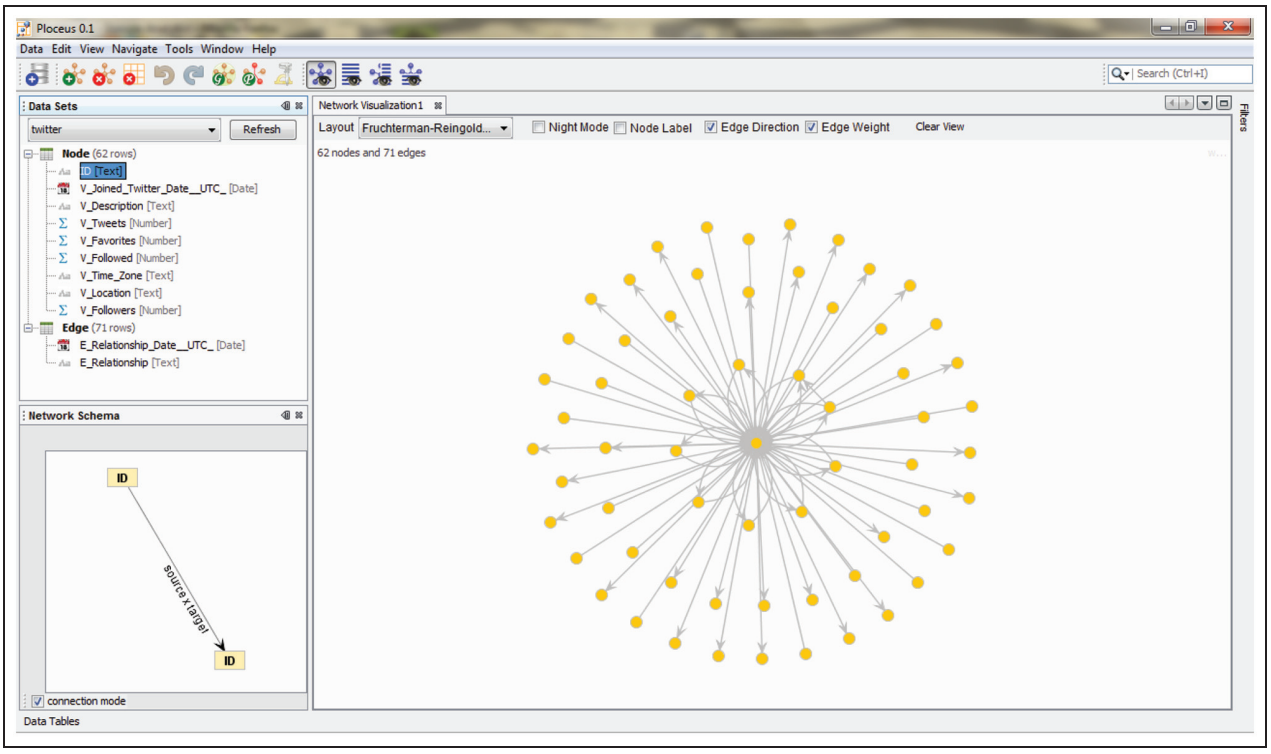


Figure 18. Using Ploceus to construct a one-mode directed Twitter network.
The arrows indicate the directions of "following" on Twitter.

pointing Ploceus to an existing relational database comprising multiple tables, it is a much more difficult inferencing problem to identify the Source and Target columns. In this case, Ploceus allows analysts to aggregate two or more different types of nodes under a self-defined node type. Similar to the default aggregation discussed in section "Operations," this "aggregate type" operation merges nodes if they share identical labels and attribute values even when these nodes are of different types.

Figure 18 shows a resulting visualization of one of the authors' own Twitter network. Ploceus represents the direction of the edges using arrows. If two nodes are connected to each other in both directions, the two edges will overlap and potentially cause confusion. Ploceus thus renders these kinds of edges as quadratic Bézier parametric curves, so that bidirectional edges do not overlap and form a distinctive visual pattern (Figure 18).

All the network operations such as adding attributes and slicing 'n dicing still also apply for one-mode directed networks. Figure 19 shows three egocentric subnetworks generated by slicing 'n dicing the network in Figure 18 using the (Relationship) dimension in the edge table: a "follower" network, a "following" network, and a "mention" network. In addition, the people nodes have an attribute "Tweets," representing the number of tweets by each user, encoded as node size.

Computing connections

Ploceus is powered by the implementation of a formal framework that systematically specifies how to compute edge connections and assign edge weights. In this section, we provide an overview of the framework for readers interested in the implementation details.

Approach and assumptions

Analysts that organize data into structured rows and columns in tables are implicitly declaring relationships between data elements. When data elements appear in the same column, they usually belong to the same type (e.g. both 142 and 16 are GroupSize in Table 1). When data elements appear in the same row, they are usually semantically related, and the specific semantics depend on the context. When Aarnio, Alicia, and OEOB appear in a single row of the White House visit logs, this co-occurrence can be interpreted as a visiting relationship between two entities: the person Alicia Aarnio visited the OEOB. When Data Mining of Digital Behavior and

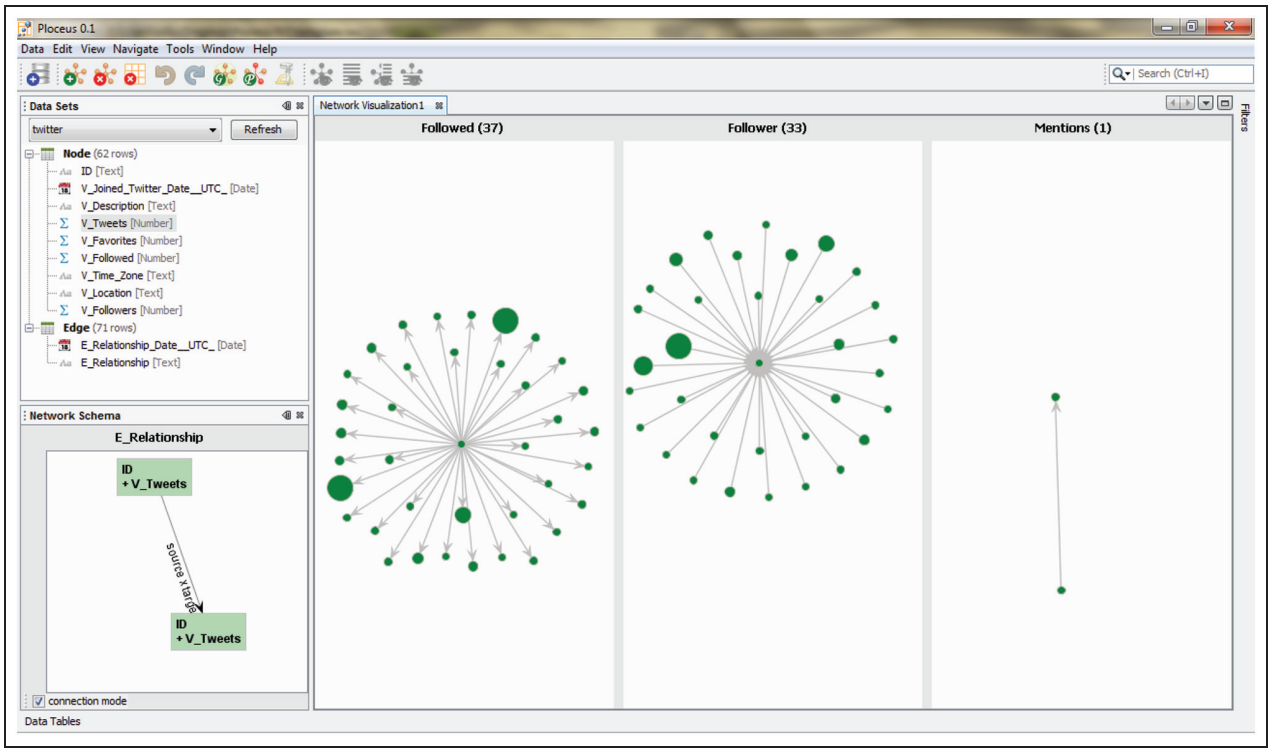


Figure 19. Slicing 'n dicing the twitter network using the (Relationship) dimension and encoding node size as the number of tweets.

2241750 appear in the same row of the NSF grant data, this co-occurrence can be interpreted as a description of an entity in terms of an attribute: the amount of the grant titled "Data Mining of Digital Behavior" is$2241750.

Our approach leverages this simple observation that *the meaning of row-based co-occurrence is context sensitive*. It is thus possible to propose a co-occurrence-based formal framework, which specifies the construction and transformation of networks, where the meaning of the graphs created will be subject to users' interpretation. Co-occurrence is undirected: when A co-occurs in a row with B, B also co-occurs with A.

We base our formal framework on the relational model[52] used widely in database theories, with basic relational algebraic operators such as selection (σ), projection (π), join ($\bowtie$), and aggregation ($\mathscr{F}$).[2] We make the following three assumptions:

- Each row in a table has a unique identifier;
- Each value in the table cells is *atomic*, that is, the value can be classified as nominal, quantitative, and ordinal, and the value cannot be decomposed into meaningful smaller units;
- We only focus on creating networks in which there are no edges connecting one node to itself.

First-order graphs

The entire formal framework is built on the fundamental notion of a *first-order graph* and transformative operations on the graph. First-order graphs are the simplest graphs or networks we can construct where each node and edge is constructed from one (1) single row only. In relational model terms, a row is a *tuple*, where one or more cell values in that row form a *subtuple*, and a table is a *relation*. When all the data needed for graph construction are present in a single table, for any given row in a table, there are two main ways to construct a node from it. We can create a node such that its label is a subtuple (e.g. the node label is "Smith, John," or a function of a subtuple (e.g. taking Size as the argument and returning "large group" as the node label if the group size is above 50, and "small group" otherwise). In a similar way, we can assign an attribute to a node based on a subtuple or a function of subtuple.

It is thus a basic idea that in the translation from a table to a graph, if the construction of a node results from only a single row of the table, the node is a *first-order* node. Two first-order nodes can have the same labels and attributes, as there may be rows containing identical values for selected table dimensions. First-order nodes are created using the relational projection operator.

Here, we introduce two important concepts, *locale* of a node and *basis* of an edge, in order to compute connections consistently when multiple tables and graph transformation are involved. The locale of a node refers to the set of tuples from which the node is constructed; the basis of an edge refers to the set of relational elements (tuples or graph nodes), which are jointly shared by the locales of two nodes. In actual implementation, the comparison of locale and computing of edges are realized using the relational selection and projection operators. The *weight* of an edge will be the *cardinality of its basis*.

In first-order graphs, for example, the locale of a node will just contain one element, which is the tuple from which the node is created. As mentioned earlier, our formalism focuses on establishing relationships based on co-occurrence in rows. Two first-order nodes are thus connected if they share the same locale. Formally speaking

$$\text{If } \exists t, locale(n_1) = locale(n_2) = t, \quad \text{then } e(n_1, n_2)$$

Our framework considers two possible cases when first-order nodes and edges are constructed from multiple tables. First, we can create two sets of first-order nodes, each constructed from a single table only, and the edges between the nodes are created by linking two tables. The notion of co-occurrence is then no longer limited to one tuple in a relation but is extended to include two or more tuples in multiple relations through a join condition specified by the analyst. Formally

$$\begin{aligned}&\text{Given } locale(n_1) = R_1 \cdot t_i \wedge locale(n_2) = R_2 \cdot t_j\\ &\text{If } (R_1 \cdot t_i \cup R_2 \cdot t_j) \in (R_1 \bowtie_\theta R_2), \text{then } e(n_1, n_2)\\ &basis(n_1, n_2) = \{(R_1 \cdot t_i, R_2 \cdot t_j)\}\end{aligned}$$

In the second and more complex case, a set of first-order nodes can be constructed such that their labels come from one table, and their attributes come from another table. We do not allow constructing node labels from multiple relations in our formalism for the purpose of simplicity. The type of join used here in constructing first-order nodes will be a left-outer-join[2] because we want to preserve all the node labels even when there are no matching attributes. The locale of the nodes is determined by the table from which the labels are constructed only.

Higher-order graphs: transformation

First-order graphs often are not at the right level of abstraction intended for exploration and analysis. For example, there may be nodes with identical labels that refer to the same entity. In section "Operations," we introduced three transformative operations: aggregation, projection, and edge weighting. We also mentioned that Ploceus aggregates nodes by labels and attributes automatically. Our formal framework specifies how these transformations affect the edges based on the notion of a locale introduced in the previous section. In aggregation, for example, assuming that the analysts have specified a function of aggregating nodes, the newly produced nodes will inherit the locales of the nodes being aggregated

$$locale(n') = locale(n_1) \cup \ldots \cup locale(n_j)$$

Two new nodes will be connected if the intersection of their locales is not empty

$$basis(n'_1, n'_2) = locale(n'_1) \cap locale(n'_2) \neq \emptyset$$

For projection on a two-mode graph with two types of nodes N and M, for example, two nodes $n_1, n_2 \in N$ are connected if they have at least one neighbor in common in M

$$\exists m \in M, e(n_1, m) \in E \text{ and } e(n_2, m) \in E \Rightarrow e(n_1, n_2)$$

According to this definition, the basis of an edge is no longer a set of tuples, but a set of nodes

$$basis(n_1, n_2) = \{m \in M | e(n_1, m) \in E \text{ and } e(n_2, m) \in E\}$$

Slicing and dicing are operations at a global level using dimensions that are orthogonal to those used in network construction. In our framework, the dimensions used in slicing and dicing serve as query conditions when nodes and edges are created through relational selection and projection operators.

Extending to one-mode graphs

We initially developed the formal framework without giving serious consideration to the possibility of extending to one-mode graphs. Scenarios discussed in section "Extending to one-mode networks" make a compelling case to broaden the scope of our framework for the construction of directed graphs from data tables describing reflexive relationships. Section "Extending to one-mode networks" presents our design rationale at the interface level; in this section, we briefly discuss the underlying theoretical logic.

Suppose we want to construct a graph showing the Twitter relationships between different users from Table 5. We first create two identical sets of first-order Person nodes from Table 5(a), called N and N_1. To construct connections between these two sets of

first-order nodes, we follow the definition discussed in section "First-order graphs"

$$\text{A node } n \in N \text{ is connected to a node } n_1 \in N_1$$
$$\text{if } locale(n) + locale(n_1) \in R_P \bowtie R_P \text{ on the condition that}$$
$$R_P \cdot \text{ID} = R_E \cdot \text{Source} \quad \text{and} \quad R_P \cdot \text{ID} = R_E \cdot \text{Target}$$

where R_P represents the person table and R_E represents the relationship table

Note that here we make a slight modification by replacing the union operator $\cup$ with the concatenation operator $+$ in $locale(n) + locale(n_1)$. Concatenation preserves the order of the values in the tuples and does not remove duplicates, thus ensuring that the order of tuple values is preserved in edge creation. We can then assign directions to edges by specifying n nodes as source nodes and n_1 nodes as target nodes. Finally, we can do an aggregation of n and n_1 nodes if they share the same label.

Implementation

Ploceus is built entirely in Java on the NetBeans Rich Client Platform.[53] It utilizes two major external toolkits and libraries: H2[54] as the underlying database for relational algebraic queries and JUNG[25] as the graph visualization and computational metrics library. All the operations supported by Ploceus are performed in real time. Simple operations such as adding nodes and creating connections are realized through Structured Query Language (SQL) queries and are scalable for up to tens of thousands of rows without significant delay. More complex operations such as projection and statistical metrics computation are more computationally expensive, and the performance can be affected with large datasets. Every subnetwork in slicing and dicing is created through a separate thread, and the performance bottleneck is at the concurrent handling of SQL queries by the underlying H2 database. Future study includes optimization of the implementation of operations.

Outstanding problems and limitations

Joining multiple tables

When computing connections between columns from different tables, we currently infer equi-join conditions by analyzing foreign key constraints between tables through a Dijkstra shortest-path algorithm.[39] We first construct a graph where the nodes are the columns in each table, and primary key columns and foreign key columns are connected. Given two columns to be connected, we then apply the shortest-path algorithm on this graph.

When a database becomes more complex in terms of ER modeling, there might be multiple reasonable equi-join conditions. It is thus the user's decision to choose the appropriate join condition. Currently, Ploceus handles this situation through a dialog, letting users interactively add relevant tables and connect the primary keys and foreign keys to specify the desired join condition (Figure 20). Related systems such as Tableau[22] take a similar approach. Tableau does not support interactive table joining during the process of exploration. Instead, at the stage of data import, analysts need to explicitly join tables to include all the data columns necessary for exploration. Tableau supports a richer set of join types and conditions (Figure 21).

Potentially, we can also use more sophisticated techniques to automatically compute a number of different join conditions and to rank them by inferring analysts' intention. The diversity of all the possible join conditions, however, can hardly be fully captured. More importantly, all these approaches do not address a fundamental issue satisfactorily; analysts must have a precise and good understanding of the concepts of relational join, primary key, and foreign key. Even if they understand these concepts well, it is still nontrivial to interpret the semantics of edges constructed as a result of joining multiple tables.

Currently, we do not have a satisfactory solution to this problem, and we doubt there will be one if the underlying data model is going to be relational. The recent emerging NoSQL databases[55] might provide an interesting angle to address this issue. Multiple related tables in relational databases are a direct consequence of the design choice to normalize data tables for the sake of minimizing redundant data representation and avoiding anomalies in data modification.[2] These goals are not emphasized in NoSQL databases. It would be interesting to explore if the problem of mandatory join specification can then be eliminated if we choose NoSQL data model (key-value pairs instead of data tables) where the joins have been done a priori, in some sense eliminating to do them explicitly.

Big/dense graphs

Scalability is an important issue in Information Visualization and Visual Analytics research. An implication of the network modeling power provided by Ploceus is that analysts can easily create both big graphs with thousands or even millions of nodes and dense graphs where the number of edges is close to the maximum possible number of edges. Visualizing and analyzing these graphs remain great research challenges. Due to the limited number of pixels available on computer screens, it is impossible to display all the nodes without resulting in cluttering or overlapping.

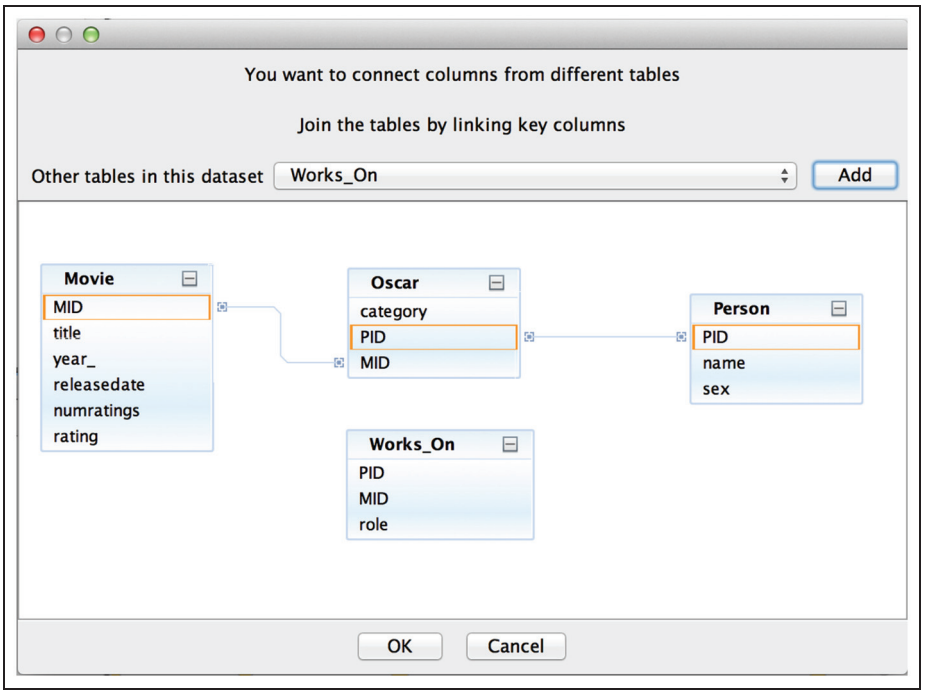


Figure 20. Users interactively add relevant tables and connect the primary keys and foreign keys to specify the desired join condition in Ploceus. In this example dataset about IMDB movies, a person (PID) and a movie (MID) can be connected in two different ways, via the **Works_On** table and the **Oscar** table, respectively.
IMDB: Internet Movie Database.

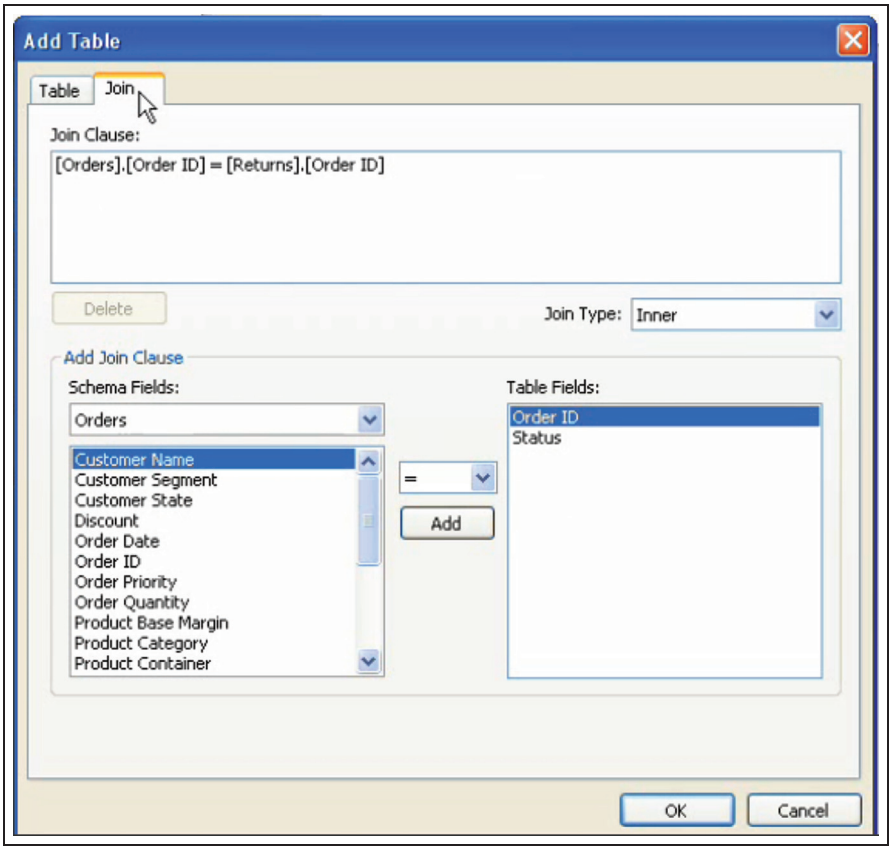


Figure 21. Dialog interface in Tableau to join multiple tables.

Dense edges also cause severe performance issues in computing graph layout. In Ploceus, the problem of scale is handled using subnetwork sampling. The interface displays bold messages to remind analysts that only a subnetwork is shown. Analysts can interactively add or remove subnetworks through search queries.

A few potential directions exist for future research on this problem. First, it is worthwhile to investigate appropriate mechanisms of network sampling. Ploceus currently samples randomly. Techniques such as the degree-of-interest functional retrieval[56] sound more promising. Users can pick a focal point and the system displays a subgraph that is of maximal interest. Second, we would like to investigate when it is actually useful to show an overview of the entire network and to articulate the user tasks involved in these situations. It may be possible that we can design alternative visual representations that provide information needed in these tasks without having to show every single node and edge. Third, one way to analyze big graphs is to use a divide and conquer strategy by breaking the graphs down into meaningful subgraphs. Filtering and slicing 'n dicing are two reasonable mechanisms to do so, and they are included in Ploceus. It may be necessary to analyze and compare multiple networks at the same time. Ploceus now organizes multiple networks in the form of a matrix, but there are potential readability and usability problems when each of these networks contains a large number of nodes and edges. While systems such as ManyNets[28] have taken a first step in the effort of facilitating visual analysis of multiple networks, more research is needed to understand and design for multiple network analysis.

Expressive power

In working with sample datasets, we have already identified situations that point to potential limitations of the current framework. For example, if we want to create a network where two organizations are connected if they have collaborated on more than two grants within the past 5 years, the set of operations described in section "Operations" is not sufficient to express such semantics of conditional connectivity.

Further study is required to understand the expressive power and limitations of this framework. Relational algebra, an established framework, is proven to be equivalent to first-order logic, and the expressive power of first-order logic is well understood.[2] In relational algebra, a set of primitive operators serves as building blocks for more complex operators. Since we are investigating a new domain here, it remains to be seen if the set of operations can serve as primitives for graph construction and if any additional operations need to be included for completeness.

User evaluation

To understand the implications of the algebraic framework and the interface design, we conducted an evaluation of the learnability and usability of Ploceus. We recruited 10 participants, including one undergraduate student; five graduate students in the areas of computer science, communication design, and ergonomics; and four working professionals in the areas of software engineering, program managing, and electrical engineering. Eight of them were knowledgeable of database technologies and SQL queries, and the other two did not have relevant experience in these technologies. All of them had never seen or used Ploceus. The goal of evaluation was to identify qualitative insights about the way people think about network visualization construction and how well Ploceus supports network modeling.

Tasks and procedures

We gave a brief introduction to Ploceus, demonstrated the main functionalities of the system and showed the participants how to construct different networks using the White House Visitor dataset (Table 1 shows sample rows). We then asked them to create visualizations and answer the following questions on the NSF dataset (Table 3 shows sample rows):

- Can you create a visualization showing the collaboration pattern between organizations on research grants?
- Which organization(s) tend to collaborate the most?
- In which year(s) do we see the most cross-organization collaboration?

We explicitly told the participants that while there were three tables in the NSF dataset, they could connect any two columns in the data management view (Figure 2) as if these columns were from the same table; Ploceus could infer the right join condition in such simple datasets.

To answer the given questions, the participants needed to create visualizations in multiple steps as outlined in the scenario in section "Scenario: analyzing cross-institution research efforts": first, add the grant IDs and the organizations as nodes; then connect these two types of nodes; perform a projection on the grant IDs so that the organizations are connected directly to each other via common grant IDs, and finally slice and dice the network using the date dimension to break down the network by year.

We asked the participants to think aloud, observed their interaction with the system, recorded their interaction history as hand-written notes, and sought their impressions and comments on the system after they completed the tasks. Our aim was to gain an understanding of how difficult the system was to learn and use, if there were any problematic design issues, and how we might be able to address the difficulties experienced by the participants.

Results and analysis

Out of the 10 participants, four did not experience any difficulty and quickly answered the three questions accurately; five participants did experiment a few times before completing the tasks successfully. Only one participant failed to discover the correct strategy within half an hour.

In addition to task performance, we were interested in understanding the participants' strategies and thinking in action in visualization construction and exploration processes. Inspired by Norman's seven stages of action,[57] we hypothesized a model of expected action sequences in the process of visualization construction and exploration (Figure 22(a)). We anticipated that using Ploceus to construct network visualizations would involve the following steps:

1. *Interpret task/form intention.* When a task is assigned, users need to be able to understand what kind of insight is being inquired; when the task is self-initiated, users form an intention to look at certain aspects of the data.
2. *Simulate appropriate visualizations.* Users mentally simulate and visualize possible representations that can provide the desired insight.
3. *Identify relevant data dimensions.* Only a few selected columns in the dataset are relevant for a given question, users need to identify which data dimensions to include in the visualization.
4. *Conceive construction strategy.* Given a desired visualization, choose appropriate operations in the Ploceus framework and understand how these operations should be combined to create the visualization.
5. *Execute construction sequence.* Users execute the conceived construction strategy and perform each operation using the Ploceus interface.
6. *Interpret visualizations.* Finally, assuming that the construction is successful, users must be able to accurately interpret the visualizations generated and provide answers to the given questions.

We used this model as a framework to analyze and interpret participants' interaction histories with Ploceus, while keeping in mind that the model needed to be validated and refined based on empirical evidence. The model was useful for deductive analysis and helped us pinpoint difficulties the participants experienced at various stages of action; on the other hand, we also identified evidence pointing to possible refinements of the model.

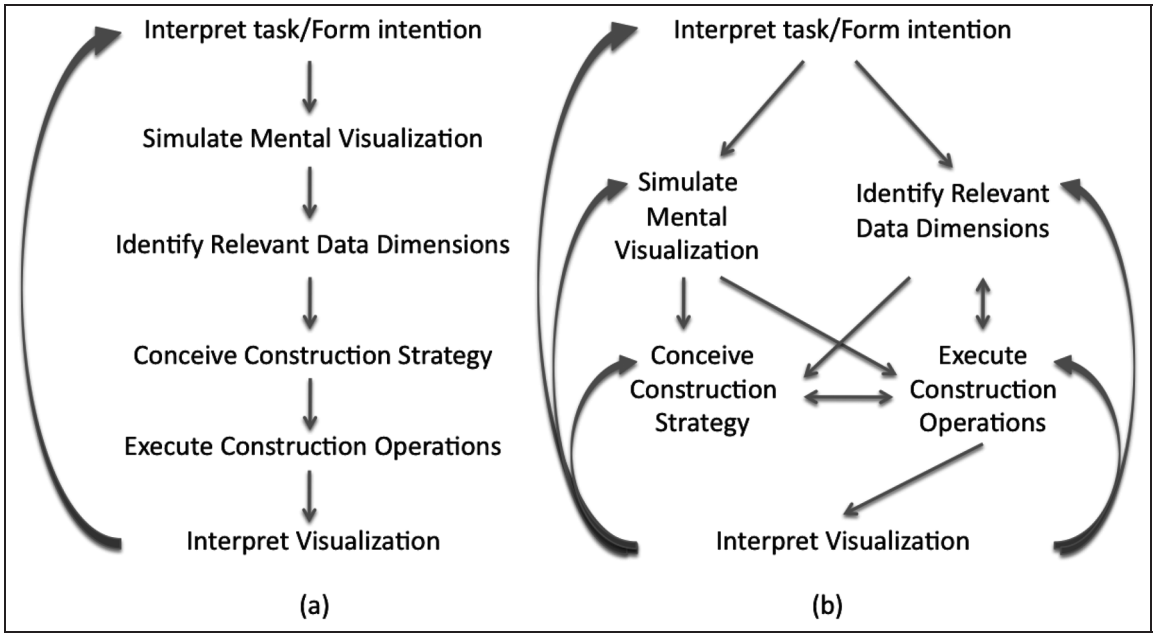


Figure 22. Models of user interaction with Ploceus for visualization construction and analysis: Model (a) represents the hypothesized model and model (b) represents the revised version based on qualitative data.

Difficulties in stages of action. One participant (P3) appeared to have understood the features of Ploceus after the demo session and quickly constructed a visualization after we gave her the questions. Interestingly, she created a network connecting researchers with organizations and then did a projection on the organizations. When we asked her to describe what the visualization was showing, she realized that she was not creating a visualization that would answer the questions. She then created a different network in which researchers are connected to each other if they come from the same organizations, and again, this visualization was not able to answer the given questions. She finally realized what went wrong and commented that "I totally mis-interpreted the question." After careful consideration of the semantics of the questions, she successfully created the intended visualization. We interpret from this process that she had difficulty *interpreting what questions were being asked*.

Some users showed signs of a lack of ability to *mentally visualize the representation* that would answer a given question. One participant (P4) successfully created a one-mode network of collaborating organizations. To answer the question, "In which year(s) do we see the most cross-organization collaboration?," he added the Date column values as nodes to the visualization and connected the dates with the organizations. He tried hard to answer the question by examining the visualization, but did not have any viable lead.

In general, identifying the relevant data dimensions was not a major difficulty for most of the users. For example, to answer Q1 ("create a visualization showing the collaboration pattern between organizations on research grants"), all participants were aware that organizations and grants needed to be part of the visualization. Three participants (P4, P6, and P7), however, were uncertain if the name column in the Person table was relevant. All of them talked about their logic in the same way: organizations collaborated if researchers from these organizations had received grants together. In this sense, researchers were crucial components in the visualization. As a result, they tried to connect researchers with grants and to connect grants with organizations. This approach gave them no viable lead. After P7 finally constructed the correct visualization through multiple trials, he commented, "I didn't realize the system is so powerful that you can directly connect organizations with grants together!"

Even if users know clearly what variables are important, they may still experience potential difficulties to conceive the appropriate steps to construct the visualization. Previous studies have shown that visualization construction is a major hurdle for novice users.[58] In observing the participants, we got similar impressions. P4 wanted to create a one-mode network by projection, but he forgot to create the connections between the two classes of nodes first, and the projection could not be performed. In this case, he could not *conceive a proper sequence of the operations* to construct the visualization.

Most of the users did not have any difficulties in translating an intention of performing an operation to an actual action. That is, they were able to pinpoint the user interface component designed to support a specific operation. One participant (P4), however, could not figure out how to perform slicing 'n dicing, even though we demonstrated this functionality for him. He commented, "Although I've seen it, I just completely forgot about it." We acknowledge that the demonstration session was relatively short, and viewing a demonstration is very different from performing the same action oneself. However, the participant was confident that he would perform much better if he had experimented with Ploceus for a longer period of time.

All users were familiar with node-link diagrams and had no difficulty in reading the visualizations generated. The major difficulty surfaced when the size of the generated network became too large. The layout algorithms included in Ploceus could be improved. The force-directed layout, for example, was slow to stabilize for a network with more than 500 nodes, and it did not show clusters inherent in a network clearly. Disconnected subnetworks tended to be pushed to the boundaries of the visualization, making the graph less readable.

Where data do not fit model: trial-and-error strategy. While the preconceived action model anchored our interpretation of user difficulties in using Ploceus, upon analyzing user interaction history, we realized that the model might be an idealized sequence of visualization construction and exploration. In many cases, the participants' actions did not adhere to the presumed action sequence. The actual construction processes were quite ad hoc.

For example, while we did find evidence that could be interpreted as users mentally visualizing their desired representations (P7, for example, reported that he wanted to construct a visualization shown in the demo session, where a two-mode network was projected into a one-mode network), some participants did not mentally visualize the appropriate visualizations before constructing these visualizations. Instead, they talked about how they would "try it out" when they decided to perform an operation and commented they were not sure what the resulting visualizations would look like. This strategy could be due to unfamiliarity with Ploceus' working mechanism but could

also be an effort of cognitive offloading. Rather than planning everything in the head, it was easier to put a thought into action. As P9 mentioned, "let me just try to see if this works, if not, I'll just start over."

As a consequence of adopting the trial-and-error strategy, the participants iterated between the stages in a nonlinear fashion, as shown in Figure 22(b). Participant P6, who failed to create appropriate visualizations within the given time frame, largely iterated in the loop of "interpret question/form intention → identify data dimensions ↔ execute construction operations ↔ interpret visualizations." A major cause of failure, as we interpreted, was that he was not able to mentally visualize an appropriate visualization and could not conceive a proper strategy to implement the visualization. As a result, he purely relied on randomly picking the operations.

Role of database knowledge

We speculated before the study that participants with background knowledge in database technologies, especially SQL queries, would understand the system better and perform the tasks with less difficulty. Qualitative evidence showed, however, this assumption was too simplistic.

Among the four participants who encountered no difficulty in completing the tasks, two had taken database classes, and the other two had no knowledge of SQL. Participant P8 was a software engineer and used SQL in his projects; yet, he spent a significant amount of time trying to understand the interface. P8 kept talking about operations he wanted to perform in terms of database concepts (e.g. "How do I do a GROUP BY?"). A major difficulty for him, as he described it, was to map the various SQL queries to the interactive operations supported in Ploceus. He even suggested that he would like to see a window showing the corresponding SQL queries whenever he interacted with the interface. It was also interesting to note that in answering Q3 ("In which year do we see the most collaboration between organizations"), although he observed that there were obviously more links in the "2003" subnetwork, he was not confident in giving a definite answer and commented that he would like to see a precise numeric value to confirm the answer. For a database expert like P8, whose thinking was deeply rooted in SQL queries, the learning curve might be higher than nonexpert users.

General impression and comments

All the participants liked Ploceus, especially the interface design. Although some participants considered constructing network visualizations nontrivial, they still agreed that the interface was consistent and the learning curve was not too high. The affordance of the network schema view was clear to all the participants, and all of them strongly favored this view.

The participants also identified features in Ploceus that they had difficulties in understanding and interpreting. More than one participant mentioned that the affordances of slicing 'n dicing shelves were not immediately clear to first-time users, but they acknowledged that after some interaction the design began to make sense for them. One participant also mentioned that the meaning of the numerical value following each slicing 'n dicing value was not clear. For example, when we slice 'n dice an organization collaboration network by year, a slice will be displayed as "2000 (282)," as shown in Figure 14, indicating the number of grants given in the year 2000.

One participant was so interested in Ploceus that he asked for some extra time after the given tasks to perform some open-ended exploration. One of the questions he wanted to know was who got the most money from NSF. He tried to create a network connecting researchers with amounts and then tried to order the amount nodes by value in the list view. While this is certainly one way to answer the question, a bar chart might be a more effective visual representation than a network visualization. In this regard, the user should have picked systems such as Tableau[22] instead of Ploceus. This observation is consistent with what Kobsa[59] has noted in his evaluation study of early InfoVis systems: users tend to use the default system or setting given to them, and it is difficult for them to initialize a change in the mindset to explore alternative representations or system settings.

Conclusion and future study

In this article, we have focused on the system design and formal framework aspects of performing network-based visual analysis and argued that our approach provides new capabilities beyond the existing study. Our contributions include the following:

- Drawing from prior study, we present a conceptual framework specifying possible operations for constructing and transforming networks from multivariate tabular data. Most of these operations are meaningful to end users who are not necessarily database experts.
- A specification of the operations based on the relational model and an implementation of the framework in relational algebra.
- The design and implementation of a system based on the framework, which integrates data manipulation with visual exploration processes.

- A discussion of the nature of high-level tasks in network-based visual analysis that may have implications for future study on visual analytics.
- A qualitative evaluation that proposes a model of how users construct and explore network visualizations using Ploceus.

This research lays the foundation for further investigations. First, there are certain features we would like to add, such as pinpointing specific data rows/columns from visualizations that explicitly illustrate the provenance of the data and integrating analytic techniques on data tables such as log-linear modeling on top of the network analysis techniques presented here. It also makes sense to provide a visual representation of users' interaction history. Such a construction history can be useful for nonexpert users to understand the consequences of their actions. As mentioned in section "Expressive power," it is worthwhile to understand the expressive power and limitations of our framework in greater detail and perhaps to examine how this framework can be applied and extended for compound graphs. Since the user evaluation presented here involves controlled tasks for a specific dataset, we would also like to gain further insights on the system's ecological validity in longer-term case studies.

Funding

This study was supported by the National Science Foundation under award IIS-0915788 and the VACCINE Center, a Department of Homeland Security's Center of Excellence in Command, Control and Interoperability.

References

1. Chen PP. The entity-relationship model—toward a unified view of data. *ACM T Database Syst* 1976; 1(1): 9–36.
2. Elmasri R and Navathe S. *Fundamentals of database systems*. 6th ed. Boston, MA: Addison-Wesley, 2011.
3. Chen C, Yan X, Zhu F, et al. Graph OLAP: a multi-dimensional framework for graph data analysis. *Knowl Inf Syst* 2009; 21: 41–63.
4. Bagui S and Earp R. *Database design using entity-relationship diagrams*. 1st ed. Boston, MA: Auerbach Publications, 2003.
5. Amar RA and Stasko JT. Knowledge precepts for design and evaluation of information visualizations. *IEEE T Vis Comput Gr* 2005; 11(4): 432–442.
6. Lee B, Plaisant C, Parr CS, et al. Task taxonomy for graph visualization. In: *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, Venice, Italy, 23 May 2006, pp. 1–5. NY: ACM Press.
7. Shneiderman B and Aris A. Network visualization by semantic substrates. *IEEE T Vis Comput Gr* 2006; 12(5): 733–740.
8. Liu Z and Stasko J. Mental models, visual reasoning and interaction in information visualization: a top-down perspective. *IEEE T Vis Comput Gr* 2010; 16(6): 999–1008.
9. NetMiner—social network analysis software, http://www.netminer.com (2010).
10. Pajek—program for large network analysis, http://pajek.imfm.si/doku.php (2010).
11. UCINET, http://www.analytictech.com/ucinet/ (2010).
12. Adar E. GUESS: a language and interface for graph exploration. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*, Quebec, Canada, 22-27 April 2006.
13. Hansen D, Shneiderman B and Smith MA. *Analyzing social media networks with NodeXL: insights from a connected world*. San Francisco, CA: Morgan Kaufmann, 2010.
14. Henry N, Fekete J and McGuffin MJ. NodeTrix: a hybrid visualization of social networks. *IEEE T Vis Comput Gr* 2007; 13(6): 1302–1309.
15. Perer A and Shneiderman B. Balancing systematic and flexible exploration of social networks. *IEEE T Vis Comput Gr* 2006; 12(5): 693–700.
16. Liu Z, Navathe S and Stasko J. Network-based visual analysis of tabular data. In: *IEEE conference on visual analytics science and technology '11*, Providence, RI, 23-28 October, 2011, pp. 41–50. Piscataway, NJ: IEEE Press.
17. Livny M, Ramakrishnan R, Beyer K, et al. DEVise: integrated querying and visual exploration of large datasets. *SIGMOD Rec* 1997; 26(2): 301–312.
18. Rao R and Card SK. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: *Proceedings of the SIGCHI conference on human factors in computing systems: celebrating interdependence*, Boston, Massachusetts, USA, 24-28 April 1994, pp. 318–322. Boston, MA: ACM.
19. Spenke M, Beilken C and Berlage T. FOCUS: the interactive table for product comparison and selection. In: *Proceedings of the 9th annual ACM symposium on user interface software and technology*, Seattle, WA, USA, 6-8 November 1996, pp. 41–50. New York, NY: ACM Press.
20. Spenke M and Beilken C. InfoZoom: analysing formula one racing results with an interactive data mining and visualisation tool. In: *Proceedings of 2nd international conference on data mining*, Cambridge University, Cambridge, UK, 5-7 July 2000, pp. 455–464. Southampton, UK: WIT Press.
21. Stolte C, Tang D and Hanrahan P. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE T Vis Comput Gr* 2002; 8(1): 52–65.
22. Tableau Software, http://www.tableausoftware.com/ (2010).
23. Stasko J, Görg C and Liu Z. Jigsaw: supporting investigative analysis through interactive visualization. *Inform Visual* 2008; 7(2): 118–132.

24. Heer J, Card SK and Landay JA. Prefuse: a toolkit for interactive information visualization. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Portland, Oregon, 2-7 Apr 2005, pp. 421–430. NY: ACM Press.
25. O'Madadhain J, Fisher D, White S, et al. *JUNG: Java universal network/graph framework*. Technical report UCI-ICS 03-17, 2003.
26. Gephi: an open source graph visualization and manipulation software, http://gephi.org (2010).
27. Auber D. Tulip: a huge graph visualization framework. In: Jünger M and Mutzel P (eds) *Graph drawing software (mathematics and visualization)*. Berlin: Springer-Verlag, 2003, pp. 105–123.
28. Freire M, Plaisant C, Shneiderman B, et al. ManyNets: an interface for multiple network analysis and visualization. In: *Proceedings of the 28th international conference on human factors in computing systems*, Atlanta, GA, 10-15 April 2010, pp. 213–222. NY: ACM Press.
29. Wattenberg M. Visual exploration of multivariate graphs. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Montreal, Quebec, 22-27 April 2006, pp. 811–819. New York: ACM.
30. Guéhis S, Rigaux P and Waller E. Data-driven publication of relational databases. In: *10th international database engineering and applications symposium*, Delhi, India, 11-14 December 2006, pp. 267–272. Piscataway, NJ: IEEE Press.
31. Bohannon P, Korth HF and Narayan PPS. The table and the tree: on-line access to relational data through virtual XML documents. In: *Proceedings of WebDB*, Santa Barbara, California, USA, 24-25 May 2001, pp. 55–60. NY: ACM.
32. Wilkinson L. *The grammar of graphics*. NY: Springer-Verlag, 2005.
33. Weaver C. Multidimensional data dissection using attribute relationship graphs. In: *IEEE symposium on visual analytics science and technology (VAST)*, Salt Lake City, Utah, 24-29 October 2010, pp. 75–82. Piscataway, NJ: IEEE Press.
34. TouchGraph Navigator: graph visualization and social network analysis software, http://touchgraph.com/navigator (2011).
35. Centrifuge systems, http://centrifugesystems.com/ (2011).
36. Heer J and Perer A. Orion: a system for modeling, transformation and visualization of multidimensional heterogeneous networks. In: *IEEE conference on visual analytics science and technology '11*, Providence, RI, 23-28 October, 2011. Piscataway, NJ: IEEE Press.
37. Weaver C. Cross-filtered views for multidimensional visual analysis. *IEEE T Vis Comput Gr* 2010; 16(2): 192–204.
38. North C, Conklin N, Indukuri K, et al. Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Inform Visual* 2002; 1(3–4): 211–228.
39. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math* 1959; 1: 269–271.
40. Latapy M, Magnien C and Vecchio ND. Basic notions for the analysis of large two-mode networks. *Soc Networks* 2008; 30(1): 31–48.
41. Cypher A and Halbert DC. *Watch what I do: programming by demonstration*. Cambridge, MA: The MIT Press, 1993.
42. Fruchterman TM and Reingold EM. Graph drawing by force-directed placement. *Software Pract Exper* 1991; 21(11): 1129–1164.
43. Kamada T and Kawai S. An algorithm for drawing general undirected graphs. *Inform Process Lett* 1989; 31(1): 7–15.
44. Meyer B. Self-organizing graphs: a neural network perspective of graph layout. In: Sue Whitesides (ed.) *Proceedings of the 6th International Symposium on Graph Drawing (GD '98)*, Montreal, Canada, 13-15 August 1998, pp. 246-262. London, UK: Springer-Verlag.
45. Cleveland WS and McGill R. Graphical perception and graphical methods for analyzing scientific data. *Science* 1985; 229(4716): 828–833.
46. Mackinlay J. Automating the design of graphical presentations of relational information. *ACM T Graphic* 1986; 5(2): 110–141.
47. Mackinlay JD, Hanrahan P and Stolte C. Show me: automatic presentation for visual analysis. *IEEE T Vis Comput Gr* 2007; 13(6): 1137–1144.
48. Ware C. *Visual thinking for design*. Burlington, MA: Morgan Kaufmann, 2008.
49. Ware C. *Information visualization: perception for design*. San Francisco, CA: Morgan Kaufmann, 2004.
50. Smith MA, Shneiderman B, Milic-Frayling N, et al. Analyzing (social media) networks with NodeXL. In: *Proceedings of the fourth international conference on communities and technologies*, University Park, PA, USA, 25-27 June 2009, pp. 255–264. University Park, PA: ACM.
51. GraphML file format, http://graphml.graphdrawing.org/ (2010).
52. Codd EF. A relational model of data for large shared data banks. *Commun ACM* 1970; 13(6): 377–387.
53. NetBeans rich-client platform development, http://netbeans.org/features/platform/ (2011).
54. H2 database engine, http://www.h2database.com/html/main.html (2011).
55. Meijer E and Bierman G. A co-relational model of data for large shared data banks. *Queue* 2011; 9(3): 30–48.
56. Van Ham F and Perer A. "Search, Show Context, Expand on Demand": supporting large graph exploration with degree-of-interest. *IEEE T Vis Comput Gr* 2009; 15(6): 953–960.
57. Norman DA. *The design of everyday things*. NY: Basic Books, 2002.
58. Grammel L, Tory M and Storey M. How information visualization novices construct visualizations. *IEEE T Vis Comput Gr* 2010; 16(6): 943–952.
59. Kobsa A. An empirical comparison of three commercial information visualization systems. In: *IEEE symposium on information visualization*, San Diego, California, 22-23 October 2001, pp. 123–130. Piscataway, NJ: IEEE Press.

Visual Analytics Support for Intelligence Analysis

Carsten Görg* University of Colorado
Youn-ah Kang† Google Inc.
Zhicheng Liu‡ Stanford University
John Stasko§ Georgia Institute of Technology

ABSTRACT

Intelligence analysis challenges investigators to examine large collections of data and documents and come to a deeper understanding of the information and events contained within them. Visual analytics technologies hold great promise as potential aids for intelligence analysis professionals. We describe our research to better understand intelligence analysis processes and analysts, learn how visual analytics can help investigators, and design visual analytics systems to serve in this role. To illustrate these ideas, we present a hypothetical intelligence analysis scenario that explores a collection of text documents using the Jigsaw system that we have created. The system combines computational analysis of document text with interactive visualizations of the document contents and analysis results. Evaluating such systems is very challenging and the article concludes by discussing potential evaluation methodologies for these types of systems.

Keywords: Visual analytics, investigative analysis, intelligence analysis, information visualization, knowledge acquisition, data exploration, case study, qualitative user study.

1 INTRODUCTION

Visual analytics is a relatively new research field that integrates the interactive visualization and exploration of data with computational data analyses [8]. Intelligence analysis has been one of the key application domains of visual analytics since the area's inception in 2004, facilitated by the creation of the National Visualization and Analytics Center by the Department of Homeland Security. An initial research roadmap [11] described challenges and goals of the new field and identified tasks, data, and analytical scenarios focused on homeland security and prevention of terrorism.

Enabling insights through the analysis of large amounts of diverse and dynamic data was the underlying grand challenge in the research agenda. As stated in [11], "The analysis of overwhelming amounts of disparate, conflicting, and dynamic information is central to identifying and preventing emerging threats, protecting our borders, and responding in the event of an attack or other disaster. This analysis process requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information in the face of rapidly changing situations to both detect the expected and discover the unexpected."

Intelligence analysis requires investigators to gather as much available data as possible in order to better understand a situation and then make judgments about the appropriate next steps to take. Two fundamental types of investigative scenarios exist within the intelligence domain: (1) targeted analysis scenarios, in which analysts are tasked with examining specific people, organizations, or incidents, as well as locations and dates, in order to either investigate past events or uncover an imminent threat, and (2) open ended, strategic analysis scenarios, in which analysts are tasked with learning as much as possible about a person, organization, country, or situation in order to gain a deeper understanding, conduct an accurate assessment, and possibly make a prediction on the likely chain of events that will occur at a later point in time. Examining and understanding large collections of textual documents plays an important role in both types of these scenarios. Analysts must gather nuggets of information within textual documents from diverse sources, ranging from reports from field agents to open source news articles. Examining textual documents is fundamentally a slow process (due to the sequential nature of reading) and it is challenging for the analysts to keep track of what they discovered and form an internal mental model that represents a coherent picture of the events, people, places, and organizations discussed in the documents. Uncovering and understanding the connections between those entities across a large collection of documents is one of the key challenges they face.

Based on a cognitive task analysis of working analysts, Pirolli and Card [9] identified a number of "pain points" in the intelligence process that are particularly challenging to human analysts. These pain points include the costs of scanning, recognizing (assessing), and selecting items for further attention; the costs of shifting attention and control; the limited span of attention for evidence and hypotheses; and the difficulty of generating alternative hypotheses. All these challenges are exacerbated when the amount of data to examine grows larger and larger. Today's "big data" technologies often make the acquisition of data easier, but they present increasing challenges to analysts who must review and investigate all that data. In this article we highlight a number of our research projects on intelligence analysis from the last five years, including an observational study to gain a better understanding of the intelligence analysis process and its characteristics, the development of a visual analytics system that integrates computational text analyses with interactive visualization in order to explore collections of documents, and an evaluation of the utility of the system via a controlled laboratory experiment as well as observational case studies of extended use of the system in the field.

2 INTELLIGENCE ANALYSIS PROCESS

Analyzing and understanding end-users needs and tasks is one of the fundamental requirements for creating useful computational tools. To better understand intelligence analysis, it is important to explore the mindset and methodologies of analysts as well as the fundamental processes they conduct. Heuer [4] examined the psychology of intelligence analysis and the types of mental reasoning analysts must engage in. In particular, he identified a number of challenges analysts must confront in the analytical reasoning process. For example, when people encounter a situation of uncertainty, they typically will develop a single hypothesis explaining the situation and will work to gather evidence confirming the hypothesis. Intelligence analysts, however, are trained to develop multiple hypotheses and seek out information that can discredit many of the hypotheses.

Many other researchers have studied the intelligence analysis process in order to construct abstract models of it. While a number of process models exist, most involve some form of iterative cycle of exploration, including steps such as data collection, pro-

*e-mail: Carsten.Goerg@ucdenver.edu
†e-mail: ykang@google.com
‡e-mail: zcliu@cs.stanford.edu
§e-mail: stasko@cc.gatech.edu

cessing, analysis and production, dissemination, and planning and direction [1].

Pirolli and Card's notional model of the sensemaking loop for intelligence analysis [9] has been widely cited and adopted by researchers within the visual analytics community. It consists of a linear set of states characterizing both data and process flow in an investigation. Analysts iterate through this process over the course of an investigation. At a high level, the model contains two primary loops: a foraging loop in which analysts collect data and evidence, and a sensemaking loop in which analysts reflect on the data in order to generate schema and hypotheses about the situation and ultimately construct a presentation of the findings. Each loop contains three stages that further refine the process and both loops are connected through an overarching reality/policy loop.

This model broadly characterizes the workflow used for analysis activities and it has guided the development of a number of computational tools. However, it abstracts substantially from analysts' work in the real world and does not provide an adequate level of detail necessary to develop tools that analysts can integrate seamlessly with their existing workflow. Furthermore, not all analysts agree that the linear structure of the model captures the way they work. Dr. Kristan Wheaton, Professor at the Department of Intelligence Studies at Mercyhurst College, proposed an alternative model in which modeling, collection, analysis, and production stages take place in parallel, just with different emphases over the course of an investigation. At the beginning, emphasis focuses on modeling and then throughout the investigation it shifts to collection, analysis, and finally production.

To better understand the analytical process and its requirements in the intelligence domain, we conducted our own qualitative user study [6]. Professor Wheaton provided us with the opportunity to observe three teams of intelligence analysts in training within the intelligence program at Mercyhurst College. The student teams each conducted an intelligence analysis project throughout an entire academic term (ten weeks). One team consisted of four undergraduate students and performed analysis on a project for which we served as a "client"; the other two teams consisted of graduate students and conducted a structured analysis on projects provided by external clients.

We found that four processes dominated the overall workflow: construction of a conceptual model, collection, analysis, and production. The study helped us to better understand some misconceptions that visual analytics researchers may harbor about intelligence analysis. For instance, analysis is typically not about finding an answer to a specific problem and it does not evolve in a sequential process. Instead, analysis is often about determining how to answer a question, what to research, what to collect, and what criteria to use. The process is often organic and parallel. Another misconception is that intelligence analysts typically operate as lone investigators, researching some problem. We, conversely, found that collaboration is commonplace and crucial, frequently being asynchronous. Also, the student analysts we observed did not seek grand, monolithic computational analysis tools. Instead, the teams used a variety of computational tools with many being small applications used for one specific purpose. They sought ways to integrate existing tools and easy-to-use new tools that leveraged existing analysis methods.

Finally, our study surfaced a number of recommendations for visual analytics technology developers:

- Externalize the thinking process - Help analysts continuously build a conceptual model
- Support source management - Enable managing both pushed and pulled information and organizing sources meaningfully
- Support analysis with constantly changing information - Integrate collection and analysis in a single system and help analysts use structured methods during collection
- Help analysts create convincing production - Support insight provenance and sanity checks of analytical products
- Support asynchronous collaboration rather than synchronous collaboration for exploratory analysis

3 Example Scenario Employing Visual Analytics

In order to better demonstrate how visual analytics can aid intelligence analysis, we present an example scenario. "The 9/11 Commission Report" is a publicly available report about the 9/11 terrorist attacks on the World Trade Center in New York. One version of the report is stored as a pdf document with 585 pages. In order to better simulate a larger collection of short intelligence reports, we split this document into 585 pages and consider each as a separate document. We use the page breaks as separators since the report does not have a natural structure that would lend itself to being split into short documents of a few paragraphs.

To illustrate this scenario, we employ the Jigsaw visual analytics system [10]. Jigsaw combines automated text analyses with interactive visualizations for exploring and analyzing collections of unstructured and semi-structured text documents. It automatically identifies entities of interest in the documents, such as people, places, and organizations, and then shows connections between those entities across the entire collection, as well as connections between documents and entities. Connections are defined by co-occurrence: if two entities co-occur in the same document, they are connected to each other as well as to that document. If entities co-occur in many documents they have a stronger connection. Even though this untyped connection model based on co-occurrence is very simple, it has turned out to be a powerful tool for investigative analyses. It works best if the documents are not too large, as it is often the case for news articles or case reports that usually span a few paragraphs.

We present the scenario from the point of view of a hypothetical intelligence analyst who is examining the document collection. To begin, the analyst imports the 585 single-page documents and runs an automatic entity identification. She uses the integrated OpenCalais webservice to identify people, locations, and organizations; the integrated GATE package to identify money entities, and built-in regular expression matching algorithms for identifying date entities. The analyst removes all entities that occur in only one document (they would not contribute to any connections) and performs a basic entity clean-up process, including removing wrongly identified entities and aliasing entities with multiple representations such as "George Bush" and "George W. Bush". The entire process results in a document collection with 369 people, 200 location, 252 organization, 12 money, and 464 date entities across the 585 documents.

The analyst begins the investigation seeking an overview of the entities. She uses the List View to display lists for Location, Person, Organization, and Money and change the list ordering from alphabetic to frequency-of-occurrence to see the most frequent entities in the document collection (Figure 1). The small bar to the left of each entity indicates the number of documents in which it occurs. Entities with aliases are shown in italic font and the aliases are displayed as tool tips, as shown for *Usama Bin Ladin*. The most frequent locations are *United States* (364 occurrences), *Afghanistan* (184), *Pakistan* (98), *New York* (77), and *Saudi Arabia* (71); the most frequent persons are *Bill Clinton* (65), *George W. Bush* (59), *Usama Bin Ladin* (59), *Richard Clarke* (50), and *George Tenet* (36); and the most frequent organizations are *al-Qeda* (233), *Central Intelligence Agency* (214), *Federal Bureau of Investigation* (181), *White House* (93), and *Federal Aviation Administration* (91).

The analyst next selects *Usama Bin Ladin* in the person list and reorders the other lists by strength of connection to the selection in order to see the entities most common with him (Figure 2, left).

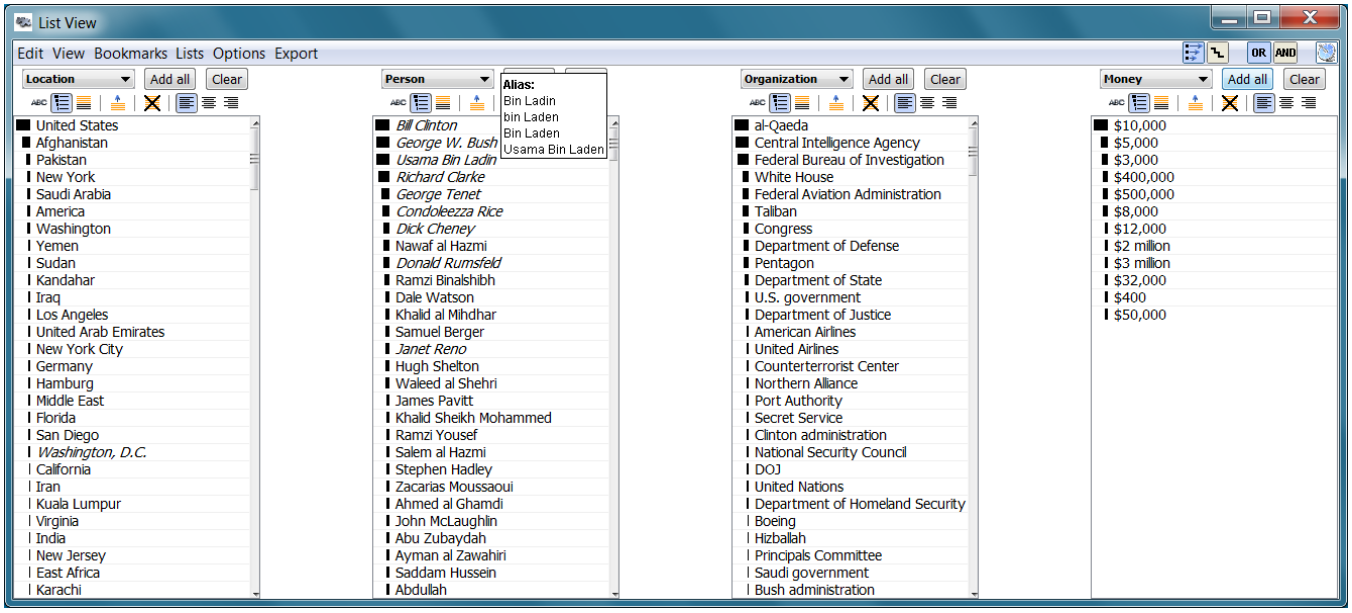


Figure 1: List View showing an overview of the 9/11 Commission Report, focusing on the Location, Person, and Organization, and Money entities. All entities are sorted by frequency.

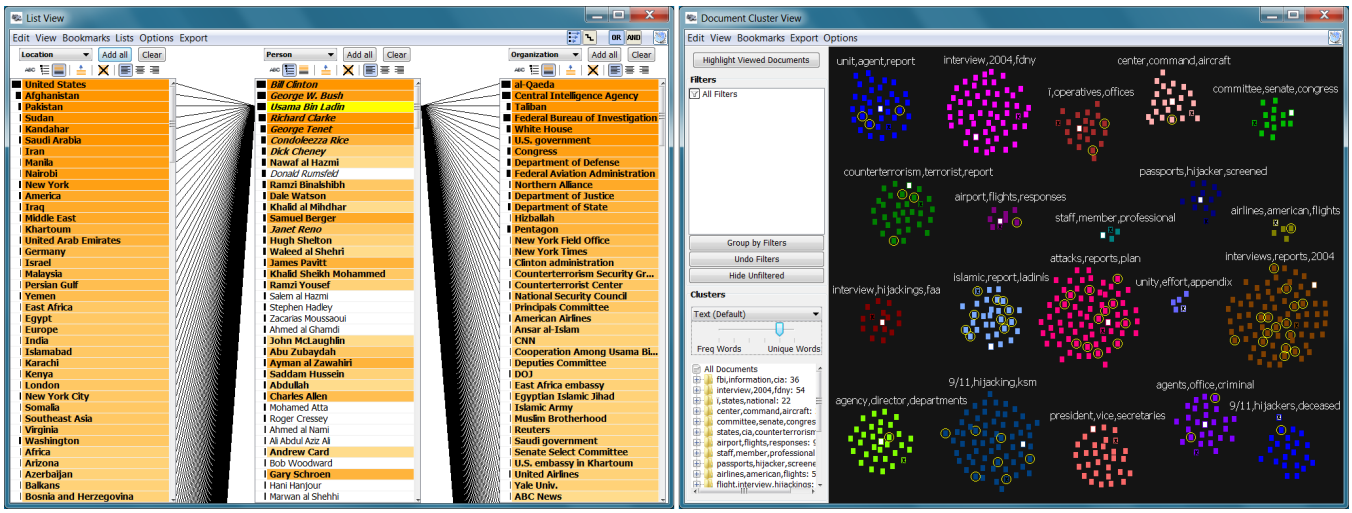


Figure 2: List View (left) showing locations, persons, and organizations connected to *Usama Bin Ladin*. Document Cluster View (right) showing different clusters of related documents (small rectangles in different colors). Documents connected to *Usama Bin Ladin* are selected (surrounded by a yellow circle).

The List View highlights entities connected to a selected entity (yellow) via an orange background. Darker shades of orange indicate or stronger (more frequent) connection. Entities that are not directly connected have a white background. *United States*, *Afghanistan*, *Pakistan*, *Sudan*, and *Kandahar* are the most connected locations to *Usama Bin Ladin*; *al-Qaeda*, *Central Intelligence Agency*, *Taliban*, *Federal Bureau of Investigation*, and *White House* are the most connected organizations. He is also strongly connected to the people *Bill Clinton*, *George W. Bush*, *Richard Clarke*, and *George Tenet*.

To better understand the themes and topics in the report, and in particular those in which *Usama Bin Laden* is mentioned, the analyst opens the Document Cluster View and displays the documents clustered by text similarity (Figure 2, right). Each document is displayed as a small rectangle and each cluster is labeled with three keywords. Text analysis algorithms integrated in Jigsaw automatically compute the clusters and summaries. The cluster summaries represent important topics in the report, including counterterrorism, hijackings, attacks, interviews, and president.

Cross-view selection and filtering are important capabilities in visual analytics systems. Since *Usama Bin Ladin* is still selected in the List View (Figure 2, left), the documents he appears in are also selected in the Document Cluster View (Figure 2, right), indicated by a yellow circle. He is connected to more than ten documents in the "islamic, report, ladin's", "attacks, reports, plan", and "interviews, reports, 2004" clusters and to seven documents in the "9/11, hijacking, ksm" cluster. The analyst also could use cross-view selection in the opposite direction for a different kind of exploration: when she selects the "president, vice, secretaries" cluster, she observes in the List View that *George W. Bush*, *Dick Cheney*, *Condoleezza Rice*, and *Donald Rumsfeld* are the most connected people to that cluster. Interestingly, *Donald Rumsfeld* is not connected to *Usama Bin Ladin* (Figure 2, left).

To learn more about Rumsfeld, the analyst opens the Graph View and explores the people and organizations connected to him using a "circular layout" approach (Figure 3, left). This approach positions the documents that mention *Donald Rumsfeld* on a circle (white

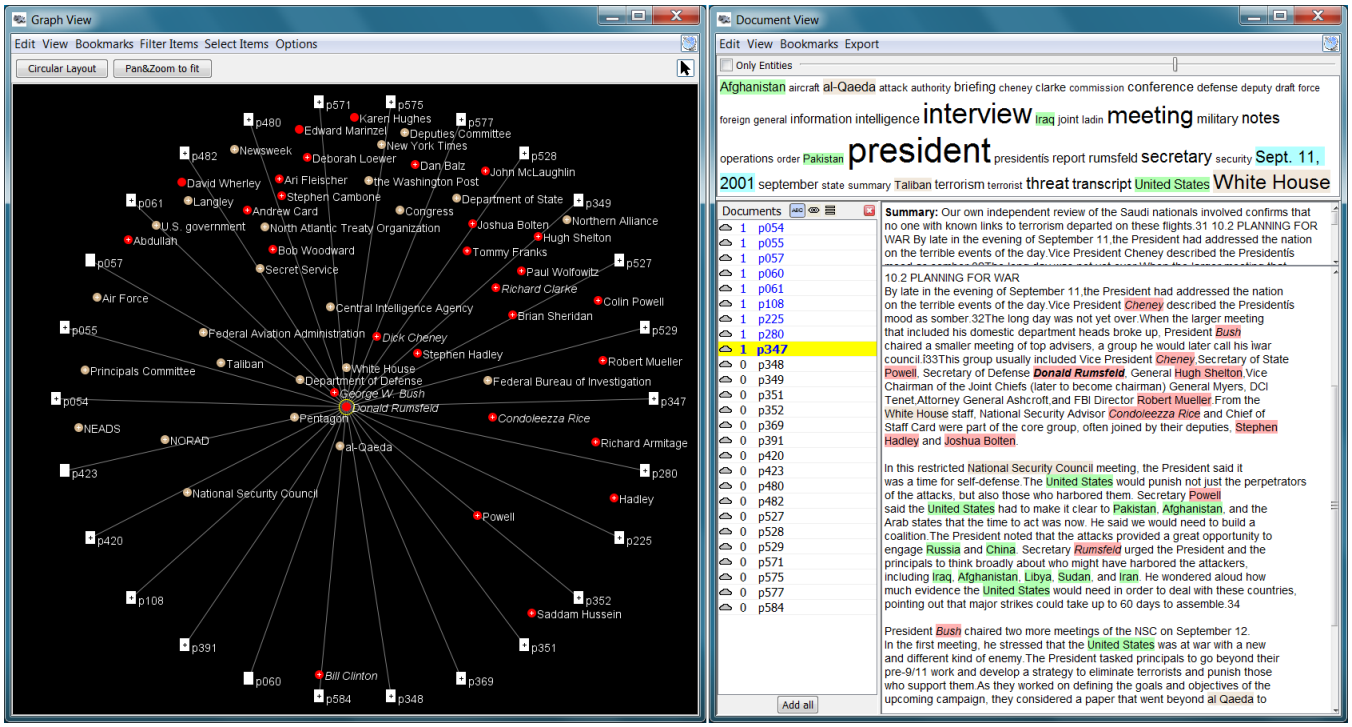


Figure 3: Graph View (left) showing a Circular Layout with documents (white rectangles) that mention *Donald Rumsfeld*. Persons (red circles) and organizations (tan circles) are positioned inside the circle of documents, the more connected they are the closer they are positioned to the center of the circle. Document View (right) showing all documents mentioning *Donald Rumsfeld*. Above the selected document's text (right) is a one sentence summary and below are the affiliated entities. The word cloud (top) summarizes all documents loaded in the view.

rectangles) and the related entities within that circle (red circles for people, tan circles for organizations). More highly connected entities are placed closer to the center of the circle. The layout shows that *Donald Rumsfeld* is strongly connected to *George W. Bush*, *Dick Cheney*, and *Stephen Hadley*, as well as to the organizations *White House*, *Department of Defense*, and *Pentagon*. The Graph View also supports interactive exploration of the connection network via expand and collapse operations. A double click on a document or an entity expands that item and brings in all other items that are connected to it. Items having additional connections that are currently not shown are indicated with a plus sign. A double click on an item that already shows all its connected items (e.g. *Donald Rumsfeld*) collapses that item and hides all its connected items.

The analyst next wants to read the documents about Donald Rumsfeld, so she opens them in the Document View (Figure 3, right). The document list (left) shows the 26 documents (pages) that mention him; documents shown in blue have already been examined, and the number in front of the document indicates how often it was displayed. The word cloud (top) summarizes the currently loaded document set using the most frequent words in those documents. The selected (yellow) document in the list is presented on the right. Above the documented text is a one-sentence summary of the document computed by a text summary analysis. To support quick scanning of documents, entities in the word cloud and in the document itself are highlighted: people in red, locations in green, dates in blue, and organizations in tan. This document (p347) talks about a restricted National Security Meeting on the night of the attacks in which "Rumsfeld urged the President and the principals to think broadly about who might have harbored the attackers, including Iraq, Afghanistan, Libya, Sudan, and Iran." (The numbers in some sentences in the document are footnote references.)

To investigate if similar documents exist in the collection, the analyst opens the Document Grid View and sorts and colors the documents by their similarity to document p347 (Figure 4, top). The shading of blue indicates document similarity: dark blue indicates similar documents, light blue indicates documents that do not have much in common with the reference document. A tooltip provides the one sentence summary of a document. A similar document p215 mentions "Bonk told Bush that Americans would die from terrorism during the next four years. During the long contest after election day, the CIA set up an office in Crawford to pass intelligence to Bush and some of his key advisors." It seems that there might have been some miscommunication in the post election transition. To understand the role of the former president *Bill Clinton*, the analyst displays his name in the Word Tree View (Figure 4, bottom). A Word Tree [12] shows all occurrences of a word or phrase across all documents in the context of the words that follow it. Each word can be explored further by a click. The Word Tree View for *Bill Clinton* shows that his "administration effectively relied on the CIA to take the lead in preparing long-term offensive plans" and that "One of the great regrets of my presidency is that I didn't get him [Bin Ladin] for you".

Sidebar: Visual Analytics Tools for Intelligence Analysis

A few commercial tools for intelligence analysis employ visual analytics techniques including Analyst's Notebook from IBM i2 (`http://www.ibm.com/software/industry/i2software`), nSpace from Oculus (`http://www.oculusinfo.com/nspace`), and Palantir's suite of systems (`http://www.palantir.com`). An extensive discussion of academic research projects employing visual analytics for understanding text and document collections can be found in [2].

The scenario exploring the 9/11 Commission Report and the images used in this article were produced using the Jigsaw visual analytics system. Jigsaw was designed to help investigators ex-

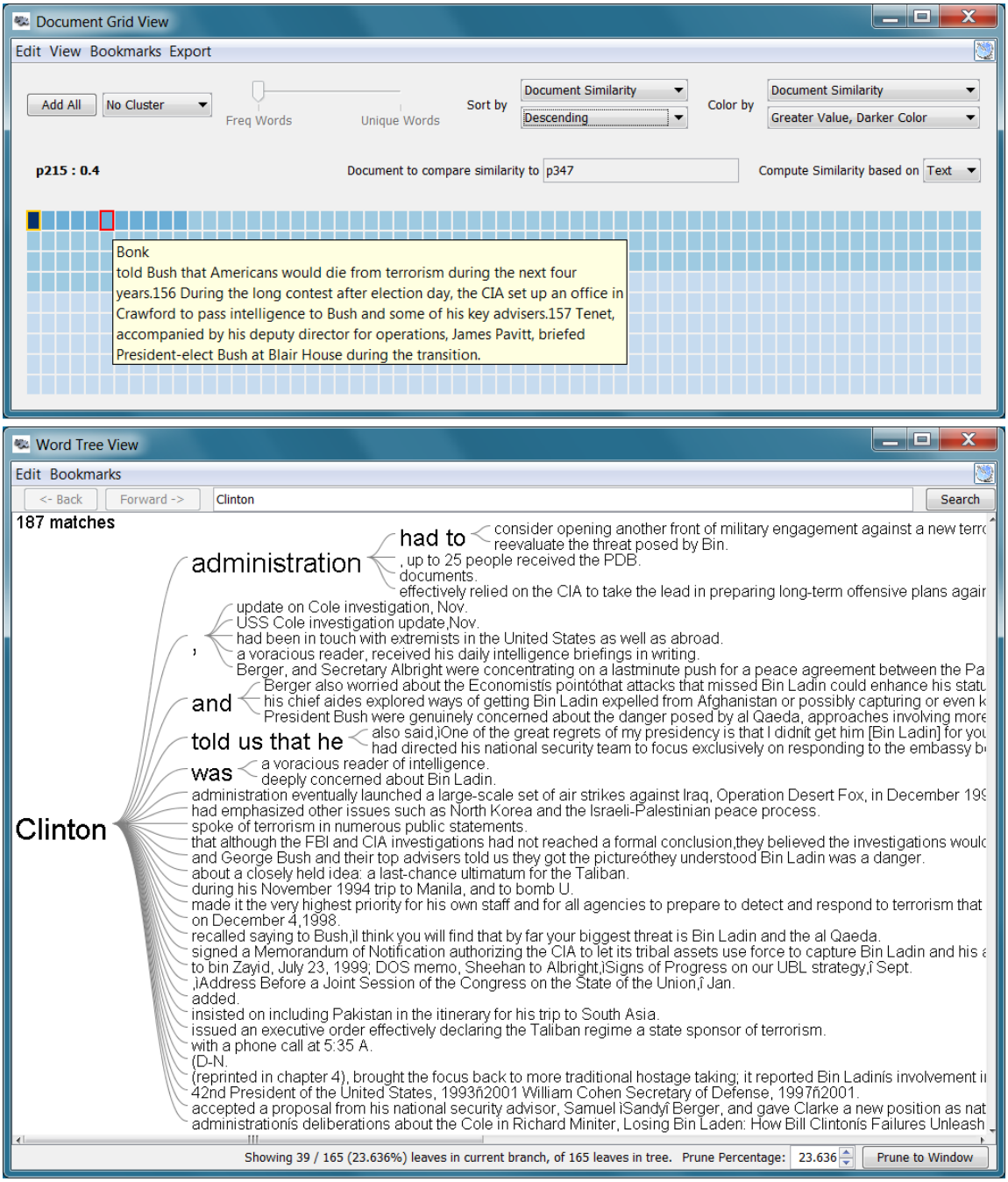


Figure 4: Document Grid View (top) with the document (small rectangle) order and shading set to correspond to the document's similarity to the selected document p347. Word Tree View (bottom) showing occurrences of the person *Clinton* and the most common phrases that follow him in sentences within the 9/11 Commission Report.

plore and understand collections of text documents, and in particular, to follow trails of ideas embedded across the documents. The system's name comes from the notion of "putting the pieces together." Early versions of Jigsaw emphasized a suite of interactive visualizations portraying the documents' contents and connections between entities in the documents [10]. More recently, we have integrated computational text analysis capabilities [2]. Beyond intelligence analysis, Jigsaw has been used to explore consumer review, academic research, fraud, investigative reporting, law enforcement, business intelligence, and email document collections. The Jigsaw system itself, as well as example datasets (including the one used in this article), tutorials, videos, and related articles are available for download on the webpage (`http://www.cc.gatech.edu/gvu/ii/jigsaw`).

4 Evaluating Systems for Intelligence Analysis

The evaluation of visual analytics systems is a challenging research area in itself: there is no consensus among researchers on how to effectively and objectively measure the contributions of a system to an analyst's generation of insights. This is especially true when intelligence analysis is the domain being studied. In our research, we have used multiple techniques to evaluate the effectiveness of the systems we have built.

In one evaluation study [5], we employed a small synthetic dataset with embedded ground truth, consisting of 50 imaginary short reports with a hidden threat. We then recruited sixteen students, divided them into four groups, and asked them to conduct an analysis of the documents and identify the hidden threat. Participants in the first group only worked with pencil and paper. They received a printout of all the reports and some blank sheets for note taking. Participants in the second group received an electronic copy of the reports and could use basic text editing software for reading and searching the documents. Participants in the third group used only the Document View of the Jigsaw system to read and analyze the document collection. This setup was similar to the previous one, providing functionality for reading and searching; however, the Document View also highlighted identified entities within the documents. Participants in the fourth group used the entire suite of visualizations in Jigsaw to conduct the analysis.

The study participants worked with the documents for 90 minutes and then wrote debriefing statements, which we compared to the ground truth and then graded for accuracy. We also conducted follow-up interviews and collected their notes. Additionally, we videotaped all the sessions and used screen capture software in the settings where participants worked on a computer. We started our analysis with an inductive approach to examine the qualitative data in order to unveil potential concepts and themes and to understand the influence of the tools in the different settings. At a later stage of our analysis we combined inductive and deductive approaches and supplemented them with observations from the video logs, screen captures, and other quantifiable data.

We found that overall the participants using the full Jigsaw system outperformed all other groups on average. Because of the small subject population this result was not statistically significant, however. We did observe four particular strategies that participants employed in their investigations. These strategies ranged from first reading all the documents very carefully to finding an initial clue and following a trail from it. The participants using Jigsaw applied three out of the four strategies and performed well using any of those strategies.

We also used Jigsaw for our own investigations and participated in a number of IEEE VAST Conference Challenges and Contests over the past few years. These contests provide synthetic document collections with embedded ground truth. Participating teams are tasked with finding a hidden threat in the documents. Working with Jigsaw on the contest datasets helped us to gain practical experience in these types of intelligence investigations, improve the system, and develop additional functionality. We further describe the influence of our participation in the VAST contests on the design and development of the Jigsaw system in another article [3].

Because we have made Jigsaw available to anyone to use in their own work, examining real world use by other people is another type of evaluation that we have employed. In order to better understand how professional analysts have been using the system and to determine its benefits and limitations in practice, we interviewed six investigators who had been using Jigsaw for an extended period of time [7]. These individuals included an aerospace engineering researcher, a business analyst investigating fraud, a doctoral candidate in Industrial and Systems Engineering studying enterprise transformation, and intelligence analysts at a national lab, the Air Force, and a police department. The goal of this study was to evaluate whether Jigsaw is helping analysts with their tasks, to understand its application to different types of documents and domains, and to identify useful features and capabilities of the system as well as missing or problematic features.

We identified a number of applications of the system across more than one participant. Many used Jigsaw to find connections and relationships between entities, one of the core goals of the system. Some used it as a search and comparison tool to more conveniently work with text documents, and many used it to gain a broader understanding or overview of their documents. Surprisingly to us, some of the participants also used Jigsaw as a communication aid to share their understanding with others. We originally created the system as an analysis tool and that application is always how we have thought of it. It was interesting to note that some of the study participants also were using it to present findings and tell a story to their colleagues.

The investigators in the study identified a number of limitations and issues with the system as well. Some of the participants wanted better ways to work with only subsets of their document collections. They wanted to be able to dynamically filter out documents in an investigation, but also maintain the ability to reintroduce filtered documents as desired. Document import was another particular challenge and often required manipulating and translating their original documents into a form that Jigsaw could better analyze. Furthermore, problems that arose in document import or in any other use of the system raised questions in the investigators' and their colleagues' minds about the accuracy of the system. They commented how any kind of issue or usability problem eroded their trust.

Our study identified a number of future objectives for Jigsaw and other visual analytics systems for document analysis. The investigators all believed that entity identification is crucial and they wanted easier and more reliable mechanisms to perform it and correct/modify it. They also sought to have more flexible mechanisms for document management activities such as import, storing, filtering, and maintaining. A number of the users wanted more quantitative and statistical analysis capabilities. For instance, they expressed a desire for more network analysis and modeling metrics. In terms of the user interface, some of the investigators wanted to be able to annotate the system views, highlight particular items, and add notes and comments on top of the visual representations.

5 Conclusion

Intelligence analysis requires people and organizations to review and assess large collections of information in order to better understand current situations and take the appropriate next steps. The sheer scale, diversity, and complexity of the information to be explored often makes such analysis cognitively demanding. Furthermore, the information is often recorded as narrative text, not quantitative data, and thus it is not as amenable to automated analysis techniques.

Visual analytics technologies that combine computational text analysis with interactive visualization provide a powerful new paradigm for helping intelligence analysts in their work. While current visual analytics systems have illustrated the potential of the field, many challenges remain. Visual analytics systems must scale to increasingly larger collections of data in order to keep up with our growing ability to log and record information. Additionally, visual analytics systems should assist investigators in the complex processes of analytical reasoning, hypothesis formulation, and decision making.

6 Acknowledgements

This research is based upon work supported in part by the National Science Foundation via Awards IIS-0414667, IIS-0915788, and CCF-0808863, by the National Visualization and Analytics Center (NVAC^{TM}), a U.S. Department of Homeland Security Program, and by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

[1] Central Intelligence Agency. *A Consumer's Guide to Intelligence*. Diane Pub Co, 1999.

[2] C. Görg, Z. Liu, J. Kihm, J. Choo, P. Haesun, and J. Stasko. Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 2013. to appear.

[3] C. Görg, Z. Liu, and J. Stasko. Reflections on the Evolution of the Jigsaw Visual Analytics System. *Information Visualization*, 2013. to appear.

[4] R. Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, 1999.

[5] Y.-a. Kang, C. Görg, and J. Stasko. How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):570–583, May 2011.

[6] Y.-a. Kang and J. Stasko. Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *IEEE VAST*, pages 21–30, Oct. 2011.

[7] Y.-a. Kang and J. Stasko. Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2869–2878, 2012.

[8] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. *Information Visualization: Human-Centered Issues and Perspectives*, pages 154–175, 2008.

[9] P. Pirolli and S. Card. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. In *International Conference on Intelligence Analysis*, May 2005.

[10] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Information Visualization*, 7(2):118–132, 2008.

[11] J. J. Thomas and K. A. Cook. *Illuminating the Path*. IEEE Computer Society, 2005.

[12] M. Wattenberg and F. B. Viégas. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.

American Journal of Geographic Information System 2014, 3(2): 75-87
DOI: 10.5923/j.ajgis.20140302.02

A Geographic Information System Model for Hurricane Track Prediction

Venkata B. Dodla[1], Sudha Yerramilli[2,*]

[1]TLGVRC, Jackson State University, 1230, Raymond Road, Jackson, MS, 39204, USA
[2]National Center for Biodefense Communications, Jackson State University, 1230, Raymond Road Jackson, MS, 39204, USA

Abstract Geographic Information System(GIS) tools have been applied to build a model for the prediction of hurricane tracks of the Gulf of Mexico region. This is an analog model based on the climatologically nature of the movement of the global tropical cyclone systems in general. The model uses information from the historical tropical cyclone track data as available from the archives. Selective GIS and structured query language tools are applied to pick up all the historical tropical cyclone systems that are present within a selective radius of the active hurricane location and use them to produce the predicted track. The GIS model requires inputs of the in situ location of the active hurricane and three optional parameters. A case study of Hurricane Katrina (2005) track prediction has shown that the GIS model could predict the track with errors comparable to those from dynamical models.

Keywords Hurricanes, Track prediction, Modelling, Geographical information system

1. Introduction

A hurricane is the most destructive of all natural disasters, unleashing energy comparable to that of an earthquake, a volcano, a tsunami, or a nuclear weapon. Like most of the geophysical events, their prediction is very difficult. Although their intensification could be estimated in terms of sea surface temperature and other environmental factors, their movement from genesis to landfall is quite irregular. "Tropical cyclone" is a generic name for low pressure systems over tropical oceans with cyclonic surface wind speeds exceeding 33 meters/sec (74 mph). They are referred to as "hurricanes" in the North Atlantic and Northeast Pacific Ocean (east of dateline) basins and "typhoons" in the Northwest Pacific Ocean. Their global annual frequency is about 85 and about 9 hurricanes in the Atlantic. Analyses of historical hurricane data clearly indicate a long-term trend of an increase of one-hurricane every 30-years. With regards to the available historical records of hurricanes available since 1840's, there could be inconsistencies due to hurricane measurements and reporting. Hurricanes were recorded as observed at landfall or reported from surviving boats in 1900s, with added reports from airplanes since 1950s and all hurricanes are identified and reported since the satellite era of 1960s. Land sea et al. (2010) suggested that the increase of hurricanes in the 20th century was due to improved quantity and quality of observations. An account of U.S. hit hurricanes is that 158 hurricanes of which 64 were major hurricanes with categories 3-5 had landfall, and Florida had the most of 57, followed by Texas with 36 and Louisiana and North Carolina with 25. The hurricane landfall regions of the United States were divided up into four main regions for frequency considerations: (1) The Gulf Coast from Brownsville, Texas to the Florida Peninsula; (2) The west coast of Florida from the Florida Panhandle to the Florida Keys; (3) The east coast of Florida from the Florida Keys to the Florida/South Carolina border; and (4) the remaining East Coast from South Carolina to Maine. Devastation from hurricanes occurs mainly around the landfall time due to heavy rain and strong winds causing damage to buildings, trees and cars and storm surges leading to coastal inundation. The Atlantic hurricane season is from June 1 to November 30, with a peak during mid-August to late October.

In the era of global warming, increase in frequency, intensity, and duration of cyclone systems has been reported suggesting more destruction from these disasters. This is of great concern as the coastal habitation is rapidly increasing, with a growth of 50% from 1980 to 2003 along the south-eastern U.S. This means that the current hurricanes at the landfall would now inflict a much greater loss of life and property than from a similar hurricane 30 years ago and if hurricane frequency and intensities were to increase the problem would be even worse. (Emanuel, 2005). All this factual information emphasizes the need for hurricane landfall prediction as precisely as possible and with as much lead time as possible as the same would be critical for disaster mitigation and management.

* Corresponding author:
Sudha.yerramilli@jsums.edu (Sudha Yerramilli)
Published online at http://journal.sapub.org/ajgis

In U.S., National Hurricane Center (NHC) is the official agency which provides the weather prediction on different scales, which include hurricane prediction. NHC hurricane predictions are based on dynamical models, which use mathematical equations for atmospheric motion and physical processes and numerical methods to solve them. This prediction method of solving differential equations is an initial value problem and dependent on the initial state of the atmosphere. At the initial time, all the model variables over the numerical model domain are to be derived from irregularly spaced observations and so vary from the real atmosphere leading to uncertainties and errors in prediction. Since the errors in the initial state tend to grow with time during the forecast, small initial errors can become very large within a few days of the forecast period. The predictions from these models are time consuming and take few hours to produce the desired forecasts, as the model integrations take a few hours on the fastest super computers of the present time. NHC is presently using both the global and regional models for hurricane prediction. At least 5 global models with variations in numerical solutions, domain resolution and representation of physical processes and three different regional models specially designed for hurricane prediction are being used. The NHC official forecast errors are noted to be 55, 102, 147, 189 and 278 miles at 12, 24, 36, 48 and 72 hours (NOAA, 2012). In addition ensemble and consensus forecasts are under evaluation. Consensus forecasts are obtained by combining the forecasts from a collection of models, whereas the ensemble consists of multiple runs of a single model or runs from different independent models. They range from simple statistical models to complex dynamical models. A brief description of only the track prediction models are presented here and more details are available at the NHC website (NOAA, 2009).

http://www.nhc.noaa.gov/modelsummary.shtml. In addition to dynamical models, statistical and statistical-dynamical models are also used to have track guidance and bench marking. Statistical models are built up on use statistical relationships, such as Climatology and Persistence Model (CLIPER5) which is based on regression relationships between parameters of movement during the previous 12- and 24-hour periods, the direction of motion, its current latitude and longitude, date, and initial intensity of the current hurricane to historical track records. At present this is used as a bench mark to compare the dynamical forecasts. Statistical-dynamical models (NHC98 for Atlantic and NHC91 for east Pacific) use statistical relationships between storm behaviour and predictors of CLIPER5 and forecast predictors of steering flow obtained from dynamical model forecasts, such as the deep-layer-mean GFS geo potential heights fields (averaged from 1000 to 100-millibars).

Geographic Information System (GIS) is a database system with software that can analyse and display data using digitized maps and tables for planning and decision-making (Matejicek, 2005). A GIS can assemble, store, manipulate, and display geographically referenced data, tying this data to points, lines and areas on a map or in a table. GIS can be used to support decisions that require knowledge about the geographic distribution of people, hospitals, schools, fire stations, roads, weather events, the impact of hazards/ disasters, etc (Yerramilli, Dodla, & Yerramilli, 2011) Any location with known latitude and longitude or other geographic grid system can be a part of a GIS. GIS is a very useful tool for many aspects of emergency management, including: emergency response, planning, exercises, mitigation, homeland security and national preparedness. In addition to its ability to manage and display data, GIS has robust modelling capabilities, allowing its users to adjust data and scenarios for prediction, planning and estimation (FEMA, 2013). The current trend in GIS is on web-based mapping. This capability can allow users to view an already created map or create maps, based on their own specifications, on their personal computers. Web-based mapping is expected to widely expand the use of GIS in the workplace, in schools, and in homes (Zarcadoolas, Boyer, Krishnaswami, & Rothenberg, 2007).

GIS tools were identified to provide a standardized platform for meteorological data analysis, with advantages of bringing multiple and large geospatial data bases together for climate research and to explore spatial patterns in meteorological data leading towards development an atmospheric information system (Wilhelmi and Brunskill, 2003). At this time, GIS tools are being used by NOAA Environmental Visualization Program and National Hurricane Center to visualize hurricane related satellite data and model outputs, and GIS based hurricane response and management through Environmental Systems Research Institute (ESRI). However, attempts have not been made so far towards the use of GIS tools to develop a hurricane prediction system. As such, this is a first attempt to develop a hurricane track prediction model based on GIS tools.

In this paper, we report the results on the development of a GIS based hurricane prediction model for application to Atlantic Ocean basin affecting the U.S. Gulf Coast. The basis of the methodology is to use analog techniques for identifying all the past hurricanes similar to the in situ hurricane considering the parameters of location and time of the year. Descriptions of the GIS model development, basic data and the results of model application to Hurricane Katrina landfall prediction are provided in Sections 2, 3 and 4 respectively. Our study is based on the premise that hurricanes have a natural trend as evident from the historical data and quick exploration of all the past hurricanes with similar characteristics may prove valuable to have track guidance almost on real time. The developed GIS model would generate the output on hurricane track within 2 minutes as compared to few hours with dynamical model prediction, thus providing valuable lead time for disaster planning and mitigation measures.

Hurricane Katrina (2005) was taken for case study as it was noted to be one of the most devastating natural disasters in United States history, and one of the five deadliest hurricanes ever to strike the United States, inflicting

catastrophic damage and enormous loss of life in Louisiana and with its effects extending into the Florida, Georgia, and Alabama. This is also the most recent hurricane, for which different dynamical models showed significant differences in the track prediction indicating the uncertainties in model prediction. At this time, it may be said that no single dynamical model could be identified as the most suitable for prediction considering their performance over the last 5 years.

2. Description of GIS Model

The geo processing framework of the GIS model development that predicts the active hurricane track from the historical data is presented in this section and the "process flow diagram" representing the design of the framework is shown in Figure 1. This model has been developed using model builder application in Arc GIS 10. The work flows connect together sequences of geo processing tools feeding the output of one tool into another tool as input. Once the logical connections of the geo processing tools are successfully established, the elements in the model attains a specific color implying the model is ready to run. The inputs to the model are represented in blue color, the tools in yellow color and the green color represents the derived data from the tool. All the outputs derived from the model are stored in a geo database.

This hurricane track prediction model developed using model builder application facilitates the automation of the whole process by storing the associations, input parameters and other data features of the various work flows involved in the geo processing tasks.

2.1. Conceptual and Logical Process of the Model

The GIS model is built upon the concept of predicting an active hurricane track and its landfall point from an historical dataset presented by NOAA for the hurricanes in the past 100 years.

The first step is to pick up the *in situ* or active hurricane location point (hereafter referred to as HLP), in terms of latitude and longitude, from the official source, National Hurricane Center. The GIS model gets initialized with the HLP input data. An active region around the HLP is defined in terms of a parameter α, chosen as the distance from the HLP. All the historical hurricane/storms within the specified radial distance of α around the HLP were identified using the buffer tool in which the 'linear unit' parameters are set as per the requirements of α. All the hurricane points falling in the radial distance of α around HLP are extracted from the historical hurricane point dataset by using clip tool. To create this new feature class, the clip tool uses the buffer distance area (α) created by the buffer tool and historical hurricane dataset as the input parameters.

To avoid/remove duplications (two or more points from the same hurricane track) from the clipped feature class and at the same time, retain the closest point to the active hurricane point, the near and dissolve tools are used. Using dissolve tool, the hurricane point features with the same identities (ids) are aggregated by choosing the dissolve field as the hurricane ids and the closest point is retained by opting the statistical field to choose the point with a minimum distance from the active hurricane point (obtained from near tool).

Once the duplications are removed, the complete track data for each hurricane point is retrieved from the historical dataset by building a SQL (Structured Query Language) query using Make Query Table. This query is structured to pull the complete track data/lifecycle of the each hurricane point in the new feature class, matching their ids with the hurricane ids in the base dataset.

As we do not need the hurricane track information prior to the time points, the track points that are present in the preceding time periods from the buffer region are excluded by building a query to pick the hurricane ids with hurr_id value greater than the hurr_id value in the feature class.

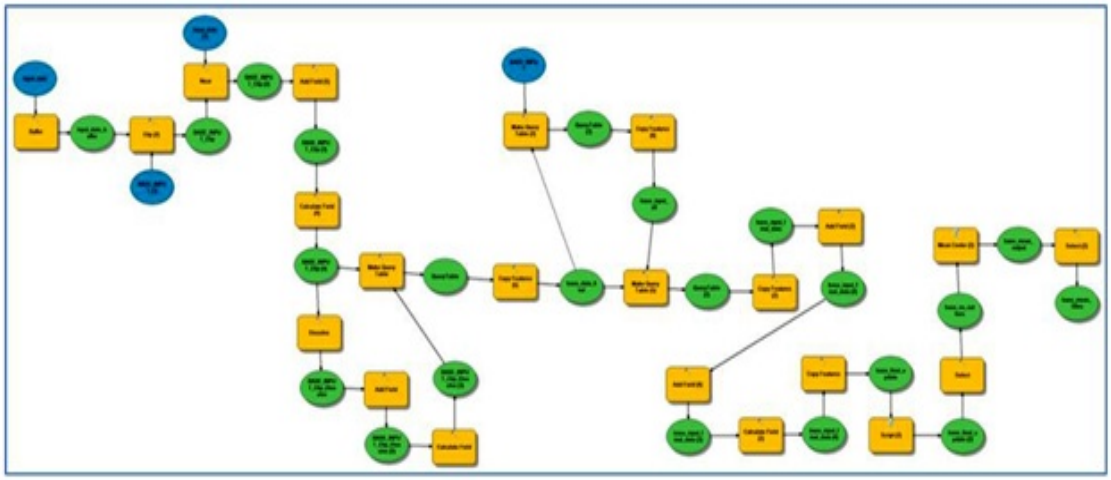


Figure 1. Flow chart depicting various steps of the GIS model built up for hurricane track prediction

With this execution, the hurricane points in the buffer region will have tracks starting from the buffer region till the end of their lifecycle. In order to assign a sequential number (time_id) to each time point in hurricane (from the starting point to the end of its life period), a custom tool is created in python script. This tool reads the hurricane ids in the database and based assigns a time_id number in a sequential/ascending order with respect to the time period.

To explain further, hurricane track points having a number value of '1' represent the first position of the track that start in the buffer area and the next points with number '2' represents the recorded position of the track after 6 hours in each hurricane. This 6-hour time period is due to the specification of the hurricane tracks at 6-hour interval coinciding with 00, 06, 12 and 18 UTC. A new feature class is created with updated time_id fields. However, this data may have outlier points due to various reasons such as errors occurred in the data entry. Select tool is used to remove the outlier points, with the imposition of a condition that the track points that fall in the range of ± 20 degrees of distance from the active hurricane point's with latitude and longitude are selected and the output is saved to a new feature class.

With this new feature class, the model predicts the projected path for the active hurricane by calculating the mean of all the hurricane tracks by grouping the time_id fields. This has been achieved by using the Mean Center tool by setting the case field parameters as time _id. The output from the mean center tool generates a possible projected path for the active hurricane, with location points given at 6 hours interval. From this projected path, a 72 hour track path is selected by choosing the mean points with time_id field less than ≤ 13 and the final output is retrieved to a new feature class.

Thus the GIS model provides a future projected hurricane track for the succeeding 72-hours, built up on the GIS tools and historical hurricane track data base as input. The GIS model has been built up using ESRI ArcGIS 10 and run on a desktop computing system with AMD Phenom Quad-core 9850 2.50GHz Processor, 4 GB RAM and 64-bit Windows 7 operating system. The model run for each experiment took 2-3 minutes of processing time.

Thus the GIS model provides a future projected hurricane track for the succeeding 72-hours, built up on the GIS tools and historical hurricane track data base as input. The GIS model has been built up using ESRI ArcGIS 10 and run on a desktop computing system with AMD Phenom Quad-core 9850 2.50GHz Processor, 4 GB RAM and 64-bit Windows 7 operating system. The model run for each experiment took 2-3 minutes of processing time.

3. Description of Hurricane Track Historical Dataset

A new global dataset of tropical cyclone tracks, IBTrACS (International Best Track Archive for Climate Stewardship), has become available for scientific community from National Climate Data Center (NCDC) of National Oceanic and Atmospheric Administration/ National Environmental Satellite, Data, and Information Service (NOAA/NESDIS). The World Meteorological Organization Tropical Cyclone Programme has endorsed IBTrACS as an official archiving and distribution resource for tropical cyclone best track data. This dataset contains the most complete information of historical tropical cyclones available, compiled from numerous tropical cyclone datasets, provided in popular formats to facilitate analysis with quality control checks of storm inventories, positions, pressures, and wind speeds. A description of the dataset is given by Knapp et al. (2010) and the data is accessible from the official website of NOAA/NESDIS/NCDC (NOAA, 2013). The data is provided in different formats: "IBTrACS Dataset: IBTrACS storm files" which is the archived format of the IBTrACS data in which all parameters are available with one storm in each data file and all other formats are derived from this set of files; "netCDF" in which all IBTrACS files are combined into one file with all variables stored in the IBTrACS Dataset except the original reports from the source agencies (e.g., original latitude, longitude, pressure, wind); and "CSV - Comma Separated Variables" which contains most of the variables which had facility to be imported to a spreadsheet or database. In addition to these formats, tropical cyclone track data points are also given in the form shape files as suitable for using with GIS software. For the present study, we downloaded the following files: ibtracs_v02r01_pts.dbf; ibtracs_v02r01_pts.prj; ibtracs_v02r01_pts.sbn; ibtracs_v02r01_pts.sbx; ibtracs_v02r01_pts.shp; ibtracs_v02r01_pts.shp.xml and ibtracs_v02r01_pts.shx files from NOAA.

This IBTrACS dataset contains complete information of the tropical cyclone tracks, and the built up of the GIS model for hurricane track prediction in this study required the use of only few parameters that are "STROMID", "SEASON", "STNUMBER", "BASIN", "STORMNAME", "OBDATE", "LATITUDE", "LONGITUDE" and "MAINSID".

3.1. Results

The built up GIS hurricane track model has been used to predict the track of Hurricane Katrina up to 72-hours prior to its landfall for validation. A brief description of the life cycle of Hurricane Katrina is provided in subsection 4.1. The GIS model has been applied first to study the sensitivity to certain parameters of buffer radius, seasonal time period and length of the historical data. This has been done through prediction of the track of Hurricane Katrina for 72-hours from "time of chosen point", which is approximately 48 hours before the landfall. The best possible values were ascertained through statistical error analysis and comparison of the "Root Mean Square Error" and "Mean Absolute Error". These results are described in subsection 4.2. Following this, the track of Hurricane Katrina was predicted at its different stages during the last 72 hours of its life cycle using the best values of the parameters. Here also we have computed the error statistics for each prediction experiment and presented mean errors.

These errors were compared with official NHC model estimations and their error statistics. These results are described in subsection 4.3.

3.2. Hurricane Katrina

Hurricane Katrina was noted to be not only one of the five deadliest hurricanes ever to strike the United States but also as one of the most devastating natural disasters in United States history. The life cycle of this hurricane spans the period during 23–30 August 2005 with landfalls, as Category 1 hurricane on the southeastern coast of Florida at around 2230 UTC 25 August and with Category 3 intensity near the mouth of the Pearl River at the Louisiana/Mississippi border at 1110 UTC 29 August (Anne, 2005). Hurricane Katrina was first identified as a tropical wave on 19 August; as a tropical depression on 23 August was designated as the cyclone Katrina at 1200 UTC 24 August with its center located at about 65 nautical miles east–southeast of Nassau. Katrina attained hurricane intensity at around 2100 UTC 25 August, had its first landfall on the southeastern coast of Florida around 6:30 PM EDT (2230 UTC) 25 August with Category 1 intensity and moved west-southwestward over the southeastern Gulf of Mexico. Hurricane Katrina attained Category-5 stage during 0600 UTC to 1800 UTC 28 August with wind speeds reaching 150 knots; moved westward and weakened rapidly after 1800 UTC 28 August and had its second landfall near Buras, Louisiana at 5:10 AM CDT (1010 UTC) 29 August and made its third landfall near Pearlington, Mississippi and Slidell, Louisiana at 10:00 AM CDT (1500 UTC) 29 August as a Category 3 hurricane with an estimated wind speed of 105 knots (Figure 2). Later Katrina rapidly weakened over land to become a Category 1 hurricane by 1800 UTC 29 August, a tropical storm by 0000 UTC 30 August, a tropical depression at 1200 UTC 30 August and transformed into an extra-tropical low pressure system by 0000 UTC 31 August.

3.3. Sensitivity Experiments

For studying the sensitivity of model prediction to the three parameters of buffer radius, data period and season length, altogether 21 experiments were performed. All these Hurricane Katrina track predictions experiments were conducted for the prediction of Hurricane Katrina starting from 1800 UTC of 27 August 2005. In the first instance, 18 sensitivity experiments were made with three choices for buffer radius as 0.5, 1.0 and 2.0 degrees; two choices for the length of the season as 3 months of July-August-September and 6 months of June – July – August – September – October - November; and length of historical data as 1842-2004, 1950-2004 and 1970-2004. The hurricane season for U.S. is noted to be for 6 months start from 1June to 30November and the 3-month period is chosen to be the month of active hurricane under study and one month before and after. The length of historical data were chosen as considering all the data (1842-2004); data since modern observations were available (1950-2004) and 30 year data period (1970-2004) preceding 2005 as 30 year data is considered as reasonable sample size for climate studies. The values for buffer radius were chosen as 0.5, 1.0 and 2.0 degrees so as to have an assessment of the sample size on homogeneity of hurricane tracks with the active hurricane. The time point of 1800 UTC 25AUG was chosen so as to obtain 48 hour track prediction up to landfall time of 1100 UTC 25AUG.

Figure 2. Hurricane Katrina Track—Google Map

Of the 18 sensitivity experiments that were conducted, the procedure was discussed for only one experiment and corresponding figures were shown. The procedural details are same for all experiments and only the results differ. Hence, the results of the experiment with buffer radius as 1.0 degree, 3-month season (July-August-September) and data period of 1950-2004 are shown. At the time of "1800 UTC 27AUGUST2005", Hurricane Katrina was located at "24.5N, 85.3W" as per the historical data source (Figure 3). As the first step, all the cyclone points within a radial distance of 1.0 degree were picked using the GIS buffer and clip tools. It was noted that a total number of "28" were identified (Figure 4a). Using GIS "near" and "dissolve" tools, duplicate points (i.e.) points of the same cyclone storm were identified and only one point nearest to the active hurricane location was retained. After this application, "14"number of points was available (Figure 4b).

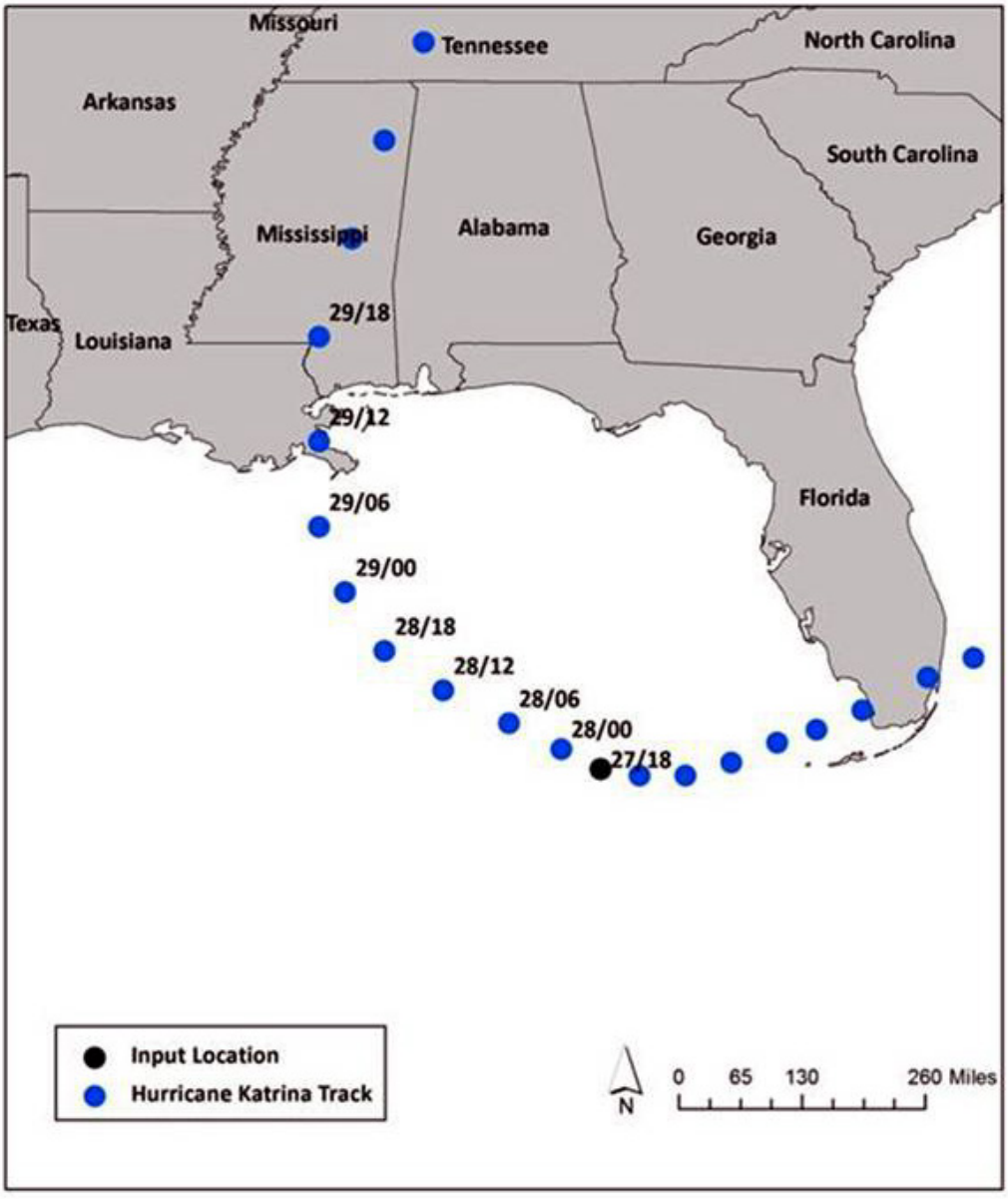


Figure 3. Hurricane Katrina track positions at different times in intervals of 6-hours. Black color circle shows the chosen time point (1800 UTC of 27 August 2005) for sensitivity experiments

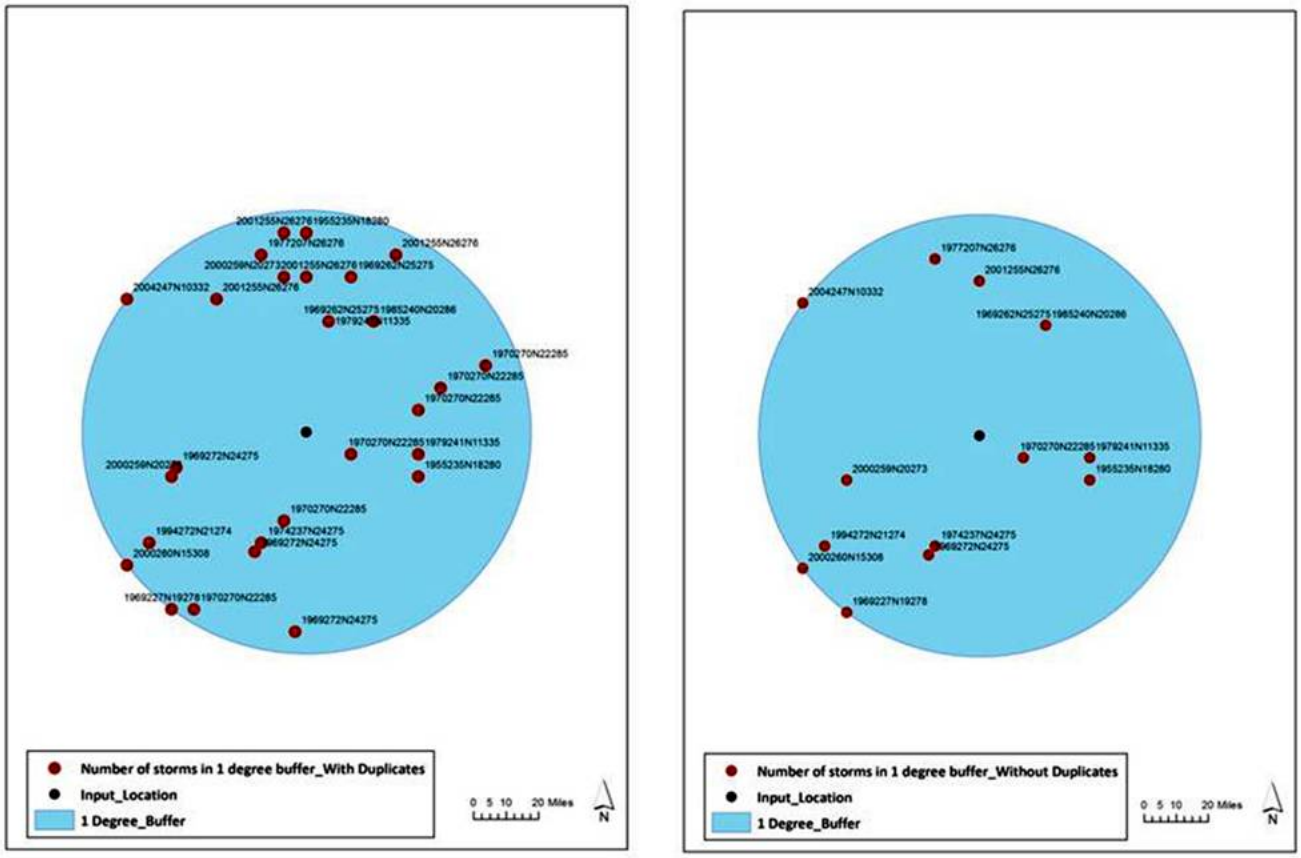


Figure 4. All hurricane locations within 1 degree buffer radius of the active hurricane point, (a) showing all points (left) and (b) after deletion of duplicate points (right)

The tracks of the cyclonic storms associated with each of these points were identified and all the subsequent time points alone were retained by discarding the prior points using SQL queries. All the individual hurricane tracks at 6-hour interval are then placed in a time sequence using a custom tool. The historical hurricane tracks as complete and subsequent to the active hurricane time point are shown in Figure 5. At this stage, all the points are checked for any outliers by imposing a check on their location to be within 20 degrees. The choice of 20 degrees is arbitrary and chosen keeping the possible distance any hurricane would travel within its life cycle from the prediction start time. "Mean Center Tool" issued to produce a geospatial mean track with location points at 6-hour interval. This is the predicted hurricane track analog (Figure 6) from the chosen time point of the Hurricane Katrina in this study. This procedure is repeated for all the 18 different combinations of experiments and the computed track distance errors are shown in Table 1. The predicted hurricane tracks from the 18 experiments are shown in Figure 7. The error statistics clearly indicate that the preceding described experiment is the best with least error of 102 and 65 miles at 24 and 48 hours respectively. This inference is based on the comparison of the errors with each class of the different sensitivity experiments. Firstly, considering the length of the hurricane season period, the errors were higher with the use of full season length of June-November as compared to selective three month period (in this case July-August-September). The errors with buffer radius of 0.5 degree with the three month period are also higher and of the same order of magnitude as with six months. The authors, through in-depth probe, identify that small sample size with 0.5 degree buffer radius is the reason for the large errors. Secondly, considering the use of the data set length, the errors were noted to be highest with the use of the full data set (i.e.) all the historical data of 1842-2004. Errors with the other two data lengths of 1950-2004 and 1970-2004 are of the nearly the same magnitude and use of the data period as 1950-2204 is noted to be better with lesser errors. Errors with respect to different buffer radii show that buffer radius of "1.0" and "2.0" degrees have errors to be in the same range whereas the errors are significantly higher with buffer radius "0.5 degrees". These observations indicate that the nature of hurricane tracks to have changes since the start of the present industrial era. Though cannot be concluded, the change in the hurricane tracks may be due to global warming due to increase of CO2 associated with industrialization. The inferior performance with buffer radius "0.5 degrees" is due to smaller sample size.

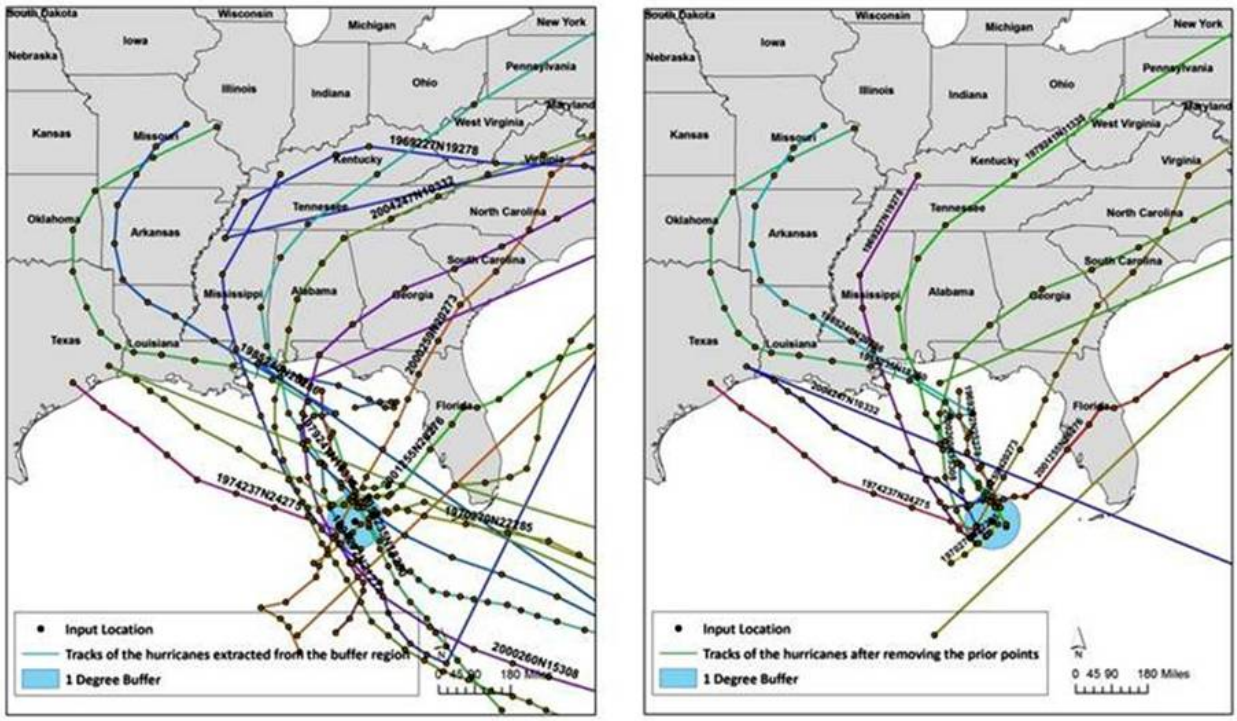


Figure 5. Hurricane tracks for the extracted locations within the buffer region, (a) left panel shows complete tracks and (b) right picture shows tracks succeeding the buffer point

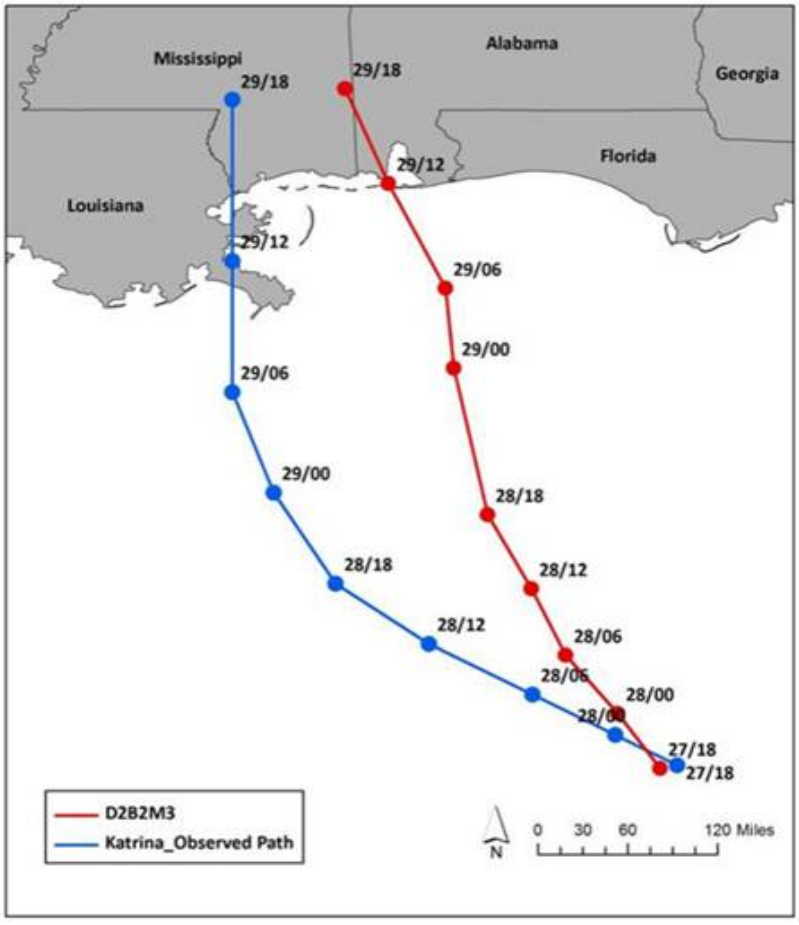


Figure 6. Observed (blue line) and GIS model predicted (red line) Hurricane Katrina track starting from 1800 UTC 27 August 2005.Track points are shown at 6-hour interval

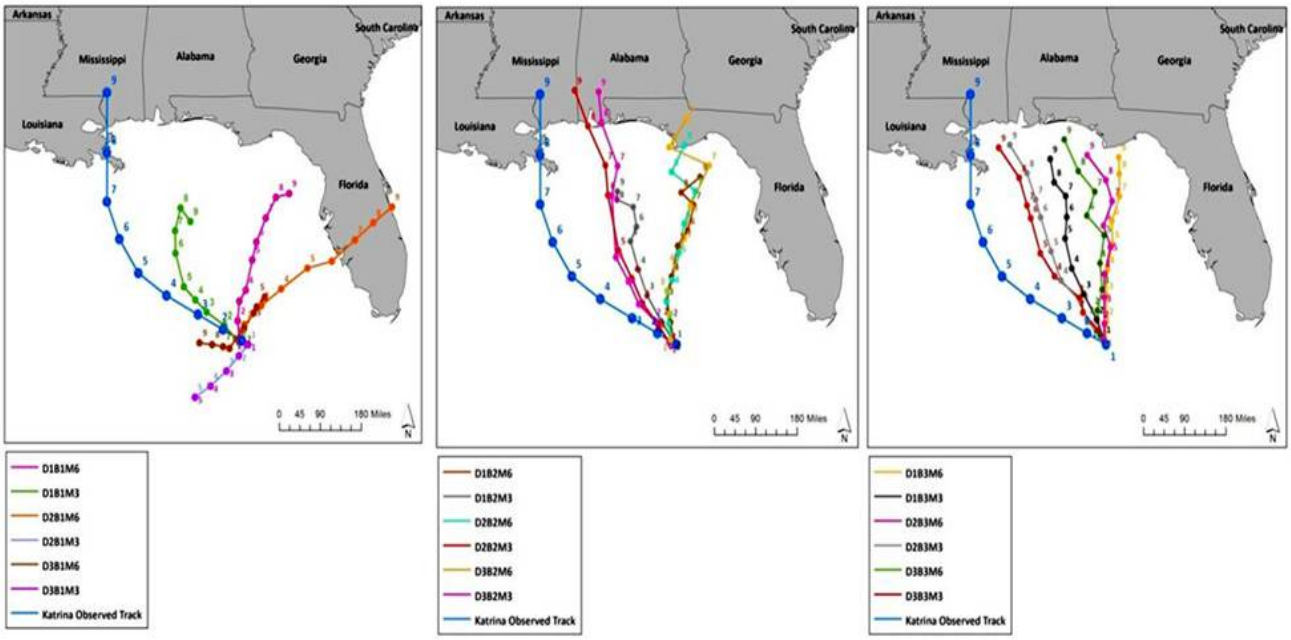


Figure 7. Hurricane Katrina observed track (blue color) and predicted tracks starting from 24.5N, 85.3W at 1800 UTC 27AUGUST2005, with buffer radius as (a) left panel: 0.5 degrees; (b) middle panel: 1.0 degrees and (c) right panel:2.0 degrees; for different values of historical data as D1=1842-2004, D2=1950-2004 and D3=1970-2004; for seasonal length as M3=July-August-September and M6=June-July-August-September-October-November

Table 1. Hurricane Katrina Track Prediction errors (miles) for the 18 experiments of sensitivity with respect to buffer radius, seasonal length and historical data period

Period (years) →	1842-2004	1842-2004	1842-2004	1950-2004	1950-2004	1950-2004	1970-2004	1970-2004	1970-2004
Buffer radius (Degrees) →	0.5	1	2	0.5	1	2	0.5	1	2
Season length (months) →	JJASON	JJASON	JJASON	JJASON	JJASON	JJASON	JJASON	JJASON	JJASON
Prediction time (hr) ↓	Track error (miles)								
0	6	6	3	4	10	11	4	15	23
6	33	33	45	44	28	39	42	41	46
12	86	75	97	127	74	92	110	85	95
18	157	146	161	228	145	153	180	159	152
24	226	214	225	334	213	220	254	230	215
30	268	262	254	417	257	237	294	275	208
36	310	295	288	486	299	274	349	332	241
42	334	276	286	526	251	262	406	245	208
48	390	334	301	577	285	245	492	278	194

Period (years) →	1842-2004	1842-2004	1842-2004	1842-2004	1950-2004	1950-2004	1970-2004	1970-2004	1970-2004
Buffer radius (Degrees) →	0.5	1	2	0.5	1	2	0.5	1	2
Season length (months) →	JAS	JAS	JAS	JAS	JAS	JAS	JAS	JAS	JAS
Prediction time (hr) ↓	Track error (miles)								
0	7	4	11	14	11	15	14	15	31
6	8	23	32	58	15	25	58	22	39
12	18	51	62	118	34	48	118	29	51
18	57	92	99	187	72	69	187	64	63
24	93	131	142	254	102	107	254	93	87
30	113	165	169		137	121		142	103
36	143	179	184		144	126		164	109
42	174	167	167		105	113		130	102
48	284	228	190		65	118		109	111

Due to difficulty to judge relative merits of the experiments with buffer radius as "1.0" and "2.0" degrees, three more experiments were conducted with buffer radius as 3.0, 4.0, and 5.0 degrees to ascertain if buffer radius higher than 2.0 degrees would lead to better track prediction and if so to optimize the value for buffer radius. All of these three additional experiments were made keeping the historical data period as 1950-2004 and seasonal length as 3 months, since these have clearly identified to be the best of choice. The predicted tracks from the six experiments with these values for historical data set and seasonal length and with six different buffer radius values of 0.5, 1.0, 2.0, 3.0 4.0 and 5.0 degrees are shown together in Figure 8 and the corresponding grouped error values in Table 2. A comparison of the results from these six experiments clearly show that buffer radius as "1.0" and "2.0" degrees show better performance than all other experiments, indicating that higher buffer radius lead to larger sample size and thus effecting the homogeneity. Higher buffer radiuses tend to have larger errors during the early period of prediction, before 24 hours and lesser errors beyond. Although "1.0" or "2.0" degrees could be taken as the optimized values for "buffer radius", keeping in view of the varying errors before and beyond 24 hours, prediction experiments of the Hurricane Katrina track starting from 9 different time points were conducted with five values for buffer radius (except 0.5 degrees), the results of which are presented in the next section.

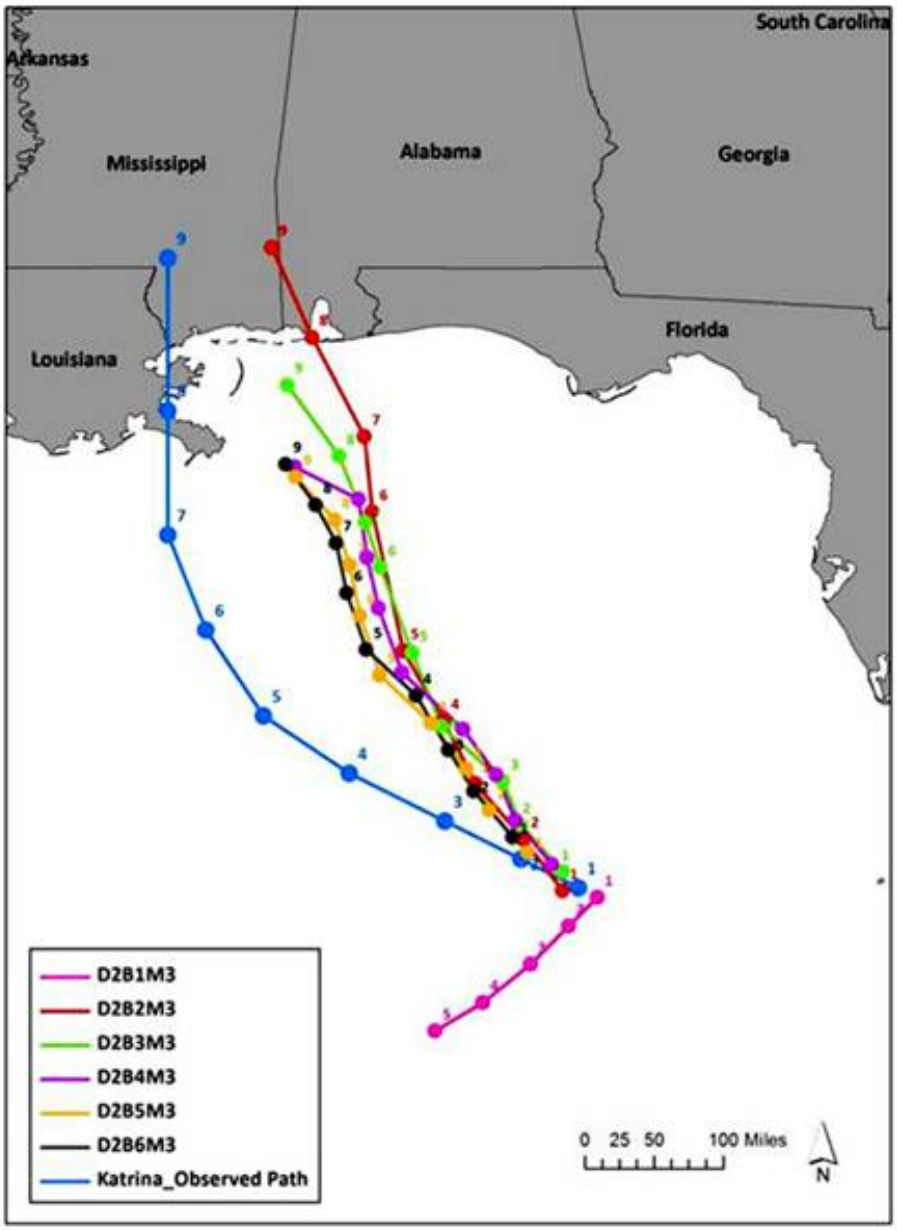


Figure 8. Hurricane Katrina observed track and predicted tracks starting from 24.5N, 85.3W at 1800 UTC 27AUGUST2005, for the experiments with different buffer radius with fixed data period as D2=1950-2004; seasonal length as M3=July-August-September. Buffer radius values are taken as B1=0.5, B2=1.0, B3=2.0, B4=3.0, B5=4.0 and B6=5.0 degrees

Table 2. Hurricane Katrina Track Prediction errors (miles) for the 5 experiments of sensitivity with different values for buffer radius

Period (years) →	1950-2004	1950-2004	1950-2004	1950-2004	1950-2004
Buffer radius (Degrees) →	1	2	3	4	5
Season length (months) →	JAS	JAS	JAS	JAS	JAS
Prediction time (hr) ↓	Track error (miles)				
0	11	15	24	42	57
6	15	25	28	41	58
12	34	48	48	40	52
18	72	69	81	65	71
24	102	107	95	81	82
30	137	121	112	99	95
36	144	126	128	118	107
42	105	113	136	131	115
48	65	118	171	177	166

3.4. Hurricane Katrina Track Prediction

Hurricane Katrina had a long life cycle with its first identification as a tropical wave on 19 August over southeast Bahamas to its final landfall on 29 August near Pearl River at Louisiana/Mississippi border. For the present study, we have chosen nine time points starting from 1800UTC 26August up to 1800UTC 28August at 6-hour time interval, such that the last prediction starting from 1800UTC 28August would provide a 24-hour prediction before the landfall. With this set up, track prediction of Hurricane Katrina using the proposed GIS model and the predicted track points were considered up to 1800UTC 29 August only as the landfall occurred around 1100UTC 29August. As mentioned in the previous section, five sets of predictions were obtained with options for the historical data period as 1950-2004, seasonal length as July-August-September, and five values for buffer radius as 1.0, 2.0, 3.0, 4.0 and 5.0 degrees.

Vector distance errors were computed as errors corresponding to different periods from 6-72 hours at 6-hour interval for each of the six experiments and presented in Table 3. It is noted that the least of the average errors are 72, 137, 188, 143 and 168 miles corresponding to 12, 18, 24, 26, 48 and 60 hours from the experiment with buffer radius as 1.0 degrees. Track errors for other experiments are noted to increase with increase of buffer radius. The predicted tracks from each of the nine time points for the best experiment only are shown in Figure 9. These values are considered reasonable in comparison with NHC (National Hurricane Center) average model hurricane track prediction errors of 46, 69, 104, 127 and 161 miles corresponding to 12, 18, 24, 26, 48 and 60 hours (NOAA, 2013). Another important fact is that dynamical model predictions from suite of models considered by NHC show a large spread with large differences in the track, as seen from model predicted tracks of Hurricane Katrina starting from 26 August (McCallum and Heming, 2006). It is also inferred, through a comparison of the GIS model produced tracks with different options with suite of dynamical model predictions, that GIS predicted tracks with different options fall well within the prediction spread of the dynamical models.

Table 3. Hurricane Katrina Track Prediction errors (miles) for the 9 experiments with prediction starting at different time points

Buffer radius (degrees) →	1.0	2.0	3.0	4.0	5.0	Ensemble Average
Prediction time (hours) ↓	Track error (miles)					
0	11	23	26	37	45	28
6	40	48	41	48	54	46
12	71	74	67	65	69	69
18	100	93	90	89	91	93
24	137	121	115	114	115	120
30	166	148	140	141	142	148
36	188	179	172	175	172	177
42	199	200	194	203	195	198
48	143	171	181	213	212	184
54	164	186	208	251	243	210
60	168	253	261	264	246	238
66	304	292	279	290	263	285
72	126	271	308	322	311	268

3.5. Prediction of Landfall Time and Location

As mentioned in the introduction, prior information of the landfall time and location are important as they provide valuable inputs to the decision support system to initiate mitigation measures regarding the annual occurring natural disaster phenomena. In view of its importance, landfall time and location were computed for the nine prediction experiments made starting from different time points in the life cycle of Hurricane Katrina for each of the two experiments with buffer radius of "1.0" and "2.0" degrees. These nine predictions could be taken as nine lead times for interpretation and use of information. The time and distance errors are shown in Table 4. The landfall errors are noted to be consistently moderate with buffer radius of 2.0 degrees than 1.0 degree radius. It is seen that the GIS model predicted the landfall point with a lag of 3-6 hours (estimated to be later than the actual occurrence) with lead time of 66-36 hours. The errors with smaller lead times have increased

owing to differences in the predicted speed of movement of Hurricane Katrina. Correspondingly, the distance errors were 233, 116, 83 and 20 miles at lead times of 60, 48, 36 and 24 hours respectively. The smaller time and distance errors with lead times of 36 and 48 hours indicate the usefulness of this model in emergency.

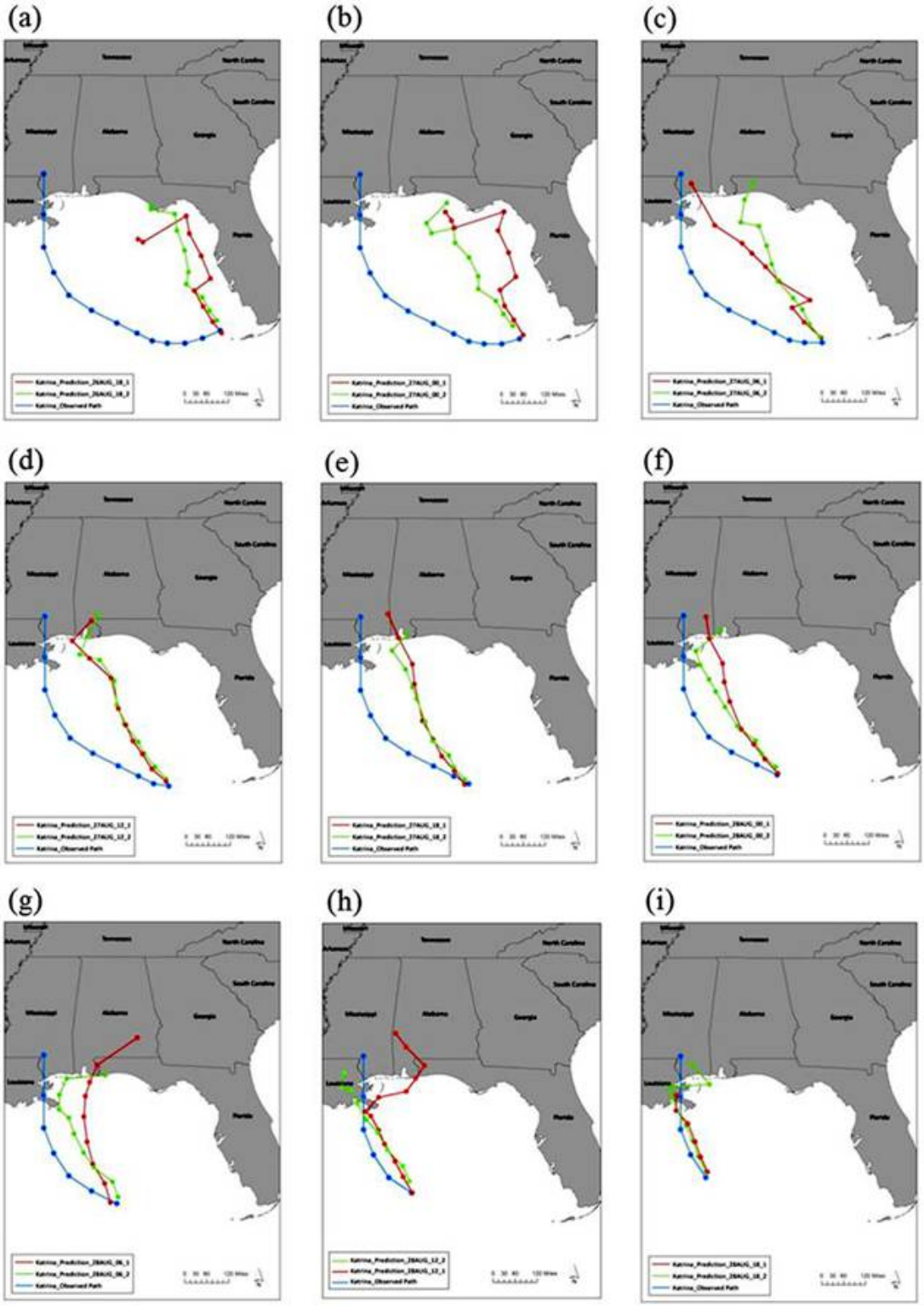


Figure 9. Observed track (blue line) and predicted tracks of Hurricane Katrina from different time points of (a) 18Z,26AUG (b) 00Z,27AUG (c) 06Z,27AUG (d) 12Z,27AUG (e) 18Z,27AUG (f) 00Z,28AUG (g) 06Z,28AUG (h) 12Z,28AUG and (i) 18Z,28AUG. Red and green lines indicate predicted tracks with buffer radius of 1.0 and 2.0 degrees respectively

Table 4. Hurricane Katrina Landfall Prediction errors (miles) for the 9 experiments starting at different time points

Experi-ment →	Starting point	Starting point location		Landfall distance error (miles)		Landfall time error (hours)	
Point ID ↓	Time	latitude	longitude	Buffer radius=1.0	Buffer radius=2.0	Buffer radius=1.0	Buffer radius=2.0
1	18Z26AUG	24.9	-82.6	266	257	-2	-1
2	00Z27AUG	24.6	-83.3	248	233	-2	+1
3	06Z27AUG	24.4	-84	74	177	3	0
4	12Z27AUG	24.4	-84.7	101	116	-1	-1
5	18Z27AUG	24.5	-85.3	107	128	+1	-1
6	00Z28AUG	24.8	-85.9	86	83	-5	-2
7	06Z28AUG	25.2	-86.7	128	74	-7	-2
8	12Z28AUG	25.7	-87.7	149	20	-5	-5
9	18Z28AUG	26.3	-88.6	15	18	-8	-5

4. Summary

This paper reveals the possibilities of hurricane track prediction over Atlantic Ocean using GIS tools. The sequential steps of the development of GIS model with different tools and applications and using historical hurricane track data of 1842-2008 were described. The model output is dependent on three different parameters of "length of input data", "length of hurricane season" and "buffer radius". Sensitivity of hurricane track prediction to these three parameters was studied through a case study of Hurricane Katrina. Results indicated that the model prediction is better with use of historical hurricane track data as of 1950-2004, hurricane season length as three months (i.e.) month of active hurricane and preceding and succeeding months and buffer radius as 1.0/2.0 degrees. The GIS model could predict the Hurricane Katrina track with errors of 46-161 miles corresponding to 12-60 hours prediction. Landfall could be estimated with a time error of 1-hour and distance error of 100 miles with a lead time of 48 hours. Although the results of this paper pertain to only one case study and require extensive validation with more number of past hurricanes, use of GIS tools and applications is amply demonstrated in the geospatial analysis and computations. Some advantages of this GIS model over the current dynamical atmospheric models are the requirement of minimal computational resources such as a desk top computer loaded with ARCGIS software and computational time of 2-3 minutes for running the entire model. This provides an opportunity to run this model at repeated intervals to update the hurricane track prediction with desirable lead times to facilitate planning and mitigation. This is a maiden attempt to apply GIS tools and applications for hurricane track prediction and has scope for further improvement through imposition of more constraints to find better homogenous historical hurricane tracks.

REFERENCES

[1] Emanuel, Kerry, 2005. Increasing Destructiveness of Tropical Cyclones over the past 30 years.Nature.436, 686-88.

[2] Knapp, K. R., M. C. Kruk., D. H. Levinson., H. J. Diamond., and C. J. Neumann., 2010. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. Bulletin of the American Meteor Society. 91, 363-376. doi:10.1175/2009BAMS2755.1

[3] Landsea, Christopher W., Gabriel A. Vecchi., Lennart, Bengtsson., Thomas R. Knutson., 2010. Impact of Duration Thresholds on Atlantic Tropical Cyclone Counts. J. Climate. 23, 2508–2519.doi: 10.1175/2009JCLI3034.1.

[4] McCallum, E., J. Heming., 2006. Hurricane Katrina: an environmental perspective Phil. Trans. R. Soc. A, 364, 2099-2115.

[5] Wilhelmi, Olga V., Jeffrey C. Brunskill., 2003. Geographic Information Systems in Weather, Climate, and Impacts. Bulletin of the American Meteor Society. 84, 1409–1414.

[6] FEMA. (2013). Applications of GIS for Emergency Management. Retrieved 2013, from www.fema.gov: http://emilms.fema.gov/is922/GISsummary.htm

[7] NOAA. (2012, Aug 2). Hurricane Preparedness - Watches & Warnings. Retrieved Oct 2013, from www.noaa.gov: http://www.nhc.noaa.gov/prepare/wwa.php

[8] NOAA. (2013). IBTrACS. Retrieved 2013, from www.noaa.gov: http://www.ncdc.noaa.gov/ibtracs/index.php ?name=ibtracs-data.

[9] NOAA. (2009, July). NHC Track and Intensity Models. Retrieved Oct 2013, from www.noaa.gov: http://www.nhc.noaa.gov/modelsummary.shtml.

[10] Waple, A. (2005). Hurricane Katrina. National Climatic Data Center.

[11] Yerramilli, A., Dodla, V., & Yerramilli, S. (2011). Air Pollution, Modeling and GIS based Decision Support Systems for Air Quality Risk Assessment. In F. Nejadkoorki, Advanced Air Pollution (pp. 295-324). InTech.

[12] Zarcadoolas, C., Boyer, J., Krishnaswami, A., & Rothenberg, A. (2007). GIS Maps to Communicate Emergency Preparedness: How Useable Are They for Inner City Residents? Journal of Homeland Security and Emergency Management, Vol 4 Issue 3.

Assessing Geographical Inaccessibility to Health Care: Using GIS Network Based Methods

Sudha Yerramilli[1] and Duber Gomez Fonseca[2]

[1]Trent Lott Center for Geospatial Technology, CSET, Jackson State University, Jackson, MS, 39204, USA
[2] College of Science Engineering and Technology, Jackson State University, Jackson, MS, 39204, USA

Abstract Disparities in the geographic accessibility to health care may be due to the location/distribution of the population and the characteristics of the transportation infrastructure relative to spatial arrangement of the health care delivery system within a region. Access to health care is a complicated concept and is largely dependent on the characteristics of the population in need of services. The most significant features affecting the health status and health outcomes involve distance between the population's geographic regions and health care facilities and the travel time taken to reach the health care delivery system. Because of Mississippi's rural nature and uneven distribution of physicians, geographic disparities exist in access to primary care services leaving women, children, elderly and general populations in underserved health care regions. The purpose of the research is to identify hot spots of vulnerable population burdened due to geographical accessibility to right kind of health services. This research investigates these features by using network-based GIS methods in ten counties with urban-rural settings. The methodology assesses the geographic accessibility of three types of critical health care facilities: obstetrician/gynaecology (Women in child bearing age); paediatrics (children) and Trauma/Burn Centers (general population). To examine, using network analyst GIS functionalities, these facilities are geocoded, and optimal travel-time based service areas were generated and pertinent vulnerable population data layers were developed. The results identified hot spots of vulnerable populations residing outside the optimal service areas, with rural regions and pregnant women bearing most of the health burden due to geographic inaccessibility. This GIS methodology equip health administrators and policy makers in providing comprehensive view of the health systems from a territorial perspective while assisting them in making conscious policy decisions.

Keywords Geographic Accessibility, Network based methods, Health Care, GIS

1. Introduction

Healthy people make health economy. Health care is the basic care provided to the populations at community level and targets better health outcomes and greater equity in health. Primary health care is the key in shaping healthy communities, improving and maintaining health of people. The health care quality is better described as a gulf for certain segments of the population, such as racial and ethnic minority groups, given the gap between actual care received and ideal or best care quality [1]. These health care disparities arise due to hitches in optimal health care functionalities such as availability, accessibility and affordability [2]. Disparities in geographic access to health care result from the configuration of facilities, population distribution, and the transportation infrastructure [3]. Geographic accessibility is considered as a critical determinant of human health around the globe (World Health Organization) than any of the vectors and significant factor contributing towards health disparities.

Access to health care is a complicated concept and is largely dependent on the characteristics of the population in need of services. The most significant features affecting the health status and health outcomes involve distance between the population's geographic regions and health care facilities and the travel time taken to reach the health care delivery system.

1.1Geographic Accessibility: Location and Health

Location and Health have been prominent features in determining accessibility rates among communities; and "location" was long considered more a determinant of health than pathogens [4]. One of the most significant factors that controls health status and largely contributes to health disparities is 'the distance to health care facilities'. The issue of equality to health care access has become a research priority in many countries [5].

The United States Department of Agriculture (USDA) has expressed concern over growing health disparities between urban and rural communities in the dimension of infant mortality, age-standardized mortality and pregnancy related mortality rates [6].

The quality of health care in rural areas with predominantly low income and minority populations largely depend on geographical access and the distance to health care facil-

ity is highly sensitive in making health care choices. Geographical accessibility is defined as the ability of obtaining health care resources that meet the health demands of the population. It infers that a community has health care accessibility if the resources meet specific characteristics such as geographic location, affordability that fit with patient's needs [7].

Geographical accessibility centres on the concept of the presence of the right type of health care service with in the optimum travel time. The concept develops from the spatial arrangement of how people/communities, facilities and transportation network are connected or configured. Analyzing this spatial configuration serves as a measurement to expose the health disparities due to lack of the 'right' health care facilities that meets the needs of local people and difficulties with increased travel times, referred as 'burden of travel', to the facilities.

1.2 Critical Health care services: Optimal Distances

Availability and accessibility to specialty health care facilities and trauma centers are the key component to US health care system because they have shown to decrease morbidity and mortality rates with their location being in optimal distances to patients needs. The issue of geographical inaccessibility to health care facilities often affects approximately 20% of nation's population in Rural America, those who live in non-metropolitan counties. The health disparities are very large, especially in terms of geographical accessibility, with sparse distribution of general hospitals and few specialty health care systems [8].

Many studies by World health researchers have exposed health disparities related to limited/timely access to required health care facilities in these under served areas [9]. A large body of literature has documented the level of geographical accessibility to three critical health care services: Obstetrician/Gynecology, Pediatrics and Trauma Centers services as primary drivers of mortality and morbidity rates and consequent health disparities [10]. Pregnant women, Children and general populations may experience greater difficulties navigating the critical health care services within designated optimal times with poor geographical accessibility. Assessing the geographical accessibility to these three critical health care services within optimal times projects the status of health care system in the region in addition to revealing the needs of medical services required by the community.

1.3 Mississippi's health care system status

The poor status of Mississippi's health care system can be revealed from the fact that all of Mississippi's 82 counties contain Designated Medically Underserved Areas as defined by the federal Health Resources and Services Administration (HRSA) [11]. Over a half of Mississippian's (54.3%) live in designated primary health professional shortage area [12]. The foundation reports that the state's burden of high mortality and morbidity rates are exacerbated due to limited geographical access to health care facilities and outpacing the national averages in the number of deaths.

The high disparity scores, related to pregnancy, child birth and prenatal care among the women in Mississippi ranks the state 47th of 50 states in births to females 15-17 years of age, 49th in child death, and 50th in low birth weight, infant mortality [10]. While the infant mortality goal for US stands at 6.0, the rate for Mississippi stands at 9.6, attributing primary cause to issues in access to health care [13]. The University of Mississippi Medical Center (UMMC), located in the state capital is the only Level 1 Trauma facility, the only burn center and the only Level III neonatal intensive care nursery in the state.

Because of Mississippi's rural nature and uneven distribution of physicians, geographic disparities exist in access to primary care services leaving women, children, elderly and general populations in underserved health care regions.

To understand and present a clear setting of health disparities due to geographical accessibility, this study, using GIS based network analyst functionalities, identifies hot spots of vulnerable populations (in need of a specialty care) residing outside the service areas of three critical health care facilities (Obstetrician/Gynecology; Trauma Centers; Pediatrics) and equip health administrators and policy makers in providing comprehensive view of the health systems from a territorial perspective while assisting them in making conscious policy decisions.

1.4 GIS Network based methods for measuring Geographic Accessibility

According to the literature, these studies on mortality rates indicate an apparent evidence of geographic effects on health and geospatial researchers have identified these patterns as an important national policy problem in public health [14][15]. The regional variations of mortality rates confirm the significance of spatial methods in measuring the accessibility patterns [16] and further supports the necessity to analyze geographical accessibility from a spatial perspective so as to reduce health disparities.

The availability of detailed spatial data coupled with the ability of Geographic Information System (GIS) to simulate real world scenarios conveys crucial information aiding the analysis with visualization reports [17] [18]. Geospatial methods offer wide range of analytical possibilities to understand the overall picture of health disparities due to geographical accessibility. The computational power of network based GIS methods integrating transportation networks, along with their attribute information such as distances, speed limits and restrictions, provides a framework to assess geographical accessibility by estimating the physical distance or travel time between a right type of health care facility and the patient in need of the service. Even though, distance and time are both important factors of accessibility, World Health Organization (WHO) recommends using travel time, rather than distance, to assess geographical accessibility.

2. Creating Network based Service areas for critical health care facilities

Defined travel time based service areas are suitable measurements for healthcare accessibility and utilization [19] [20]. This research utilizes GIS Network Analyst tools to model geographical access to specific health care service scenarios in central Mississippi 10 county region, which is characterized with urban and rural regions. Network Analyst tools provide potential for greater use of estimations of travel time that can easily calculate service areas (buffers) and include or exclude populations that exist within the defined boundaries. Creating service areas requires accurate road network data with associated information on speed limits and restrictions. The availability of populations (with demographical profiles) at a finest geographical unit (census block level) will enhance examining the geographical accessibility of services.

The framework for measuring geographical accessibility centers on two components

1. Defining optimal travel time service area for each of the three critical health care facilities by building network dataset of transportation routes in GIS
2. Identifying pertinent vulnerable populations at a census block level, outside the optimal service area for the needed critical health care.

2.1 Datasets: Road Network, Specialty Health care and population data

The road network dataset collected from Census Bureau, 2010 provides high quality, detailed road network data for all the state of Mississippi in vector GIS format and provides information on speed limits and restriction for each and every edge and node which is critical to model real time service areas based on travel time estimates. The three critical specialty healthcare facilities: Obstetrician/Gynecology, Pediatrics and Trauma Centers are collected from Mississippi Primary Health Care Association and are geocoded, based on their physical addresses, to a point layer as a part of dataset. Block level demographic census data obtained from U.S. Census bureau was used in this research to model access to health care facilities. The demographic profiles were categorized as women in childbearing age, children under age of 12 and general population.

2.2 Assumptions

When representing real world scenarios, a number of assumptions are required for modeling. Centered on numerous population-based studies [21], to model travel time, this research makes assumptions on potential unpredictable factors that influence travel and the lack of geo-referenced data. Therefore, the GIS-based scenarios created for the three critical facilities presents an average situation. This research adopts assumptions that the travel conditions are similar through out the study region and therefore the variable factors influencing the travel conditions such as weather, traffic patterns etc. are kept constant. The research is limited to data availability and data processing capabilities.

2.3 Defining optimal travel time based service areas for health care with specialty

Defining travel-time based service areas cannot be generalized (as 30 minutes or 1 hour) as every health need has an optimal time to access specialty health care for lower mortality rates. Numerous health and mortality related studies have provided the following optimal travel times appropriate for each of the three critical health care facilities that define the health care system status of any region (Table 1).

Table 1. Defining optimal travel time for critical health care facilities

Critical Health care	Optimal Travel time studies	Optimal travel time used for this research
Obstetrician/Gynecology	An optimal time from home to a needed health care of 20 minutes or more is associated with an increased risk of mortality and adverse outcomes for pregnant women [22].	20 minutes
Pediatrics	For any child health care or primary medical care, an area is designated as HPSA if no care is accessible with in 30 minutes of travel time [23].	30 minutes
Trauma Centers	In case of emergencies and access to trauma centers, an increase in optimal travel time of more than 30 minutes is associated with 1% increase in the mortality rates [24].	30 minutes

2.4 Modelling accessibility using optimal travel time service area

To create travel time based services areas for each of the critical facilities in the study region, ArcGIS Network Analyst was employed. Prior to creating services areas, the major task was to convert the road network dataset into a geospatial network dataset. Each line segment was assigned a travel time cost attributes using the line segment's length (distance), speed limit and travel impedance factors. The converted network dataset assigns travel time along a line segment based on the distance and speed limit with additional cost in minutes calculated for turn delays. The generated network dataset was loaded into Arc GIS (10.0) along with health care locations and census block level demographic data layers.

The travel-time based service areas defined for each of the critical facilities were created as polygons using 'Service area' function: a 20-minute service area for obstetrician/Gynecology, 30-minute service areas for pediatrics and 30-minute service area for all the trauma centers were generated.

To assess the level of health care access to population, each service area is clipped to the demographic spatial distribution at a census block level. The underserved census

blocks are identified by inversing the selected clipped areas of the 20-minute, 30-minute service areas through out the study region. To create population data layer for each health care, a block's population was assigned to the underserved area only when the centroid fell within the bounds of underserved area polygon. This framework involves considering: 1) Obstetrician/Gynecology: women in child bearing age (CBA) (18-55 years), 2) Pediatrics: children under 12 years, 3) trauma/burn centers: General populations for creating these population data layers.

3. Results and Discussion

Initial examination of the results indicates that there are few significant gaps in accessibility to some of the health care service. The travel time approach adopted by this study identifies population to have varied level of access to needed health services along the transportation routes. A clear distinction in geographic access to three critical health care emerged between urban and rural populations within the study region. To illustrate the hotspots, the highly populated census blocks outside the optimal travel times are categorized under various levels.

The results are presented under the following sections to discuss the geographical accessibility of population to the critical care facilities.

3.1 Geographical access to Obstetrician/Gynecology

Figure 1 shows the spatial distribution of obstetrician/Gynecology health care facilities and the proximity area that can be accessed with in 20-minute to the facility by road.

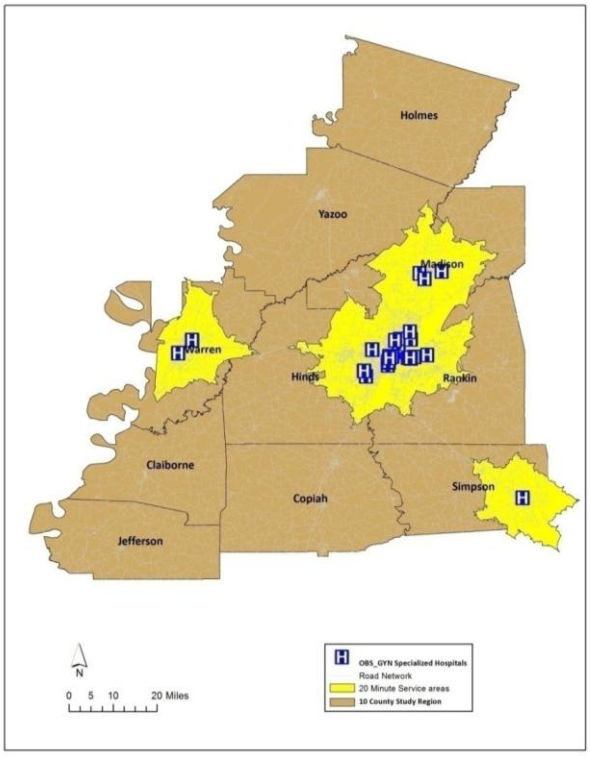


Figure 1. Spatial distribution of obstetrician/gynecology health facilities

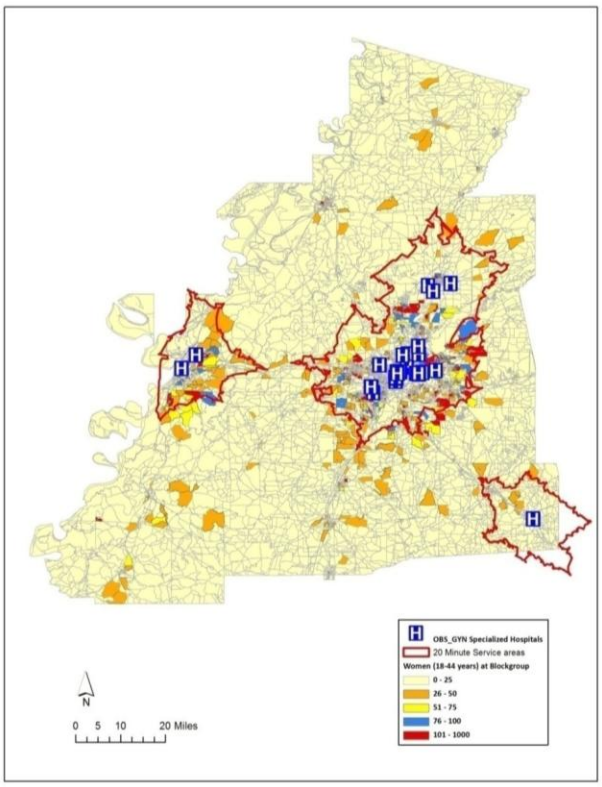


Figure 2: Optimal 20-minute health service areas and spatial distribution of women in childbearing age

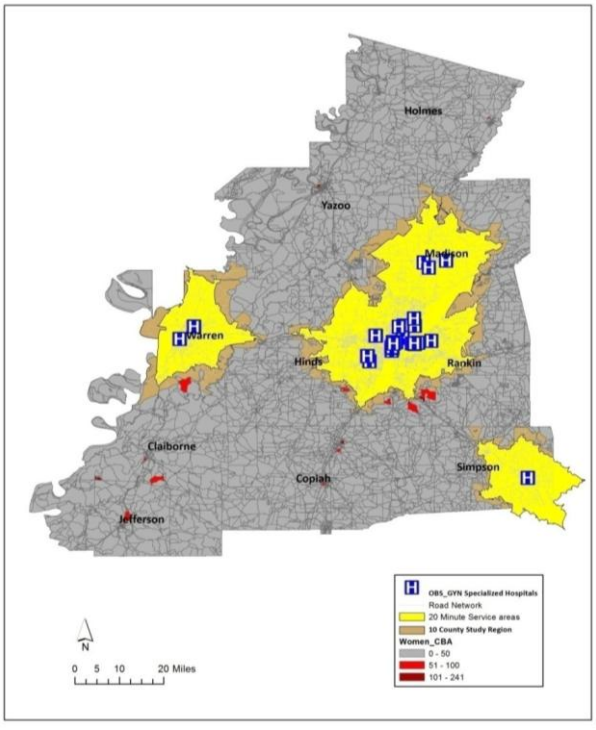


Figure 3: Hotspots of concentrations of women in childbearing age outside the optimal service regions

There are 26 OBS/GYN facilities, with uneven geographical distribution, mostly concentrated in Hinds, Rankin and Madison counties (Figure 1). The population data layer shows that around 30% of the total women in childbearing age are residing outside the optimal travel times, with rural counties taking most (80%) of the burden (Figure 2). Clai-

borne County designated as a Rural County according to Census, bears the block with maximum number (>275) of women in childbearing age with limited geographical accessibility to Obstetrician/Gynecology centers (Figure 3).

The low mortality rates associated with optimal travel times makes these rural areas vulnerable. With the underserved rural population clearly exceeding the urban population numbers, the rural women often take the burden of low mortality and high infant death rates with poor geographic access. A total of 10% of the childbearing age women is identified to be located in hot spots, with high populations concentrations and low geographical accessibility (Figure 3 and Table 2).This geographical inaccessibility, to a large extent, affects how rural residents view the long-term sustainability of their communities.

Table: 2 Urban and Rural concentrations of population outside optimal service areas.

Total number of OBS/GYN Clinics	**Total Women CBA outside 20min**		**Blocks with women in CBA outside 20-minute optimal service area**	
	Urban	Rural	Urban	Rural
26	4897	21788	1215	1720

3.2 Geographic access to Pediatrics

Figure 4 shows the spatial distribution of pediatric health centers and a service area of optimal 30-minutes travel time from each facility (Figure 4).

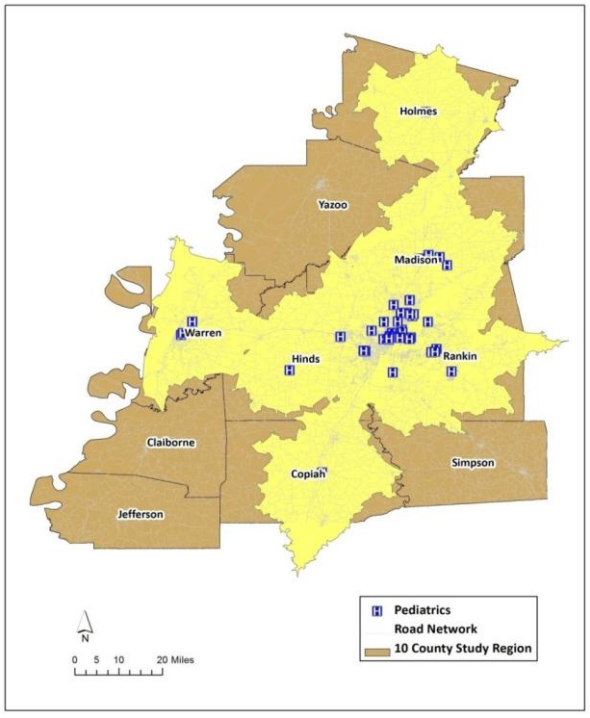


Figure 4: Spatial distribution of Pediatric health facilities

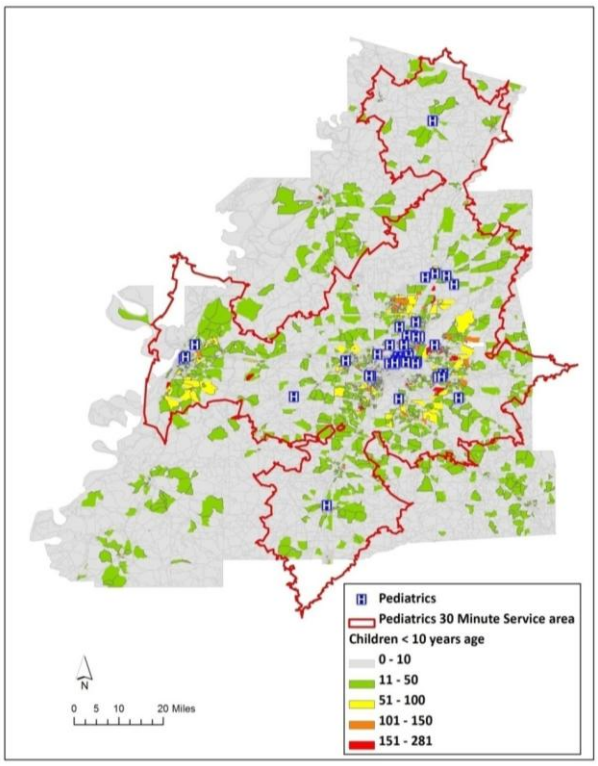


Figure 5: Optimal 30-minute service areas and spatial distribution of children

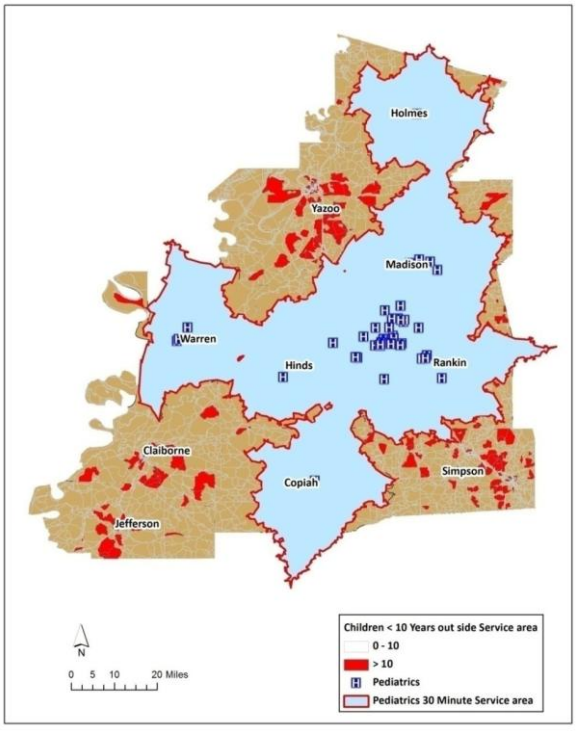


Figure 6:Hotspots of concentrations of childresn outside the optimal service regions

There are 48 Pediatric specialized clinics. The study area has an even spatial distribution of the facilities covering majority of the population both in urban and rural regions. The results reveal that more than 90% of the children are living within the optimal 30-minute travel times (Figure 5). However, the population data layer illustrated some hotspots

outside optimal travel times showing blocks with high children population concentrations and poor accessibility to health care (Figure 6). Interestingly, most of these hot spots are found in urban counties presenting information to alert health officials in decision-making process (Table 3).

Table: 3 Urban and Rural concentrations of population outside optimal service areas.

Total number of Pediatric Clinics	**Total Children outside 30min 10660**		**Block groups with Children and outside optimal 30 min service area**	
	Urban	Rural	Urban	Rural
48	2965	7695	1500	900

3.3 Geographic access to Trauma and Burn Centers

Four levels of trauma centers (I, II, III, and IV) and burn centers are geocoded and mapped in the study region (Figure 7). A 30-minute service area generated using transportation network routes is presented in Figure 8.

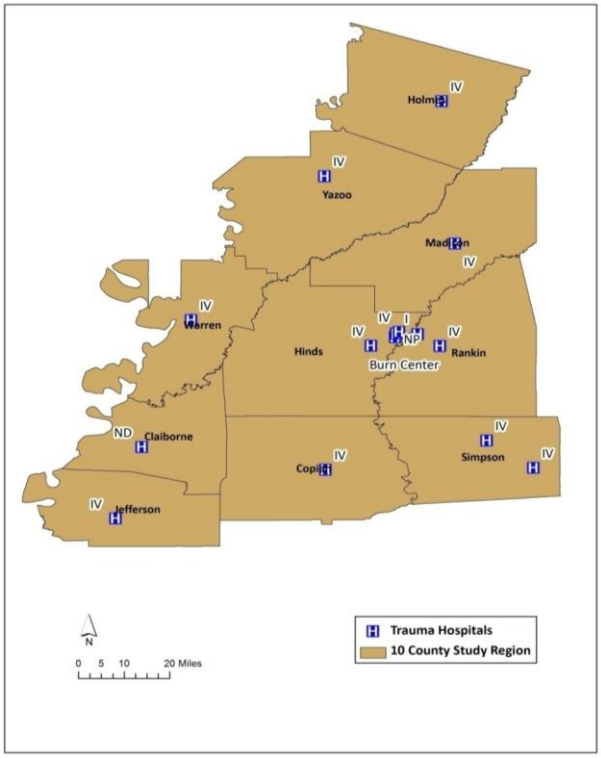


Figure 7:Spatial Distribution of Truama and burn centers

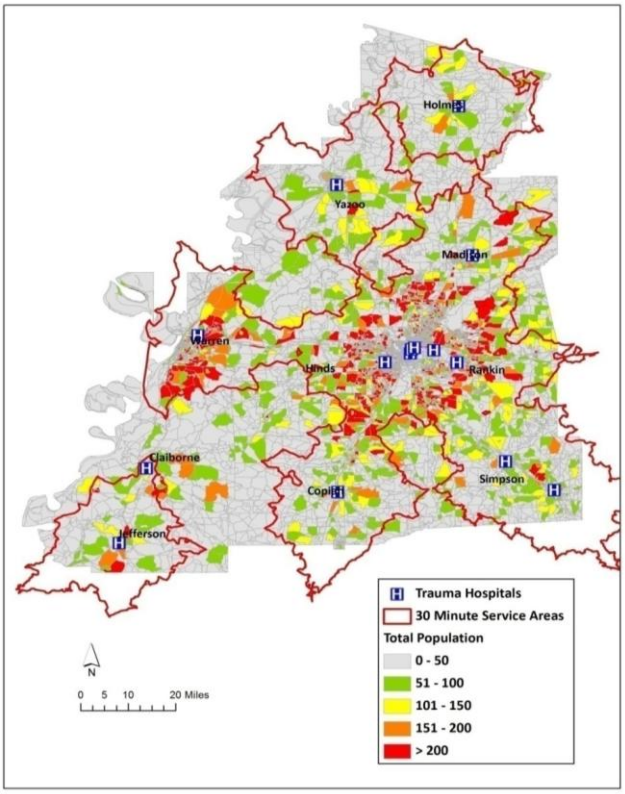


Figure 8: Optimal 30-minute service areas and spatial distribution of general population

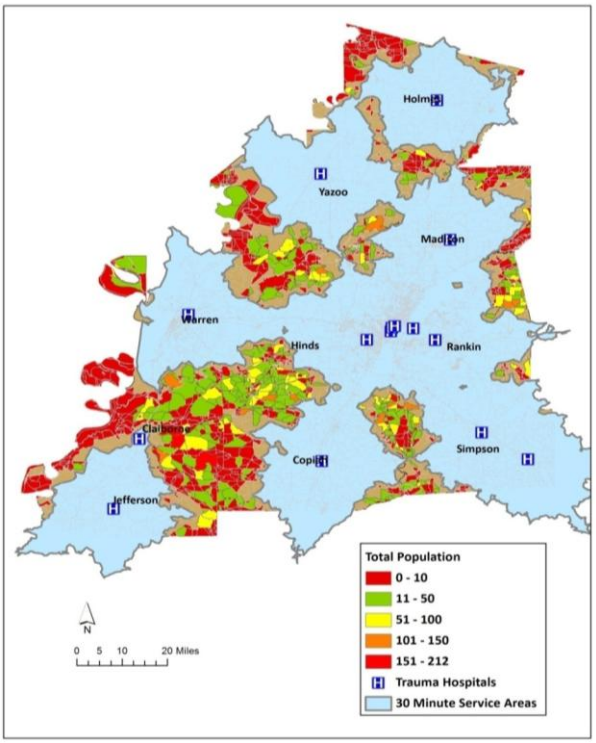


Figure 9: Hotspots of concentrations of general population outside the optimal service regions

There are 16 trauma centers and 1 burn center located in the study region. Even though the spatial distribution of these facilities appears to be evenly distributed (Figure 9), the GIS network analyst generated 30-minute service areas revealed

the real scenario of inaccessible areas to these emergency facilities. Rural counties host almost all of the hotspots; falling outside the optimal travel times, taking the total burden of geographic inaccessibility issues (Table 4).

With only one burn center, the removal of this facility leaves the state of MS vulnerable putting the entire population under no service zone and looking to neighbouring states for service.

Table 4: Urban and Rural concentrations of population outside optimal service areas.

Total number of Trauma centers (Level I-IV)	**Total Population outside optimal 30min service area**	
	Urban	Rural
16	0	25448

This methodology provides an advantageous approach in assessing community's health care needs from a geographical perspective. Geographical inaccessibility is clearly seen among rural populations. Pertinent to the relative health care facility type the model extracts the spatial locations of: women in the childbearing age; children under the age of 10 years; and population above 65 years. The results identify hot spots of vulnerable populations residing outside the service areas and equip health administrators and policy makers in providing comprehensive view of the health systems from a territorial perspective while assisting them in making conscious policy decisions. Assessing the geographic inaccessibility from various health specialty perspectives provides the requirement of right kind of services to the right kind of people in need.

4. Conclusion

The most significant features affecting the health status and health outcomes involve distance between the population's geographic regions and health care facilities and the travel time taken to reach the health care delivery system. This research investigates these features by using network-based GIS methods. Spatial distribution of health care facilities in terms of size/number, type and location are geographically analyzed to observe the feasibility of individual's access to desired services, which in turn impacts overall wellbeing of the communities. This study presents the use of GIS in analyzing health care needs and optimal geographical access in evaluating and planning future health care locations; and provides spatial decision support information for an efficient health care system.

This study has demonstrated the power of GIS in modeling the travel times for health care facilities and spatially evaluating the hotspots where people take the burden of mortality rated due to geographic inaccessibility. The geospatial simulation of variations in optimal time when specific health services are required provides crucial information, especially planning for non-urban areas.

REFERENCES

[1] Mayberry, R. M., David , N. A., Qin, H., & Ballard, D. J. (2006). Improving quality and reducing inequities: a challenge in achieving best care. Proceedings of Baylor Health Care System.19(2), p. 103. Dallas, TX: Baylor University. Medical Center.

[2] Steinwachs, D. M., & Hughes, R. G. (2008). Health Services Research: Scope and Significance. In H. RG, Patient Safety and Quality: An Evidence-Based Handbook for Nurses (p. 8). Rockville, MD, USA: Agency for Healthcare Research and Quality.

[3] Delamater , P., Messina , J., Shortridge , A., & Grady, S. (2012). Measuring geographic access to health care: raster and network-based methods. International Journal ofHealth Geographics, 11 (1), 15.

[4] Ricketts, T. C. (2002). Geography and Disparity in Health. University of North Carolina at Chapel Hill, Cecil G. Sheps Center for Health Services Research. Chapel Hill: Institute of Medicine of the National Academies.

[5] Haggertya, J. L., Robergeb,, D., Lévesque, J.-F., Gauthier, J., & Loignon, C. (2014). An exploration of rural–urban differences in healthcare-seeking trajectories: Implications for measures of accessibility. Health & Place, 28, 92-98.

[6] Carol Jones, T. P. (2009, August 10). Health Status and Health Care Access of Farm and Rural Populations. Economic Information Bulletin No. (EIB-57) , p. 72.

[7] J.F. Levesque, M. H. (2013). Patient-centred access to health care: conceptualising access at the interface of health systems and populations. International Journal of Equity Health, 12, 16-28.

[8] Duke Center for Health Policy and Inequalities Research. (2011, September 2). Geographich Access Barriers. Retrieved May 19, 2014, from U.S.Health Policy Gateways: http://ushealthpolicygateway.com/vi-key-health-policy-issues-financing-and-delivery/k-barriers-to-access/geographic-access-barriers/

[9] Thomas C. Ricketts. (2002). Arguing for Rural Health in Medicare: A Progressive Rhetoric for Rural America. North Carolina Rural Health Research and Policy Analysis Center, The University of North Carolina at Chapel Hill. Chapel Hill: Cecil G. Sheps Center for Health Services Research.

[10] Center for Health Policy Research. (2009). PUTTING WOMEN'S HEALTH CARE DISPARITIES ON THE MAP: Examining Racial and Ethnic Disparities at the State Level. University of California. Los Angeles: The Henry J. Kaiser Family Foundation.

[11] U.S.Department of Health and Human Services. (2014, 5 23). Shortage Areas: HPSA by State & County. Retrieved 5 23, 2014, from HRSA: http://hpsafind.hrsa.gov/HPSASearch.aspx

[12] Kaiser Family Foundation StateHealthFacts. (2010, April). Mississippi: Health Professional Shortage Areas. Retrieved May 23, 2014, from State Health Facts: http://www.statehealthfacts.org/profileind.jsp?cat=8&sub=156&rgn=26

[13] Mississippi State Department of Health. (2011). 2011 Mississippi Infant Mortality Report. Jackson: Mississippi State Department of Health.

[14] Pickle, L. W. (1996). Atlas of United States Mortality. U.S. Department of Health and Human Services . Hyattsville: U.S. Department of Health and Human Services .

[15] Kindig, D. A. (2002, Apr 10). Death rate variation in US subpopulations. Bulletin of the World Health Organization , pp. 9-15.

[16] Cossman, R. E. (2003). Mapping high or low mortality places across time in the United States: a research note on a health visualization and analysis project. Health Place , 9 (4), 361-369.

[17] Yerramilli, A., Dodla, V., & Yerramilli, S. (2012). GIS based Decision Support Systems for Air Quality Risk Assessment. In F. Nejadkoorki, Advanced Air Pollution (p. 307). Craotia: Intech Open Science.

[18] Yerramilli, S. (2013). Potential Impact of Climate Changes on the Inundation Risk Levels in a Dam Break Scenario. ISPRS International Journal of Geo-Information , 110-134.

[19] Lin G, A. D. (2002). Examining distance effects on hospitalizations using GIS: a study of three health regions in British Columbia, Canada. Environment and Planning A, 34, 2037-2053.

[20] Birkin M, C. G. (2005). GIS for business and service planning. In G. M. Longley PA, Geographical Information Systems: Principles Techniques Management and Applications. New York: John Wiley & Sons.

[21] Witlox, F. (2007). valuating the reliability of reported distance data in urban travel behaviour analysis. Journal of Transport Geography , 172-183.

[22] Ravelli, A. C. (2011). Travel time from home to hospital and adverse perinatal outcomes in women at term in the Netherlands. BJOG: An International Journal of Obstetrics & Gynaecology , 457-465.

[23] Eda Unal, S. C. (2008, 11 01). Health care access in Indiana. Retrieved 05 28, 2014, from Purdue University: http://www.pcrd.purdue.edu/documents/publications/PCRD-R-5.pdf

[24] Jon Nicholl, J. W. (2007). The relationship between distance to hospital and patient mortality in emergencies: an observational study. Emergency Medicine Journal , 665-668.

2014

MISSISSIPPI STATE DEPARTMENT OF HEALTH RESOURCE GUIDE
DEFINE, LOCATE, AND REACH VULNERABLE, AND AT-RISK POPULATIONS IN AN EMERGENCY

Wayne Vaughn, Vickie Perry, Sudha Yerramilli, Dayakar Philip Nittala and Albert Williams

MSDH and Jackson State University

3/19/2014

Public health consequences may be direct or indirect and can affect both a local population's health and its health infrastructure. The direct consequences of a public health disaster are counted in the number of injuries and fatalities occurring as a result of the incident. Indirect public health consequences can include exacerbation of mental and chronic health conditions (such as asthma, chronic heart disease, depression, and diabetes) or injuries sustained while cleaning up after an incident. Disasters can take many different forms, and the duration can range from an hourly disruption to days or weeks of ongoing destruction. Hurricanes and tropical storms are among the most powerful natural disasters because of their size and destructive potential. Tornadoes are relatively brief but violent, potentially causing winds in excess of 200 mph. Both earthquakes and tornadoes strike suddenly without warning. Flooding is the most common of natural hazards, and requires an understanding of the natural systems of our environment, including floodplains and the frequency of flooding events.

State Wide Description

According to the U.S. Census Bureau, there are approximately 2,950,000 people living in the state's 46,856 square miles (117,573 km), of which, 40,374 square miles (101,292 km) of it is land and 1,494 square miles (3903.79 km) of it is water. The state is primarily rural with 82 counties divided into 9 districts with the largest districts being districts II, V, and IX. The total counties with the corresponding districts are given below.

District 1:

Coahoma County, Desoto County, Grenada County, Panola County, Quitman County, Tallahatchie County, Tate County, Tunica County, and Yalobusha County

District II:

Alcorn County, Benton County, Itawamba County, Lafayette County, Lee County, Marshall County, Pontotoc County, Prentiss County, Tippah County, Tishomingo County, and Union County

District III:

Attala County, Bolivar County, Carroll County, Holmes County, Humphreys County, Leflore County, Montgomery County, Sunflower County, and Washington County

District IV:

Calhoun County, Chickasaw County, Choctaw County, Clay County, Lowndes County, Monroe County, Noxubee County, Oktibbeha County, Webster County, and Winston County

District V:

Claiborne County, Copiah County, Hinds County, Issaquena County, Madison County, Rankin County, Sharkey County, Simpson County, Warren County, and Yazoo County

District VI:

Clarke County, Jasper County, Lauderdale County, Leake County, Neshoba County, Newton County, Scott County, and Smith County

District VII:

Adams County, Amite County, Franklin County, Jefferson County, Lawrence County, Lincoln County, Pike County, Walthall County, and Wilkinson County

District VIII:

Covington County, Forrest County, Greene County, Jefferson County, Jones County, Lamar County, Marion County, Perry County, and Wayne County

District IX:

George County, Hancock County, Harrison County, Jackson County, Pearl River County, and Stone County

Natural Hazards

Earthquakes:

Earthquakes occur as a result of the violent shifting of plates under the earth's surface. The shifting of these plates releases stress along geologic faults. Earthquakes are more difficult to predict than other natural hazards and may strike at any moment without warning. This makes earthquakes very dangerous. The numbers of earthquakes known to have been centered within Mississippi boundaries are relatively small allowing little data to be reported about this particular hazard for the state of Mississippi. However, Mississippi has been affected by shocks that occurred in neighboring states.

Flooding / Flash Flooding:

Flooding occurs once the water level in a given area reaches a point where the area that is normally dry land becomes submerged. Floods are dangerous and happen quickly. This leads to major damage and can cause a significant loss of lives if at-risks population in these areas is not addressed properly for such an emergency.

Hurricanes and Coastal Storms:

Hurricanes are weather systems that develop near the equator of the earth and will usually take a path that will lead it north of its point of origin. Hurricanes are grouped into five categories according to the wind speed of the storm.

Thunderstorms and Tornadoes:

Thunderstorms are derived from an air mass that becomes so unstable that it overturns violently. Depending upon its severity, thunderstorms include rain showers that may cause flash floods, lightning, hail, tornadoes and wind gusts of 58 mph or more. Tornadoes often viewed as the most violent storms are derived from severe thunderstorms, but may also form when very warm, moist air rises into cold air. Tornadoes are not restricted to any geographical location; however it is apparent that some parts of the Mississippi are more susceptible to tornadoes than others. The most damaging of these storms spawn from super cells or violent rotating columns of air that forms as a funnel cloud and reaches the ground. When this condition occurs the tornado's wind speeds can cause minimal or massive amounts of damage especially to crops and structures.

Past History of Public Health Emergencies

The chart below lists public health emergencies caused by naturally occurring events, technological hazards or human-related events that have previously occurred in the county.

Incident	Extent of Injuries, Deaths or Damages	District Affected
River Flood	$1.6 million Mitigation Grant pending, $2 million property and infrastructure damage	District 1
MS River Flood	$33.5 million (estimate)	District 3
River Flood	$2,000,000 Property Damage, 2 Fatalities	District 5
Flood	$1.0 Million Property Damage (figure includes 4 counties)	District 9
Flood	$1,500,000 Property Damage	District 5
Flood	$10,065,000 Property Damage, 5 Fatalities,	District 1
Flood	1 Fatality, $220,000 Property Damage	District 2
Flash Flood	$5,562,000 Property Damage, 1 Fatality, 80 displaced for 2-3 days	District 1
Flash Flood	$3,400,000 Property Damage	District 2
Flash Flood	$3,545,000 property damage, $2,150,000 crop damage	District 3
Flash Flood	$3,825,000 Property Damage	District 4
Flash Flood	$6,170,000 Property Damage, $2,100,000 Crop Damage	District 5
Flash Flood	1,575,000 Property Damage	District 6
Flash Flood	$3,000,000 Property Damage, $100,000 Crop Damage	District 7

Incident	Extent of Injuries, Deaths or Damages	District Affected
Flash Flood	$2,385,000 Property Damage	District 8
Flood/Flash Flood	$900,000 Property Damage (figure includes 3 counties) $555,000 Property Damage	District 9
Winter Storm	$756,000 Property damage	District 1
Winter Storm	$517,000 Property Damage	District 2
Winter Storm	No injuries/fatalities, Damage estimate not available	District 3
Winter Storm	$500,000 Property Damage	District 4
Ice Storm	>$666,514 damages, 22 days w/o power to some	District 1
Ice Storm	$99,000 property damage	District 2
Ice Storm	$1.3 Million Property Damage	District 3
Ice Storms	$4,098,000 Property Damage	District 4
Ice Storm	$1,300,000 Crop Damage	District 5
Ice Storm	1.3 Million Property Damage	District 6
Ice Storm	$3.0 Million Property Damage	District 7
Hail/ Straight Line Wind	$609,866 Property Damage, $25,000 Crop Damage	District 1
Hail	$500,000 Property Damage	District 2
Hail	$5.0 Million Property Damage, $1,000,000 Crop Damage	District 3
Hail	$540,000 Property Damage, 50,000 Crop Damage	District 4
Hail	$850,000 Property Damage	District 5

Incident	Extent of Injuries, Deaths or Damages	District Affected
Hail Storm	300,000 Property Damage	District 6
Hail	$650,000,000 Property Damage	District 7
Hail	$1,325,000 Property Damage	District 8
Lightning	$260,000 Property Damage	District 1
Lightning	$350,000 property damage	District 2
Lightening	$70,000 Property Damage	District 4
Lightening	$570,000 Property Damage,	District 5
Lightening	100,000 Property Damage	District 6
Lightning	$280,000 Property Damage	District 7
Lightening	$300,000 Property Damage	District 8
Tornado	$93,854,000 Property Damage; 45 Injuries, 3 Fatalities	District 1
Tornado	$36,965,075 property damage, 7 Fatalities, 73 Injuries	District 2
Tornado	$11,490,000 property, $2,350,000 crop damage, 37 Injuiries, 1 Fatality	District 3
Tornado	$32,616,953 Property Damage, 2.0 Million in Crop Damage, 163 Injuries, 26 Fatalities	District 4
Tornado	$53,712,000 Property Damage, $1,200,000 Crop Damage, 4 Fatalities, 79 Injuries	District 5
Tornado	6,700,000 Property Damage 751,000 Crop Damage 2 fatalities 52 homes destroyed	District 6
Tornado	$1,650,000 Property Damage,	District 7

Incident	Extent of Injuries, Deaths or Damages	District Affected
	$500,000 Crop Damage	
Tornado	$6,225,000 Property Damage	District 8
Tornado	$3.0 Million Property Damage	District 9
High Winds	$510,000 Property Damage	District 1
High Winds	$116,000. 1 Fatality	District 2
High Winds	$800,000 Property Damage	District 3
High Winds	$530,000 Property Damage	District 4
High Winds	$7,820,000 Property Damage, 1 Injury, 4 Fatalities	District 5
High Winds	1 Fatality, 3.6 Million Property Damage	District 6
Strong Wind	$150,000 Property Damage (figure also includes Walthall County)	District 7
Strong Wind	$5,000 Property Damage	District 9
Thunderstorm	$1,445,000 Property Damage, 2 fatalities	District 1
Thunderstorm	$643,000 property damage	District 2
Thunderstorm	$6,373,000 Property Damage, $210,000 Crop Damage, 1 Fatality	District 3
Thunderstorms	$3,100,000 Property Damage	District 4
Thunderstorm	$66,505,000 Property Damage, $300,000 Crop Damage	District 5
Thunderstorm	2,365,000 Property Damage	District 6
Thunderstorm	$75,000 Property Damage	District 8

Incident	Extent of Injuries, Deaths or Damages	District Affected
Thunderstorm Winds	$805,000 Property Damage, $150,000 Crop Damage	District 7
Thunderstorm Winds	$2,620,000 Property Damage, $100,000 Crop Damage	District 8
Thunderstorm Wind	$ 20 Million Property Damage	District 9
Hurricane	2142 Fatalities, $47.7 Trillion Property Damage, $13.5 Trillion Crop Damage (All Numbers are Mississippi wide)	District 1
Hurricane	238 Fatalities, $5.9 Billion Property Damage, $1.5 Billion Crop Damage (All Numbers are Mississippi wide)	District 3
Hurricane	238 Fatalities, $5.9 Billion Property Damage, $1.5 Billion Crop Damage (All Numbers are Mississippi wide)	District 4
Hurricane	$5,900,000 Property Damage, $1,500,000 Crop Damage, 238 Fatalities (Both covering over 12 counties),	District 5
Hurricane	238 Fatalities, $5.9 Billion Property Damage, $1.5 Billion Crop Damage (All Numbers are Mississippi wide) $2.6 Million Property Damage, $2.2 Million Crop Damage (All Numbers are Mississippi wide)	District 6
Hurricane	$7.4 Billion Property Damage (figure includes eight counties), $1.5 billion crop damage, figure includes 48 counties, $1,070,000 Infrastructure; 15 Deaths, 104 Injured	District 7
Hurricane	15 Dead, 104 Injured, $5.9 Billion Property Damage, $1.5 Billion Crop	District 8

Incident	Extent of Injuries, Deaths or Damages	District Affected
	Damage (figure includes 48 counties)	
Hurricane	$250 Million Property Damage (figure includes 5 counties)	District 9
Hurricane/Typhoon	$7.4 Billion Property Damage (figure includes 8 counties)	District 9
Drought	$200,000 Crop Damage	District 1
Drought	$1,720,000 Crop Damage, $350,000 Property Damage	District 3
Drought	$4,150,000 Crop Damage	District 4
Drought	$6,000,000 Crop Damage	District 5
Drought	2,500,000 Crop Damage	District 6
Drought	$1,050,000 Property Damage, $1,300,000 Crop Damage (figure includes 42 counties)	District 7
Drought	$650,000 Crop Damage	District 8
Helicopter Crash	No Deaths, 2 severe injuries, diesel fuel clean-up	District 1
School Bus Accident	30 injuries, 1 fatality	District 3
Heavy Snow	$3,500,000 Property Damage	District 3
Heavy Snow	$1,100,000 Property Damage	District 4
Heavy Snow	$500,000 Property Damage	District 8
Ice Storm	$99,000 property damage	District 2
Ice Storm	$360,000 Property Damage	District 3
Ice Storms	$4,098,000 Property Damage	District 4

Incident	Extent of Injuries, Deaths or Damages	District Affected
Ice Storm	$1,300,000 Crop Damage,	District 5
Ice Storm	1.3 Million Property Damage	District 6
Ice Storm	$3.0 Million Property Damage	District 7
Tropical Storm/Katrina	Wind Damage	District 2
Tropical Storm	$500,000 Property Damage	District 3
Tropical Depression	$200,000 Property Damage, 1 Fatality	District 4
Tropical Storm	$13.4 Million Property Damage (figure includes eight counties)	District 7
Tropical Storm	$50,000 Property Damage	District 8
Tropical Storm	$9.0 Million Property Damage (figure includes 4 counties)	District 9
Fire/Explosion	Anel Engineering – Propane Explosion $500,000 Damage	District 3
West Nile – Mosquitoes		District 3
Gas Leak	Tennessee Gas Plant Grille, MS – Partial Evacuation of City	District 3
Transportation	MDOT bridge repair, 3 Fatalities, 5 Injuries	District 4
Straight Line Winds	$100,000 property damage	District 2
Straight line Winds	200,000 property damage	District 7
Hazmat	Tanker stuck by train & Gasoline Rollover, Evacuation of 200 people, 30 Injuries, 1 Fatality	District 5
Snow & Ice	$400,000 Property Damage	District 5

Incident	Extent of Injuries, Deaths or Damages	District Affected
Temperature Extreme	77,000 Crop Damage	District 6
Downburst	60,000 Property Damage	District 6
Extended Cold	$100,000 Property Damage (figure includes eight counties)	District 7
Heat	$77,000 Crop Damage	District 7, District 8
Williams Pipeline	N/A	District 8
Chemical Plant Explosion	3 Injured. Damage to the First Chemical Corporation Plant and an Adjacent Facility	District 9
Storm Surge	$11.3 Billion Property Damage (figure includes 3 counties)	District 9

1. Introduction to Public Health Hazard Vulnerable Analysis

The Mississippi State Department of Health (MSDH) promotes the functional skills of persons, who have disability or at risk vulnerable population through Public Health Vulnerability Analysis (HHVA). This resource guide is designed to help local health departments and districts improve their public health preparedness planning by including vulnerable and at-risk populations. This resource guide will identify and assess how public health services may address at-risk populations. Specifically, this can identify and locate vulnerable and at-risk populations and improves preparedness planning using population specific resources. To reach every person is one of the major goals for emergency preparedness and response.

1.1 At-Risk population

The term at-risk populations is used to describe individuals or groups whose needs are not fully addressed or who feel they cannot comfortably or safely use the standard resources offered during preparedness, response, and recovery efforts. This document describes a process that will help planners to define, locate, and reach at-risk populations in an emergency. GIS maps and tools are used to provide resources in planning that would offer time saving assistance for public health and emergency management planners in their efforts to reach at-risk populations.

List as vulnerable populations:

- Special Medical Needs Populations
- Children
- Community based technology dependent clients (e.g. life support equipment, oxygen, etc)
- Developmentally Disabled Clients to include Children (e.g. independent supported living, small group homes)
- Dialysis Clients
- Disabled populations (sensory, physical, mental)
- Economically disadvantaged populations
- Elder populations
- Migrant populations
- Non-English speaking populations

- Populations residing in residential shelters (e.g., battered spouses, homeless, etc.)
- Pregnant women
- Specialty care populations (e.g. radiation/oncology clinics, methadone clinics, institutional settings, etc)

Classifying and grouping the at-risk populations into very broad categories can be an effective and allows public health personnel to examine the vulnerability and resources in an emergency. These groups include people who are physically or mentally disabled (e.g., blind, deaf, hard-of-hearing, have learning disabilities, mental illness or mobility limitations), people with limited English language skills, geographically or culturally isolated people, homeless people, senior citizens, pregnant women, and children.

- Disabled Population
  - Disability (18-64 Years) Group
- Economic Disadvantage
  - Uninsured Population
  - Single Parenting
- Language and Literacy
  - Non-English Speaking
- Isolation (cultural, geographic, or social)
  - Rural Population
  - Mobile Homes
  - Farm Workers
- Age
  - Children under 5 Years
  - Age above 65 Years
  - Pregnant Women

1.2 Support Services

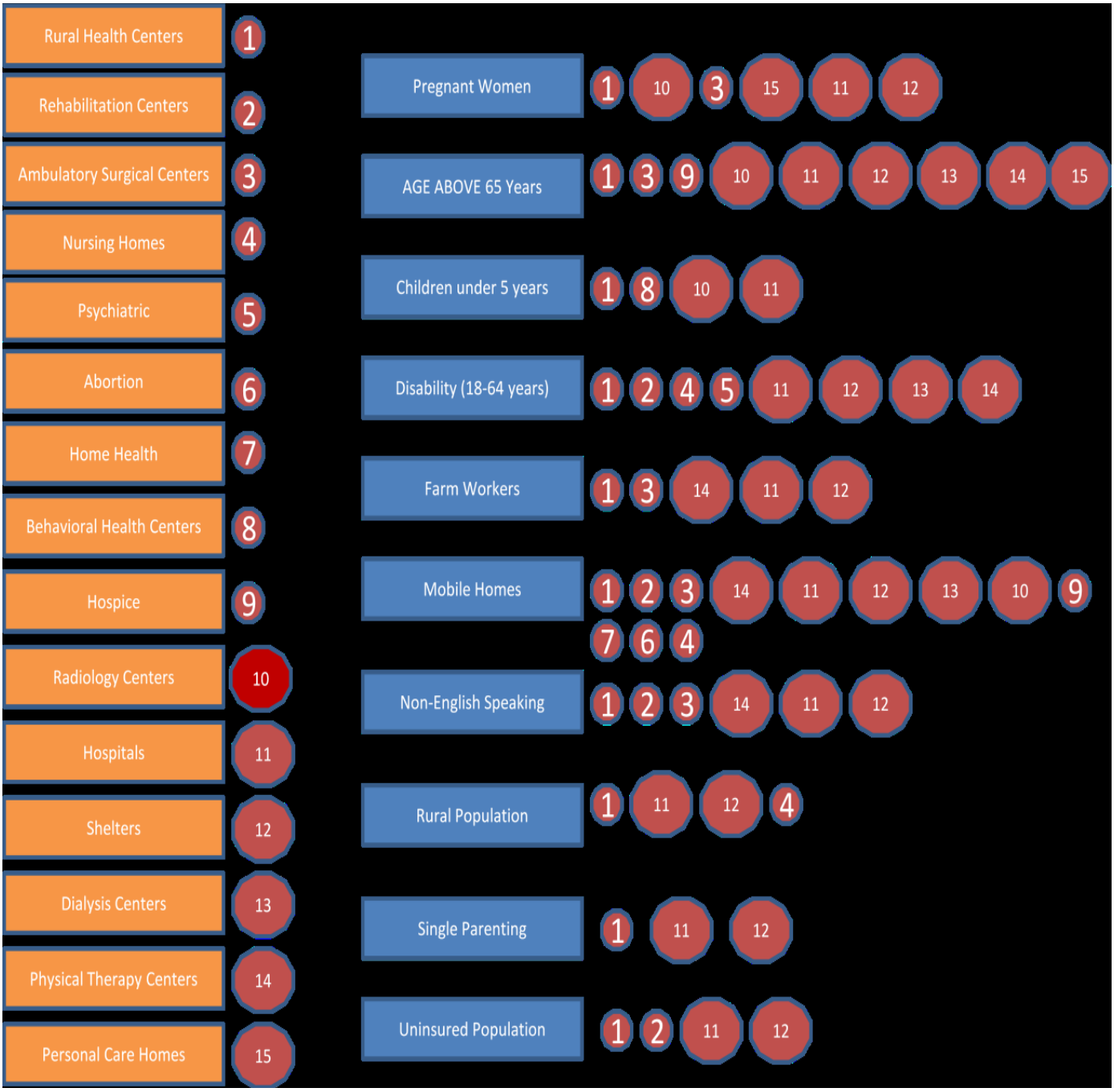
Emergency preparedness for response to natural or manmade hazards relies on identifying different support services (support service type) and mapping these services to different at-risk populations. This provides necessary information to public health personnel to know which at-

risk population groups are targeted to different specialized support services. Different types of support services used were:

- Rural Health Centers
- Rehabilitation Centers
- Ambulatory Surgical Centers
- Nursing Homes
- Psychiatric Centers
- Abortion Centers
- Home Health
- Behavioral Health Centers
- Hospice
- Radiology Centers
- Hospitals
- Shelters
- Dialysis Centers
- Physical Therapy Centers
- Personal Care Homes

1.3 ***Grouping At-Risk Populations to Support Services***

At-risk population is identified based on the guidelines mentioned in Centers for Disease Control (CDC) and Mississippi State Department of Health and the data was obtained from Census 2010. Based on State Department of Health resource guides, support services were identified by the type of service they provide and are numbered. All the support services used or accessed by a particular at-risk population are listed and are grouped together. The table below shows list of at-risk population grouped by the type of support service they uses.


Support Services
At-Risk Population
Grouping
Rural Health Centers 1
Rehabilitation Centers 2
Ambulatory Surgical Centers 3
Nursing Homes 4
Psychiatric 5
Abortion 6
Home Health 7
Behavioral Health Centers 8
Hospice 9
Radiology Centers 10
Hospitals 11
Shelters 12
Dialysis Centers 13
Physical Therapy Centers 14
Personal Care Homes 15
Pregnant Women 1 10 3 15 11 12
AGE ABOVE 65 Years 1 3 9 10 11 12 13 14 15
Children under 5 years 1 8 10 11
Disability (18-64 years) 1 2 4 5 11 12 13 14
Farm Workers 1 3 14 11 12
Mobile Homes 1 2 3 14 11 12 13 10 9 7 6 4
Non-English Speaking 1 2 3 14 11 12
Rural Population 1 11 12 4
Single Parenting 1 11 12
Uninsured Population 1 2 11 12

2. Resource Guide

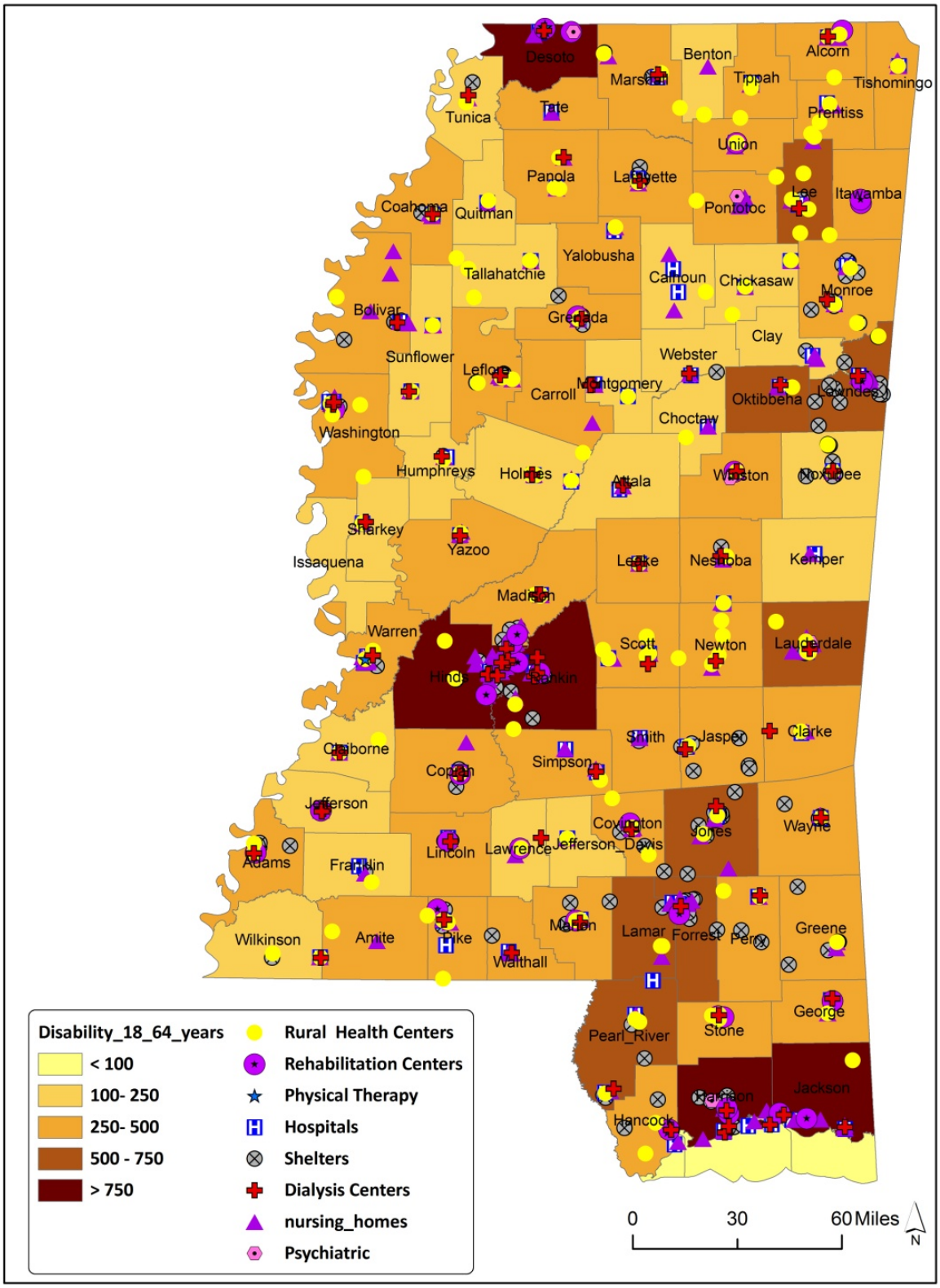
2.1 Disabled Population

People with different conditions that affect sensory, cognitive, or physical, capabilities are often considered disabled. The Americans with Disabilities Act defined disability as a mental or physical impairment that substantially limits one or more major life activities. Many individuals characterized as disabled are highly independent and they may require little or no additional assistance beyond that provided to the general public during emergencies. In contrast, countless people who are not typically considered disabled require additional assistance in various emergency situations. Lack of essential medications and medical care significantly increases the risk and at most importance should be given in recognizing these populations in an emergency. In this category elderly population are also included because of the increased prevalence of chronic diseases and functional limitations. People who are visually impaired, hearing impaired, cognitively impaired, physically limited and/or disabled, Elderly, and those dependent upon medical care, equipment, and/or medications are considered disabled populations.

Visually Impaired Individuals are blind or visually impaired and are vulnerable in emergency events as a result of their limited ability to perceive visual messages and to visually assess unfamiliar environments. Visual impaired individuals will miss visual cues, such as hand signals, colors, and flashing lights and they may also be unaware of vital emergency information that is disseminated only in visual formats. When mapping visually impaired individuals to different services public health personnel should considered this group as they are at increased risk of injuries, particularly if they become separated from their service animals or assistive devices.

Hearing Impaired individuals are deaf or hearing impaired is limited in their ability to hear environmental sounds and, in some cases, to communicate verbally. People with hearing impairments will be particularly vulnerable if they are unable to hear alarms or spoken announcements. They are unable to access and receive instructions and vital emergency information and unable to call for help and communicate with first responders or search-and rescue-personnel. When mapping hearing impaired individuals to different services public health personnel should considered this group as they are at increased risk of injuries as they

Map displaying Disability population densities between 18-64 years with Supporting Services



communicate using sign language or lip reading. Thus in emergency situations that result in limited use or loss of their communication aids hearing impaired individuals are particularly vulnerable.

- Rural Health Centers
- Rehabilitation Centers
- Nursing Homes
- Psychiatric Centers
- Hospitals
- Shelters
- Dialysis Centers

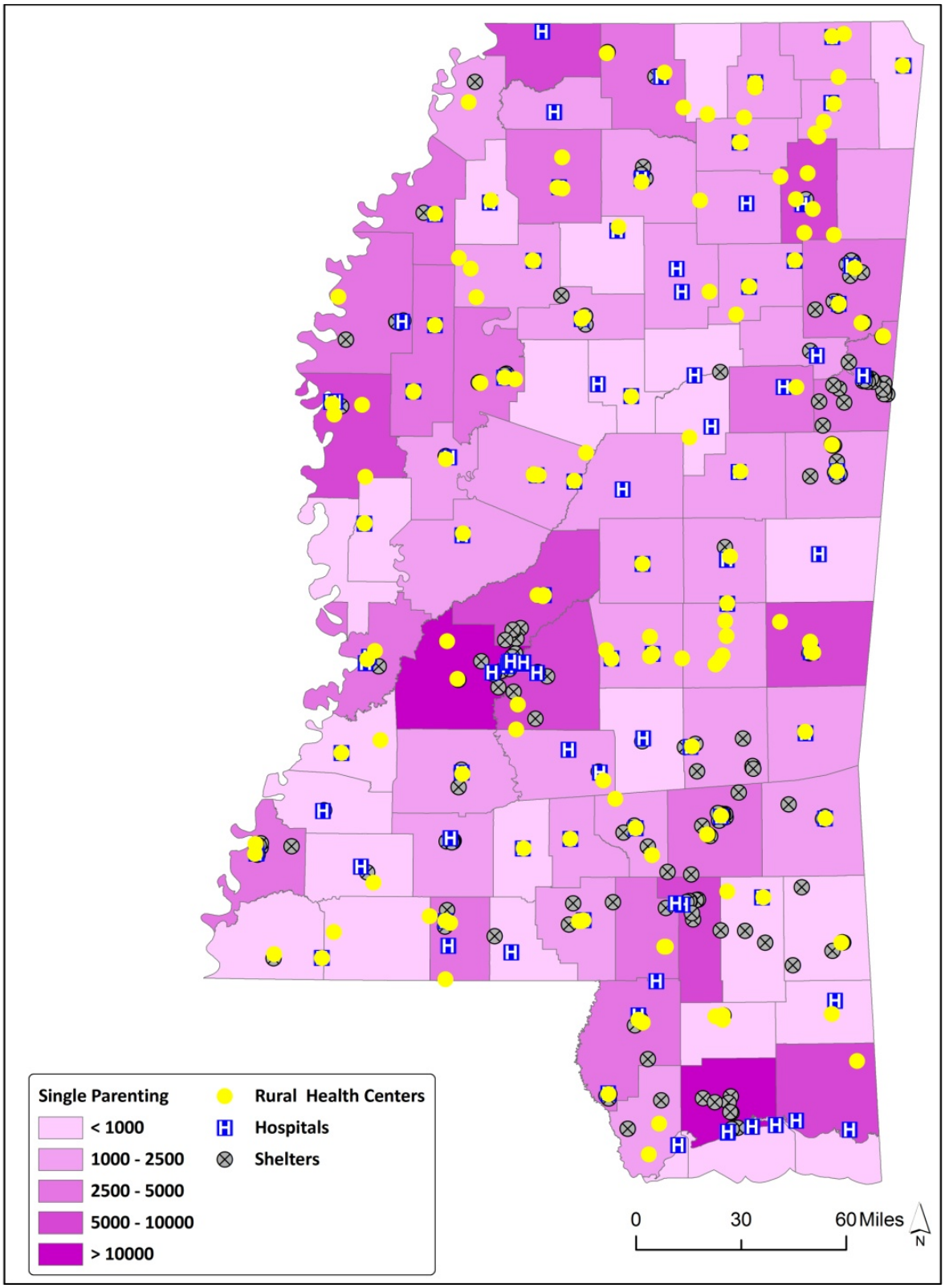
Physical Therapy Centers The support services grouped for the disabled population of age group 18 to 64 years are illustrated above. The support services were geocoded and map layers were overlaid on the disabled population of age group 18 to 64 years.

2.2 Economic Disadvantage

Economic disadvantage address at-risk population using poverty as a criteria. Economic disadvantage is very broad because many people that fall into other categories also live at or below the federal poverty level. When individuals are placed at risk because of both limited language and economic disadvantage the risk is compounded and planning efforts should reflect that risk. Economic disadvantage does not completely impair the ability of an individual but it can significantly affect if the individual does not have the resources.

People who are economically disadvantaged can be reached through traditional communication channels during disasters. The biggest barriers to receiving and acting on health information for this population are often the lack of resources to respond and a lack of awareness of possible threats to their health and well being. Other population groups from the categories of those who are at-risk because of limited language proficiency, disability, isolation (cultural, geographic or sensory), and age should be consider when mapping the economically disadvantaged people living in poverty. Uninsured population and single parent are the two groups that were mapped with different support services.

Single Parenting with Hospitals and Shelters Support Services



Mapping this group will require identifying different support service locations listed below.

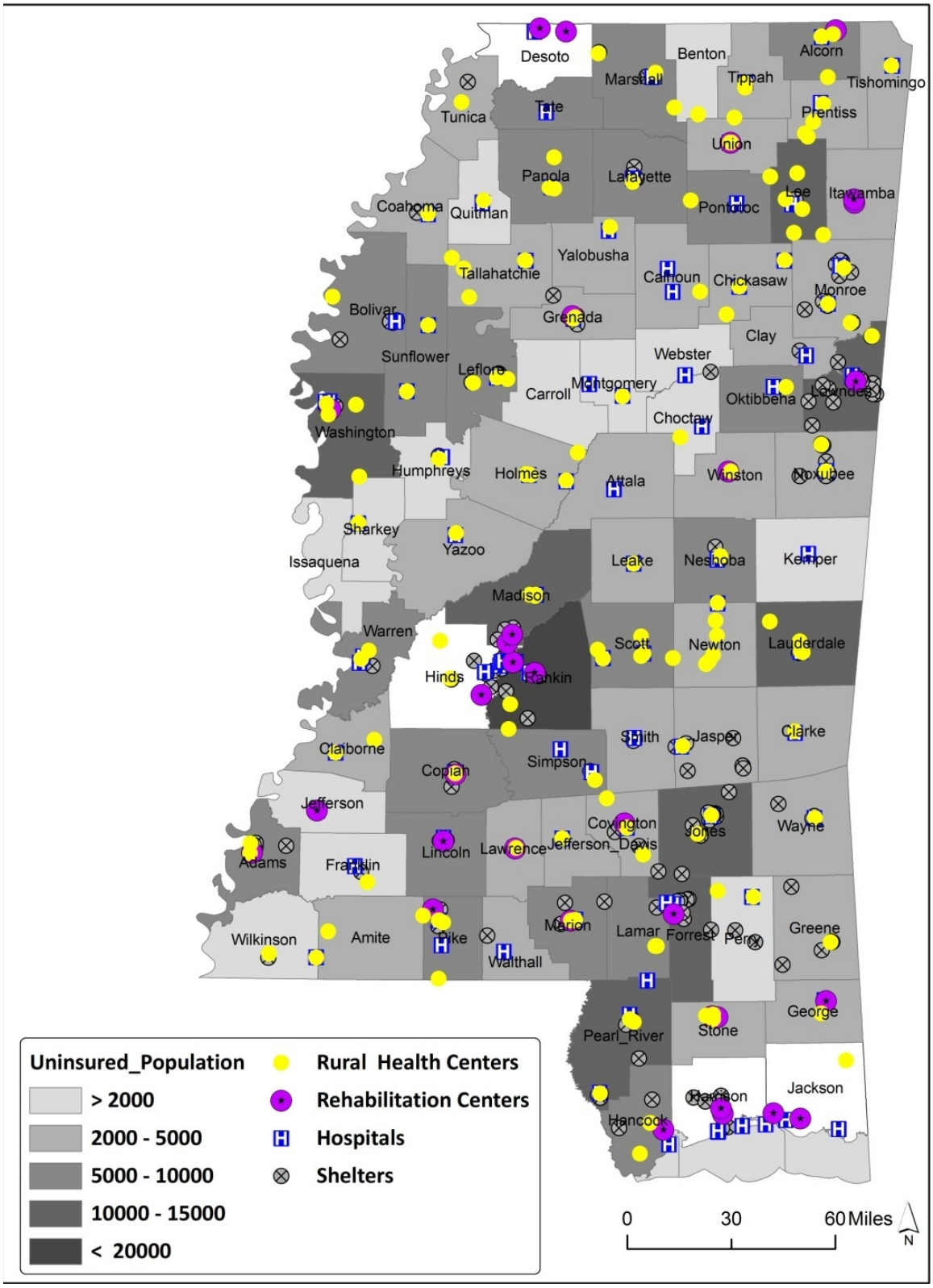
The support services grouped for the economic disadvantage group of uninsured population and single parent listed above.

- Rural Health Centers
- Rehabilitation Centers
- Ambulatory Surgical Centers
- Hospitals
- Shelters

Single parents face challenges because they have no one to share their responsibilities to care for those who are dependent on them. This increased responsibility can impair their ability to plan for emergencies or carry out public health directives, and it can be emotionally overwhelming. Many uninsured populations are members of one or more of the vulnerable population groups.

Single parents and uninsured populations may be faced with daily survival crises and may be unable to take the necessary steps to be prepared during emergencies. The support services were geocoded and map layers were overlaid on the single parent populations and uninsured group separately.

Uninsured with Rehabilitation, Hospitals and Shelters Support Services



2.3 Language and Literacy

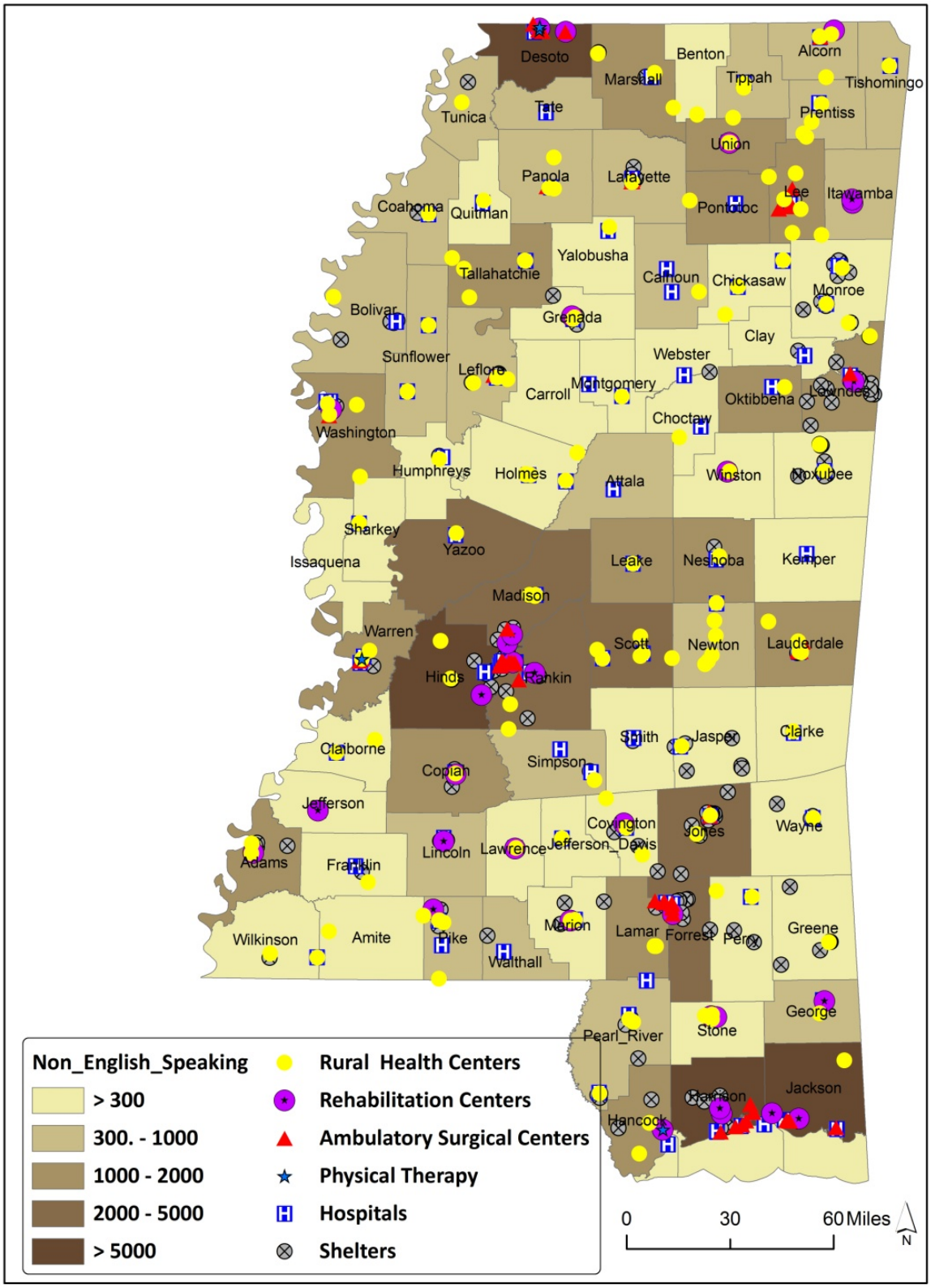
This category includes people who have a limited ability to read, speak, write or understand English, have low literacy skills, or who cannot read at all. It is important to consider language and literacy to ensure that everyone can understand the information and use appropriate resources. People who share a common language and culture often live in the same communities. Within this broad population cultural differences in healthcare and medical practices vary significantly from the mainstream population. Specific cultural and linguistic identifiers are important in defining at-risk populations. Hispanics often define themselves according to national origin. They speak different dialects and have different cultural practices.

The support services grouped for the language and literacy disadvantage group population are listed below.

- Rural Health Centers
- Rehabilitation Centers
- Ambulatory Surgical Centers
- Hospitals
- Shelters
- Physical Therapy Centers

Individuals with limited English proficiency have difficulty understanding both written and verbal information in English. The importance of the ethnic media in reaching people who speak little or no English is still underestimated by most health and emergency planners. Identifying and mapping demographically significant groups of individuals with no or limited English proficiency or those with very low literacy levels will better help public health personnel in allocating resources.

Non English Speaking with Support Services



2.4 Isolation – Rural Population/Mobile Homes/Farm Workers

2.4.1 Rural Population

Isolation is the biggest challenge to reach people with special needs. People can be isolated if they live in rural areas or in the middle of a densely populated urban core. Rural populations who live in sparsely populated communities are also considered at risk population. In urban areas, people can be isolated because of language, lack of education, cultural practices, chronic health problems, fear, lack of transportation or access to public transit systems, unemployment, and other factors. People who live in rural areas are at-risk because they live near farms and raw food supplies, power facilities, and U.S. military facilities.

In rural areas, residents within a certain distance usually know each other. It is vital for emergency professionals to know where these people are located in order to create strategies to reach them in day-to-day communication and especially in an emergency. The support services grouped for the rural population are listed above.

- Rural Health Centers
- Nursing Homes
- Hospitals
- Shelters

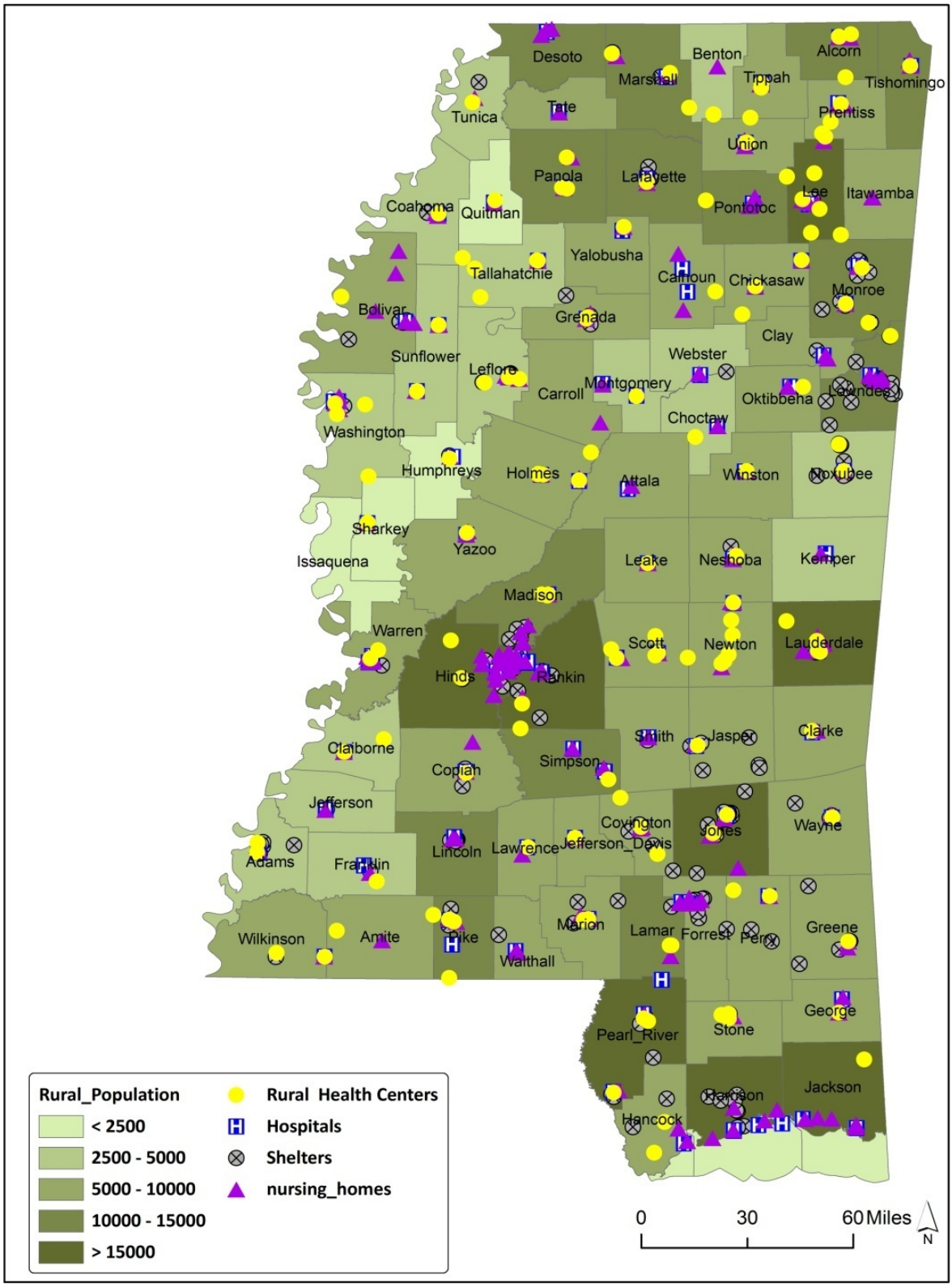
2.4.2 Farm Workers

Increased migration has social, political, and economic consequences for migrating groups, as well as for their sending and host societies. Farm workers are low aid, uninsured employees in an extremely hazardous industry, and they provide an essential service for U.S. society.

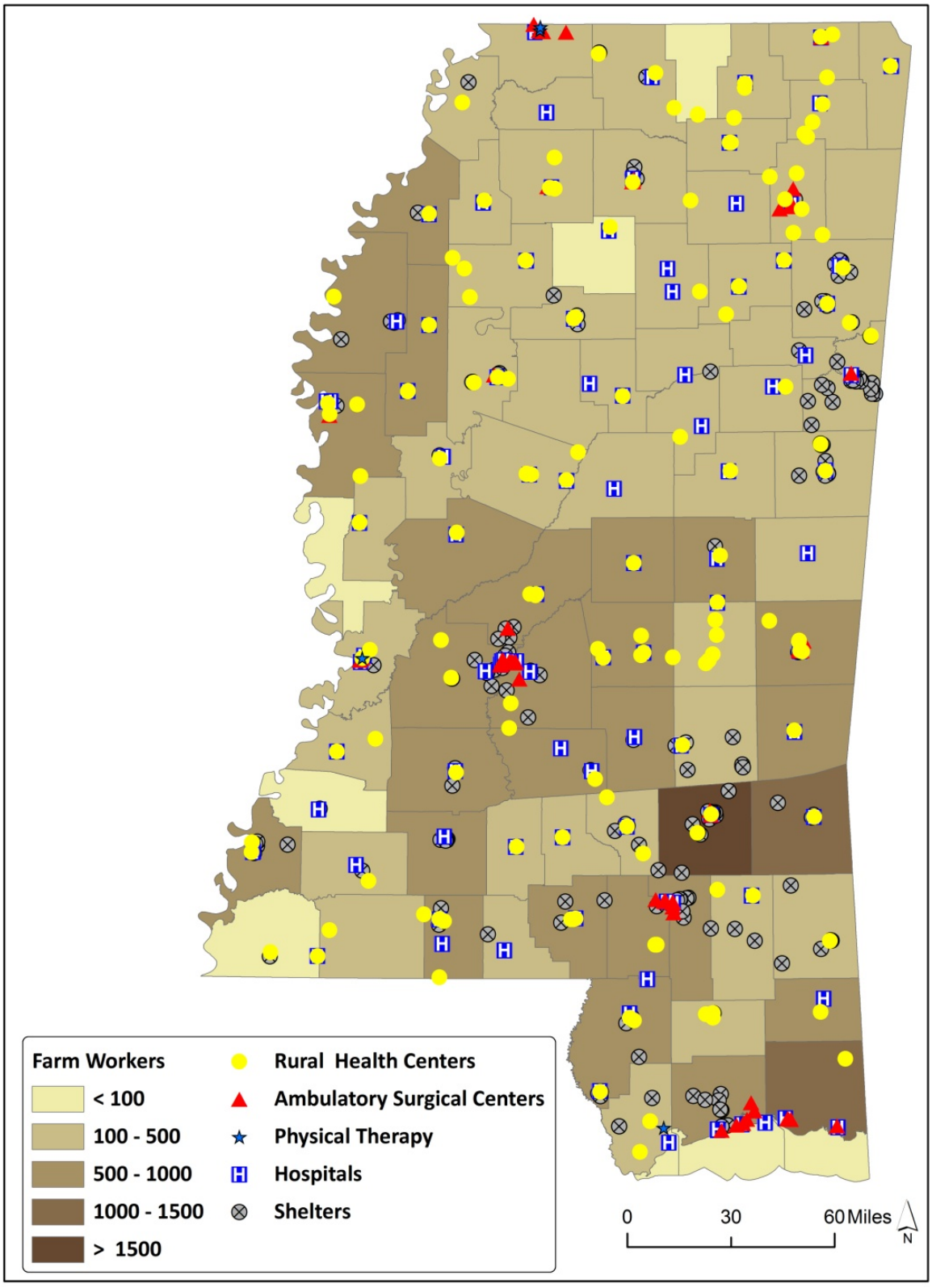
These individuals have a range of responsibilities, from planting, cultivating, grading, and sorting products, to inspecting commodities and facilities. They may work with food crops, animals, or plants. Farm workers are at-risk population and should be considered in emergency preparedness. The support services grouped for the rural population are listed below.

- Rural Health Centers
- Hospitals
- Shelters
- Ambulatory Surgical Centers

Rural Population with Hospitals, Shelters, and Nursing Homes Support Services



Farm Workers with Ambulatory, Physical Therapy, Hospitals and Shelter Support Services



2.4.3 Mobile homes

People in lower income often live in the most vulnerable housing and lack the resources to undertake recommended loss-reduction or evacuation measures. Mobile homes, also most often occupied by lower income residents, are the most dangerous places in a wildfire or windstorm.

Individuals and families often live in mobile homes are less able to withstand disasters. Each kind of hazard or event has distinct physical characteristics, and the spatial data needed to assess vulnerability for mobile homes are determined by the hazard's physical qualities. The support services grouped for the rural population are listed below.

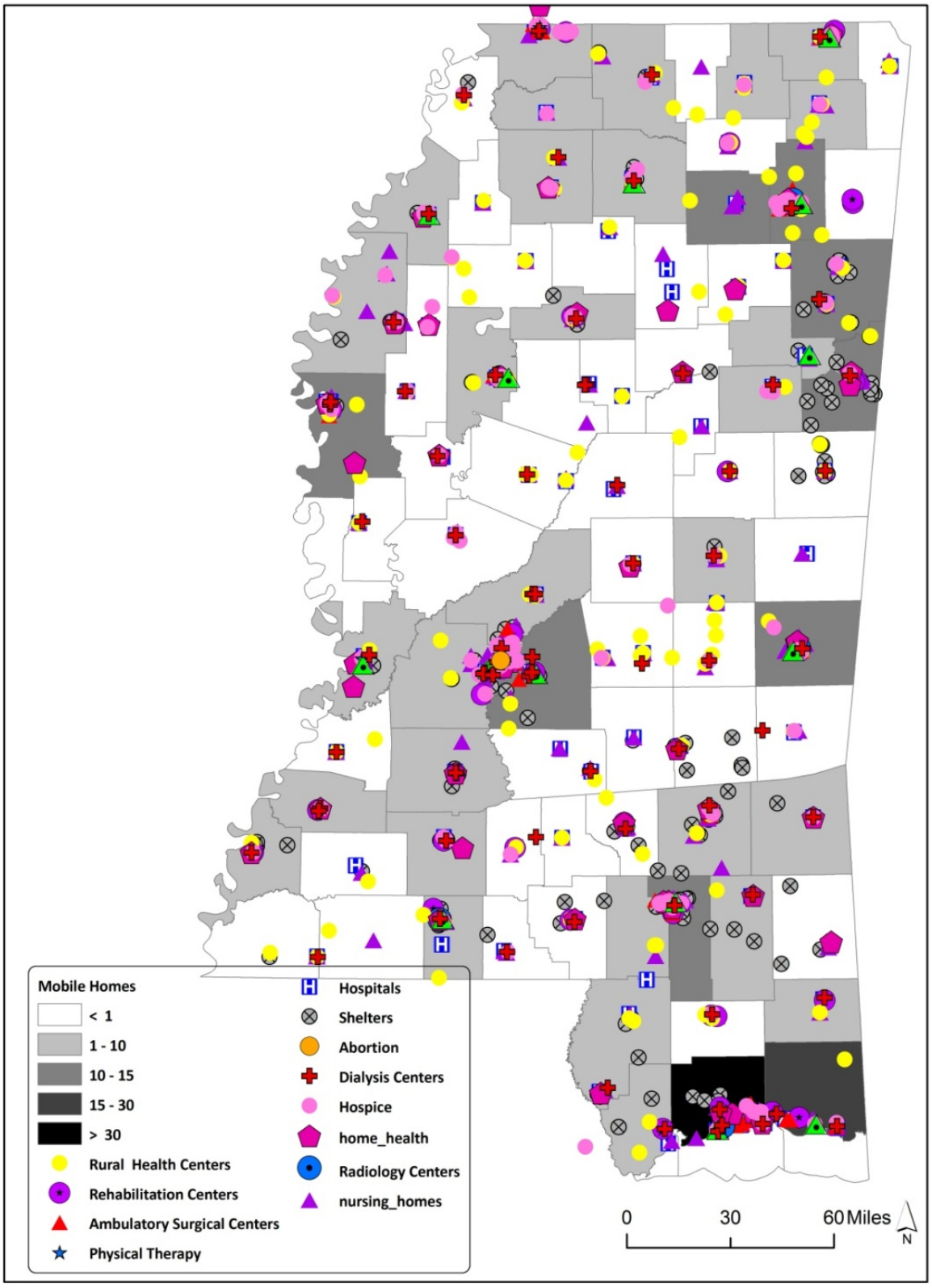
- Rural Health Centers
- Hospitals
- Shelters
- Behavioral Health Centers
- Hospice
- Radiology Centers
- Dialysis Centers
- Physical Therapy Centers

2.5 Age At-Risk Population

2.5.1 Elderly Population age 65 years

Although many elderly people are competent and able to access health care or provide for themselves in an emergency, chronic health problems, limited mobility, blindness, deafness, social isolation, fear, and reduced income put older adults age 65 and above at an increased risk during an emergency. Some frail elderly, however, have hearing, sight, speech, physical, and cognitive impairments that can prevent them from understanding and responding to public health information and emergency directions.

Mobile Homes Residents with Support Services



The support services grouped for the elderly population of age 65 years and above are listed below.

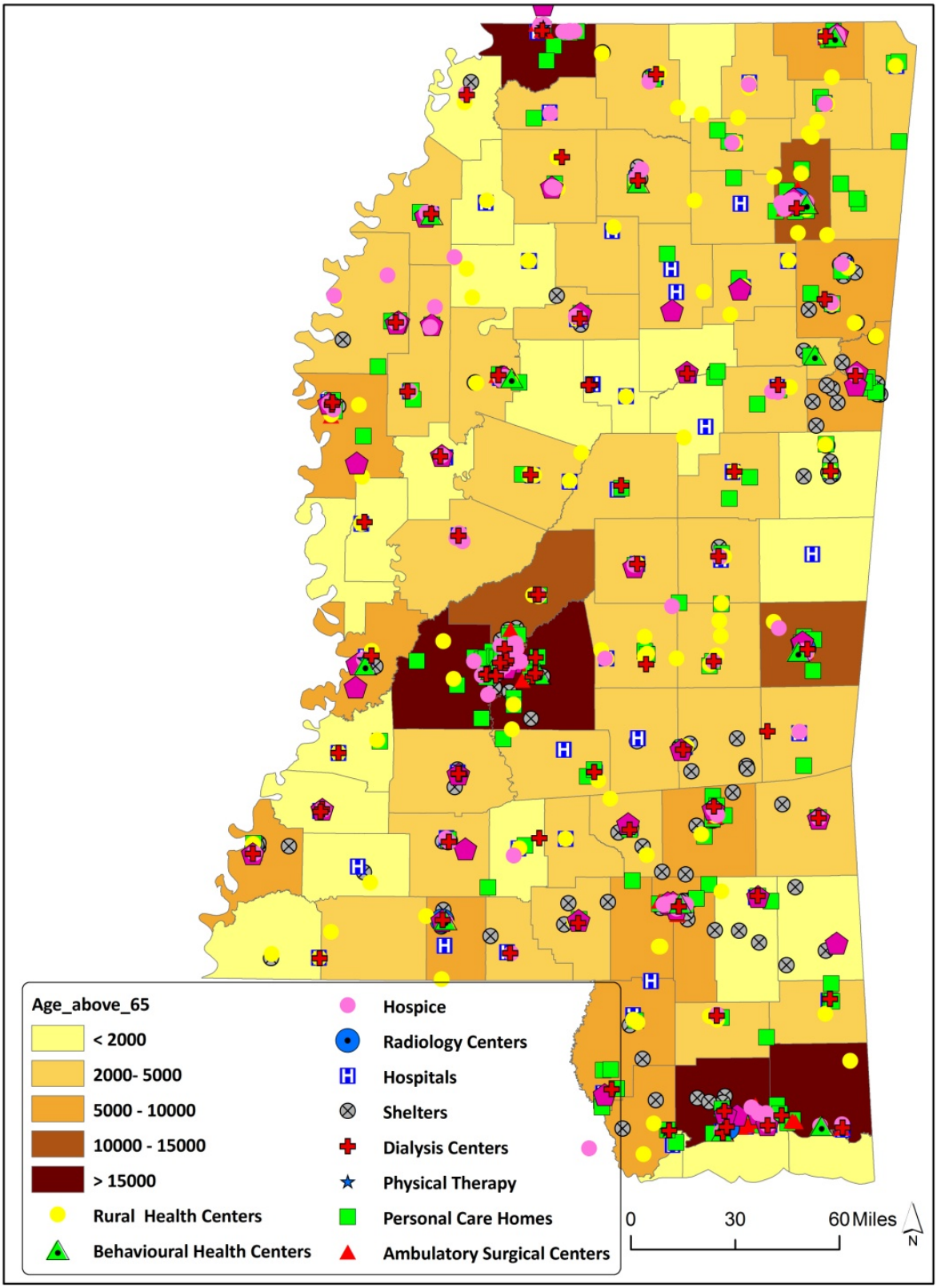
- Rural Health Centers
- Hospitals
- Shelters
- Nursing Homes
- Home Health
- Hospice
- Radiology Centers
- Dialysis Centers
- Physical Therapy Centers
- Personal Care Homes

It is important to focus on people who are 65 years with certain vulnerabilities during emergency preparedness within this broader population. They lack functional capacity both cognitive or the ability for self care and need special attention in emergency preparedness plans. The above map focuses on some of the supporting services to be considered during health hazard assessment for the population above 65 years of age.

2.5.2 Children age 5 years and below

Infants and children under the age of 5 can also be at-risk, particularly if they are separated from their parents or guardians. There are also increasing numbers of children who are home alone after school. Children have unique physical, mental, and social needs, particularly in the case of emergencies or disasters.

Elderly Population Age 65 Years and Above with Support Services

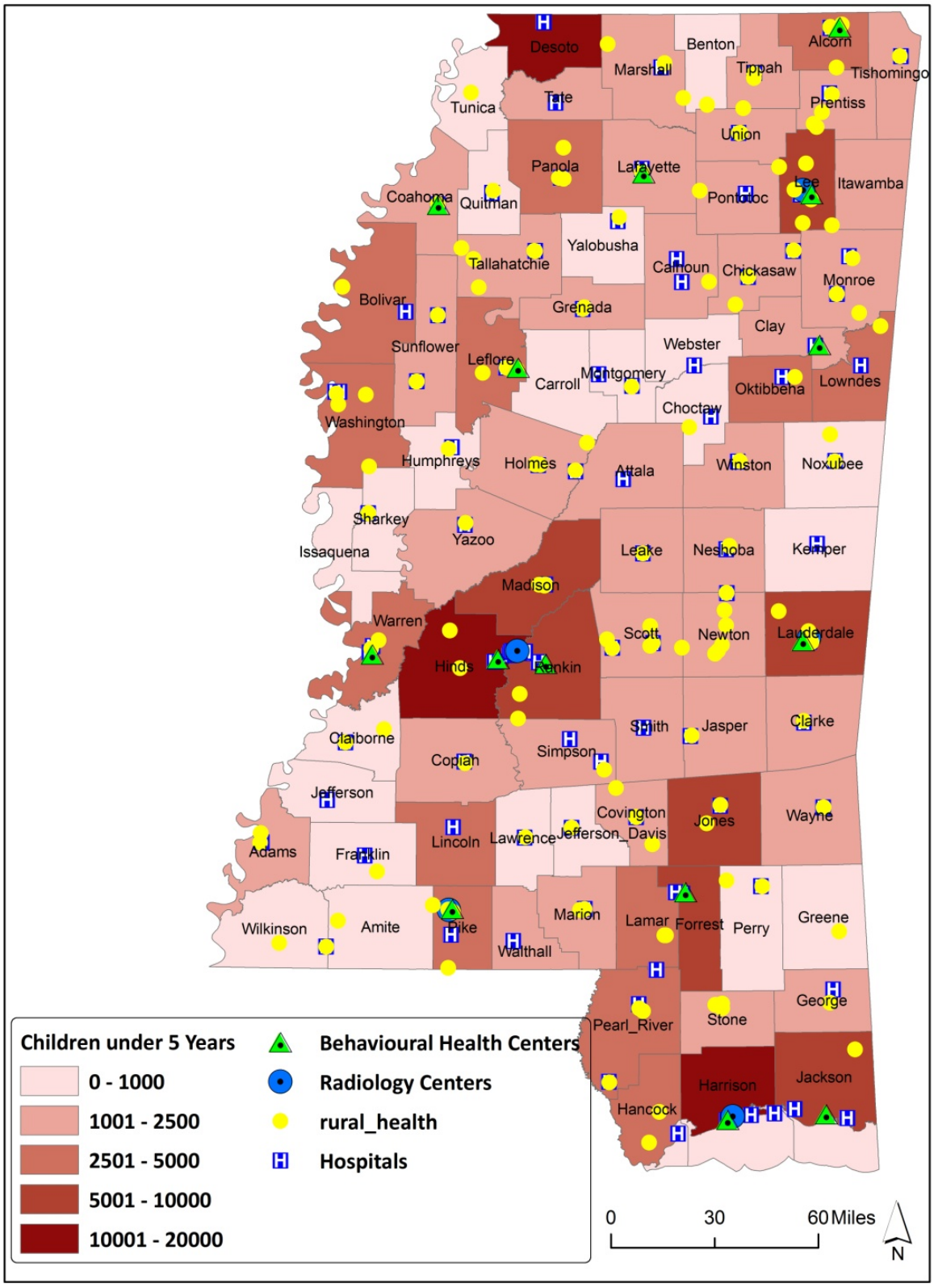


Although the needs may differ by age and developmental state, children's needs do differ from those of adults, and these differences must be considered for emergency preparedness and response. The above map focuses on some of the supporting services to be considered during health hazard assessment for the population above 65 years of age.

The support services grouped for the children population of age 5 years and under are listed below.

- Rural Health Centers
- Hospitals
- Shelters
- Radiology Centers

Children Population Age 5 Years and Under with Support Services



Ambulatory Surgical Centers in Mississippi

Ambulatory Surgical Centers	Address	County	Phone Number
Baptist Desoto Surgery Center	391 Southcrest Circle Suite 1000	Southaven, MS 38671	(662) 349-0910
Better Living Clinic Endoscopy Center	3000 Halls Ferry Road	Vicksburg, MS 39180	(601) 638-9800
Cedar Lake Surgery Center	1720 B Medical Park Drive	Biloxi, MS 39532	(228) 702-2000
Center for Digestive Health	589 Garfield	Tupelo, MS 38801	(662) 377-5800
Coleman Eye and Laser Surgery	2005 Highway 82 West	Greenwood, MS 38930	(662) 455-4523
Columbus Endoscopy Center, Inc.	600 Leigh Drive	Columbus, MS 39705	(662) 327-7525
Columbus Orthopedic Outpatient Center, LLC	640 Leigh Drive	Columbus, MS 39705	(662) 328-7123
Comprehensive Pain Management, PLLC	2089 South Ridge Drive	Tupelo, MS 38801	(662) 407-0801
Delta Gastroenterology,	9140 Hwy 51 North	Southaven, MS 38671	(601) 280-8222
Digestive Disease Center of Hattiesburg	100 Methodist Blvd.	Hattiesburg, MS 39402	(601) 450-5158
East Mississippi Endoscopic Center, LLC	1926 23rd Avenue	Meridian, MS 39302	(601) 485-1131
Endoscopy Center of North Mississippi	1206 Office Park Drive	Oxford, MS 38655	(662) 234-9888
Eye Care Surgery Center of Olive Branch, LLC	6947 Crumpler Blvd.	Olive Branch, MS 38654	(901) 255-5625
Eye Care Surgery Center of Southaven	7600 Airways Blvd., Suite D	Southaven, MS 38671	(662) 349-7477
Eye Laser/Surgery Center of Columbus	634 Leigh Drive	Columbus, MS 39705	(662) 328-1254
Eye Surgical Center of MS, LLC	053 River Oaks Drive	Flowood, MS 39232	(601) 969-1430
Gastrointestinal Associates Endoscopy	106 Highland Way, Suite 101	Madison, MS 39110	(601) 355-5123
Gastrointestinal Associate Endoscopy	1405 North State Street, Third Floor	Jackson, MS 39205	(601) 354-1234
Gastrointestinal Associates Endoscopy	1815 Mission 66	Vicksburg, MS 39180	(601) 355-1234
GI Diagnostic & Therapeutic Center of the Midsouth	7668 Airways Boulevard, Building #2	Southaven, MS 38671	(662) 349-6950
Gulf Coast Outpatient Surgery Center	2781 C T Switzer Sr. Drive; Suite 101	Biloxi, MS 39531	(228) 594-2900
Gulf South Surgery Center	1206 31st Avenue (Box 1778)	Gulfport, MS 39501	(228) 864-0008
Hattiesburg Clinic Ambulatory Surgery Center	415 South 28th Avenue	Hattiesburg, MS 39401	(601) 579-5033

Head and Neck Surgery Center	107 Millsaps Drive	Hattiesburg, MS 39402	(601) 268-5131
Hogan Surgical Center	351 Cowan Road	Gulfport, MS 39507	(228) 896-1120
Institute for Spinal Pain Treatment Center	#1 Lincoln Parkway, Suite 106	Hattiesburg, MS 39402	(601) 264-2115
Jackson Eye Institute & ASC	2500 Lakeland Drive, Suite B	Flowood, MS 39208	(601) 933-1197
Laurel Surgery and Endoscopy Center, LLC	1710 West 12th Street	Laurel, MS 39440	(601) 369-2021
Lowery A. Woodall Outpatient Surgery Facility	105 South 28th Avenue	Hattiesburg, MS 39402	(601) 288-1072
MAE Physicians Surgery Center, LLC	1190 North State Street, Suite 102	Jackson, MS 39202	(601) 968-1790
Magnolia Endoscopy Center	3050 Corder Drive	Corinth, MS 38834	(662) 284-9902
Meridian Surgery Center	2100 13th Street	Meridian, MS 39301	(601) 485-4443
Mid South Pain Treatment Center, LLC	122 Airways Place	Southaven, MS 38671	(662) 349-9990
Mississippi Coast Endoscopy & ASC	2406 Catalpa Avenue	Pascagoula, MS 39567	(228) 696-0818
Mississippi Eye Surgery Center	3432 Bienville Blvd.	Ocean Springs, MS 39563	(228) 875-6658
Mississippi Foot & Surgical Center, LLC	1915 Dunbarton Drive	Jackson, MS 39216	(601) 982-3338
Mississippi Surgical Center	1421 North State Street; Suite 101	Jackson, MS 39202	(601) 353-8000
New Gulf Coast Surgery Center, LLC	3882 Bienville Blvd.	Ocean Springs, MS 39564	(228) 872-6290
New South Neuro Spine, LLC Pain Center	2470 Flowood Drive	Flowood, MS 39232	(601) 664-1213
North Mississippi Ambulatory Surgery Center	589 Garfield Street	Tupelo, MS 38801	(662) 377-4700
North Mississippi Pain Management Center	Longtown Medical Park 4381 South	Tupelo, MS 38801	(662) 844-5477
North Mississippi Spine Center	109 Eureka St., Suite B	Batesville, MS 38606	(662) 563-7728
Ocean Springs Surgical & Endoscopy Center	3301 Bienville Blvd.	Ocean Springs, MS 39564	(228) 872-8854
Oxford Surgery Center	499 Azalea Drive	Oxford, MS 38655	(662) 234-7979
Pain Management Center, LLC	One Layfair Drive, Suite 400	Flowood, MS 39208	(601) 936-8800
Pain Treatment Center of Laurel, LLC	404 South 13th Avenue	Laurel, MS 39440	(601) 425-9042
Pain Treatment Center, LLC	106 Asbury Circle	Hattiesburg, MS 39402	(601) 268-8698
Plastic Surgery Center of Hattiesburg (PSCH), LLC	40 Franklin Road	Hattiesburg, MS 39402	(601) 296-3405

Plastic Surgery Center of Meridian	5002 Highway 39 North, Bldg. D	Meridian, MS 39301	(601) 481-7070
Popp's Ferry Out-Patient Surgery Center, LLC	431 Bertucci Blvd.	Biloxi, MS 39531	(228) 385-2020
Priemier Endoscopy Center of Jackson	501 Marshall Street/Suite 201	Jackson, MS 39202	(601) 352-2273
Rayner Eye Clinic Surgical Center	1308 Belk Drive	Oxford, MS 38655	(662) 234-6551
Regional Surgical Center, Inc.	2525 Highway 1 South, Suite A	Greenville, MS 38701	(601) 335-1103
South Mississippi Maxillofacial Surgery Center	1760 Medical Park Drive	Biloxi, MS 39532	(228) 388-5925
South Mississippi Surgery Center	39 Franklin Road; Suite 100	Hattiesburg, MS 39402	(601) 296-3800
Southern Eye Surgery Center	1420 South 28th Avenue	Hattiesburg, MS 39402	(601) 264-3937
Southern Surgery Center	3688 Veterans Memorial Hwy; Suite 100	Hattiesburg, MS 39401	(601) 554-7525
St. Dominic Ambulatory Surgery Center	970 Lakeland Drive, Suite 15	Jackson, MS 39216	(601) 984-8800
Surgicare of Jackson	760 Lakeland Drive	Jackson, MS 39216	(601) 362-8700
The DeSoto Eye Surgery Center, LLC	726 East Goodman Road, Suite B	Southaven, MS 38671	(662) 349-1959
The Eye Surgery and Laser Center, LLC	501 Marshall Avenue, Suite 604	Jackson, MS 39202	(601) 985-9120
The Plastic Surgical Center of MS	2550 Flowood Drive; Suite 101	Flowood, MS 39232	(601) 939-5544
The Runnels' Plastic Surgery Center, LLC	1057 River Oaks Drive	Flowood, MS 39232	(601) 939-9778
Total Pain Care, LLC	1001 14th Street	Meridian, MS 39301	(601) 482-9224
Tupelo Surgery Center	3353 North Gloster Street	Tupelo, MS 38804	(662) 407-0334

Dialysis Units in Mississippi

Dialysis Unit	Address	County	Phone Number
Bay Springs Dialysis Unit	14 Bay Avenue	Bay Springs, MS 39422	(601) 764-6427
BMA of Southwest	1856 Hospital Drive	Jackson, MS 39204	(601) 371-2896
Central Dialysis of Hazlehurst	232A North Caldwell Drive	Hazlehurst, MS 39083	(601) 894-5509
Central Dialysis of Magee	211 First Street	Magee, MS 39111	(601) 849-3053
Central Dialysis of Yazoo City	716 Grand Avenue	Yazoo City, MS 39194	(662) 746-4172
Central Dialysis Unit	381 Medical Drive	Jackson, MS 39216	(601) 981-9652
Central Dialysis Unit of Forest	1151 Highway 35 South	Forest, MS 39074	(601) 469-3390
Central Dialysis Unit of Kosciusko	107 Ridgewood Circle	Kosciusko, MS 39090	(662) 289-3000
Central Dialysis, Inc. - Canton	Highways 16 & 43	Canton, MS 39046	(601) 981-9652
Collins Dialysis Unit	15 Covington Ridge Place	Collins, MS 39428	(601) 765-2711
Columbia Dialysis Unit	Route 2 Box 52	Columbia, MS 39429	(601) 731-1234
DRG Fayette	225 Community Drive	Fayette, MS 39269	(601) 786-6673
FMC Dialysis Services of Rankin County	141 Gateway Drive	Brandon, MS 39042	(601) 591-0350
Frensenius Medical Care – SMKC-Diamonhead	4495 East Aloha Drive, Suite 1	Diamondhead, MS 39525	(228) 255-6679
Fresenius Medical Care - Port Gibson	123-A McComb Avenue	Port Gibson, MS 39150	(601) 437-3707
Hattiesburg Clinic Dialysis	5909 Hwy 49, Suite 10	Hattiesburg, MS 39402	(601) 296-2960
Laurel Renal Dialysis Center	3105 Hwy 15 North	Laurel, MS 39440	(601) 288-2601
Lucedale Dialysis	652 Manilla Street	Lucedale, MS 39452	(601) 947-8980
Mid-Delta Kidney Center, Inc.	1212 East Railroad Avenue	Greenville, MS 38703	(662) 332-7100
NRI- Brandon	101 Christian Road	Brandon, MS 39042	(601) 824-9764
NRI- Canton	620 East Peace Street	Canton, MS 39046	(601) 354-3062
NRI- Carthage	312 Ellis Street	Carthage, MS 39051	(601) 267-6859
NRI- Hazlehurst	201 North Haley Street	Hazlehurst, MS 39083	(601) 894-5509
NRI Jackson - North	571 Beasley Road	Jackson, MS 39206	(601) 957-1999
NRI Jackson - South	2460 Terry Road, Suite 27-J	Jackson, MS 39204	(601) 373-9154
NRI- Jackson Southwest	1828 Raymond Road	Jackson, MS 39204	(601) 373-7897
NRI- Lexington	22579 Depot Street	Lexington, MS 39095	(662) 834-3587
Ocean Springs Dialysis	12 Marks Road	Ocean Springs, MS 39564	(228) 875-6660
Pachuta Dialysis Unit	180 East Main Street	Pachuta, MS 39347	(601) 77-3012
Pearl River Renal Dialysis Center	318 B Highway 43 South	Picayune, MS 39466	(601) 264-6000
RCG of Mayersville	129 Court Street	Mayersville, MS 39113	(662) 873-2272
Renal Care Group Aberdeen	308 Highway 8 West	Aberdeen, MS 39730	(662) 369-6149
Renal Care Group Belzoni	16451 Highway 49	Belzoni, MS 39038	(662) 247-2251
Renal Care Group Brookhaven	534 Irby Drive	Brookhaven, MS 39601	(601) 833-9720
Renal Care Group Centreville	205 East Main Street	Centreville, MS 39631	(601) 645-9099
Renal Care Group Clarksdale	2010 North Street	Clarksdale, MS 38614	(662) 627-4786
Renal Care Group Cleveland	222 North Pearman Avenue	Cleveland, MS 38732	(662) 843-6965
Renal Care Group Columbus	92 Brookmore Drive	Columbus, MS 39701	(662) 327-9208
Renal Care Group Corinth	810 Alcorn Drive	Corinth, MS 38834	(662) 287-9577
Renal Care Group Eupora	207 Meadowlane Street	Eupora, MS 39744	(662) 258-6509
Renal Care Group Greenville	2001 S. Medical Park Drive	Greenville, MS 38703	(662) 378-2454
Renal Care Group Greenwood	609 Tallahatchie Street	Greenwood, MS 38930	(662) 453-5208

	(PO Box 4740)		
Renal Care Group Grenada	35 West Monroe Street	Grenada, MS 38901	(662) 226-8229
Renal Care Group Holly Springs	1325 Highway 4, East	Holly Springs, MS 38635	(662) 252-6210
Renal Care Group Indianola	627 Highway 82 West	Indianola, MS 38751	(662) 887-5155
Renal Care Group Louisville	462-A East Main Street	Louisville, MS 39339	(662) 773-6565
Renal Care Group Macon	703 North Washington Street	Macon, MS 39341	(662) 726-9866
Renal Care Group McComb	1404 White Street	McComb, MS 39648	(601) 684-6380
Renal Care Group Meridian	2221 Hwy 39 North	Meridian, MS 39301	(601) 485-4817
Renal Care Group Natchez	312 Highland Blvd (PO Box 2024)	Natchez, MS 39120	(601) 446-8060
Renal Care Group Newton	121Old 15 Loop	Newton, MS 39345	(601) 683-9485
Renal Care Group Oxford	1760 Barron Street	Oxford, MS 38655	(662) 234-3412
Renal Care Group Philadelphia	105 Office Drive	Philadelphia, MS 39350	(601) 656-0282
Renal Care Group Sardis	200 East Frontage Road	Sardis, MS 38666	(662) 487-3938
Renal Care Group Southaven	7620 Southcrest Parkway, Suite #5	Southaven, MS 38671	(662) 349-2548
Renal Care Group Starkville	104 West Garrard Road	Starkville, MS 39759	(662) 324-8300
Renal Care Group Tupelo	1031 South Madison Street	Tupelo, Ms 38801	(662) 841-3637
Renal Care Group Vicksburg	105 Keystone Circle	Vicksburg, MS 39180	(601) 634-6057
Renal Care Group-Winona	410 Highway 82	Winona, MS 38967	(662) 283-6353
Richton Dialysis Unit	507 Front Street North	Richton, MS 39476	(601) 288-2601
Silver Creek Dialysis	21 Emu Street	Silver Creek, MS 39663	(601) 296-2961
Singing River Dialysis	4907 Telephone Road	Pascagoula, MS 39567	(228) 762-0701
South MS Kidney Center North Gulfport	2525 33rd Street	Gulfport, MS 39501	(228) 864-0009
South MS Kidney Center of Biloxi	784 Vieux Marche Mall	Biloxi, MS 39530	(228) 436-9204
South MS Kidney Center of D'Iberville	10374 Lamey Bridge Road	D'Iberville, MS 39532	(228) 392-1300
South MS Kidney Center of Gulfport	4300 A West Railroad Street	Gulfport, MS 39501	(228) 864-0009
South MS Kidney Center of Orange Grove	11531 Old Highway 49	Gulfport, MS 39503	(228) 832-9293
Tunica Dialysis	1821 US Hwy 61 North	Tunica, MS 38676	(662) 322-6010
Tylertown Dialysis Unit	4820 Plaza Drive	Tylertown, MS 39667	(601) 222-0311
UMC Pediatric & ESRD Qualified Adult OPT Clinic	2500 North State Street	Jackson, MS 39216	(601) 984-4100
University Hospital & Clinics Outpatient Dialysis	350 W. Woodrow Wilson, Suite 479	Jackson, MS 39213	(601) 815-6345
Waynesboro Dialysis Unit	950 Matthew Drive	Waynesboro, MS 39367	(601) 735-5858
Wiggins Dialysis Unit	503 First Street	Wiggins, MS 39577	(601) 928-2999

Home Health Centers in Mississippi

Home Health Center	Address	County	Phone Number
Amedisys Home Health of Meridian	2900 North Hills Street, Suite A	Meridian, MS 39305	(601) 484-3293
Amedisys Home Health of Biloxi	925 Tommy Munro Drive, Suite K	Biloxi, MS 39532	(228) 388-4144
Amedisys Home Health of Collins	18 Melody Lane	Collins, MS 39428	(601) 765-8316
Amedisys Home Health of Vicksburg	2080 S, Frontage Road, Suite 105	Vicksburg, MS 39180	(601) 619-3670
Baptist Memorial Home Care & Hospice North MS	126 Hwy 51 North (PO Box 1429)	Batesville, MS 38606	(662) 578-8402
Camellia Home Health	2080 S. Frontage Road, Suite 103	Vicksburg, MS 39180	(601) 638-6606
Camellia Home Health and Hospice	133 Mayfair Road (PO Box 1956, Hattiesburg, MS 39403)	Hattiesburg, MS 39402	(601) 268-0408
Camellia Home Health of the Gulf Coast	1001 Howard Avenue	Biloxi, MS 39530	(228) 374-2273
Comfort Care Home Health	2260 Highway 15 N. (PO Box 607, Laurel MS 39441-0607)	Laurel, MS 39440	(601) 425-7521
Continue Care Home Health	401 Bailey Drive (PO Box 186)	Hollandale, MS 38748	(662) 827-2226
Continue Care Home Health II	803 East Sunflower Road (PO Box 432)	Cleveland, MS 38732	(662) 846-7693
Deaconess Home Care-Region I	108 Lundy Lane (PO Box 16929, Hattiesburg, MS 39404-6929)	Hattiesburg, MS 39401	(601) 268-1842
Deaconess Home Care-Region II	105 Whitebrook (PO Box 3550, Brookhaven, MS 39603-7550)	Brookhaven, MS 39601	(601) 835-1145
Deaconess Home Care-Region III	105 Whitebrook (PO Box 3550, Brookhaven, MS 39603-7550	Brookhaven, MS 39601	(601) 835-1145
Delta Community Home Health Agency	300-B South Street	Cleveland, MS 38732	(662) 843-7533
Delta Regional Medical Center Home Health Agency	300 South Washington (PO Box 5247)	Greenville, MS 38704	(662) 725-1200
Forrest General Home Care	1414 South 28th Avenue	Hattiesburg, MS 39402	(601) 288-4344
Gentiva Health Services	101 N. Industrial Road, Suite C (PO Box 4208, Tupelo, MS 38803	Tupelo, MS 38801	(662) 844-9725

Gentiva Health Services	(PO Drawer W/104 Legion Avenue)	Calhoun City, MS 38916	(662) 628-6657
Gentiva Health Services	106 Riverview Drive	Flowood, MS 39232	(601) 362-7801
Gentiva Health Services	189 Park Creek Drive (PO Box 8609)	Columbus, MS 39705	(662) 327-9560
Gentiva Health Services	208 West Green Street (PO Box 186)	Hazlehurst, MS 39083	(601) 894-2701
Gentiva Health Services	2600 Old North Hill Street	Meridian, MS 39305	(601) 484-6726
Grenada Lake Medical Center Home Health Agency	960 Avent Drive	Grenada, MS 38901	(662) 227-7545
Home Choice Health Services, Inc.	2606 Corporate Avenue, Suite 201	Memphis, TN 38132	(901) 380-4404
Intrepid USA HealthCare Services	2175 Business Center Drive #1	Memphis, TN 38134-3108	(901) 213-2520
Kare- In- Home Health Services	10281 Corporate Drive	Gulfport, MS 39503	(228) 604-2155
Magnolia Regional Health Center Home Health & Hospice	2034 East Shiloh Road	Corinth, MS 38834	(662) 293-1405
Marion General Hospital Home Health Agency	906 Sumrall Road (PO Box 630)	Columbia, MS 39429	(601) 740-2201
Medgar Evers Home Health Agency	210 Gilchrist Street (PO Box 699)	Fayette, MS 39069	(601) 786-3955
Methodist Alliance Home Care	6400 Shelby View Drive, Ste. 101	Memphis, TN 38134	(662) 429-2175
Mid-Delta Home Health Agency	405 Hayden Street (PO Box 373)	Belzoni, MS 39038	(662) 247-1254
Mid-Delta Home Health of Charleston	620 South State Street, Suite 3	Clarksdale, MS 38614	(662) 624-4910
Mississippi Home Care of Jackson	817 East River Place, Suite 201	Jackson, MS 39202	(601) 352-5065
Mississippi Home Care of Vicksburg	1911-C Mission 66, Suite C	Vicksburg, MS 39180	(601) 629-0015
Mississippi Homecare of Eupora	8 North Dunn	Eupora, MS 39744	(662) 258-4339
Mississippi Homecare of Picayune	127-A West Canal Street	Picayune, MS 39466	(601) 749-9101
Mississippi Homecare of Richton	205 Elm Avenue E	Richton, MS 39476	(601) 788-2912
North MS Medical Center Home Health Agency	812 Garfield	Tupelo, MS 38801	(662) 372-2499
Pro-Care Home Health	#20 Bay Avenue (PO Box 527)	Bay Springs, MS 39422	(601) 764-2081
Saad's Nursing Service of MS	10598 D'Iberville Blvd, Suite B	D'Iberville, MS 39450	(228) 482-5622
Sat-Home Health Agency, Inc. of Greenwood	205 Walthall (PO Box 1783)	Greenwood, MS 38930	(662) 453-8420
Southeast Home Health Agency Region 8C	503 McInnis Avenue (PO Box 489)	Leakesville, MS 39451	(601) 394-2694

PUBLIC HEALTH DISTRICT VIII			
Southwest Home Health Agency Region 7A PUBLIC HEALTH DISTRICT VII	317 Highland Boulevard, Suite M	Natchez, MS 39120	(601) 445-4350
St. Luke Home Health Services	1504 Aston Avenue (PO Box 1103)	McComb, MS 39648	(601) 249-4260
Sta-Home Health Agency, Inc.	406 Briarwood Drive, Bldg. 200	Jackson, MS 39206	(662) 956-5100
Sta-Home Health Agency, Inc. of Carthage, Inc.	616 Highway 35 South (PO Box 366)	Carthage, MS 39051	(662) 267-9770
Sunflower Home Health	1000 North LF Packer	Ruleville, MS 38771	(662) 756-4676
Tender Loving Care	11010 Highway 49, Suite 4	Gulfport, MS 39503-4191	(228) 831-9821
Tombigbee Home Health Agency Region 4B PUBLIC HEALTH DISTRICT IV	797 South Jackson Street (Rt 1, Box 1049)	Houston, MS 38851	(662) 456-3791
Tombigbee Home Health Agency Region 4C PUBLIC HEALTH DISTRICT IV	400 A Wilkin Wise Road, Suite 1	Columbus, MS 39701	(662) 328-6158
Wayne General Hospital Home Health Agency	920 Matthew Drive (PO Box 1249)	Waynesboro, MS 39367	(601) 735-5500
Wesley Home Health	229 Methodist Blvd.	Hattiesburg, MS 39402	(601) 268-8450

Hospice Centers in Mississippi

Hospice Center	Address	Place	Phone Number
Compassionate Hospice Care of Southern MS	113 Jefferson Davis Blvd	Natchez, MS 39120	601 442-6800
Alliance Hospice, Inc. (was Heavenly Hospice)	127 Pratt Drive	Corinth, MS 38834	662 286-9833
Magnolia Regional Health Center Home Health & Hospice Agency	2034 East Shiloh Road	Corinth, MS 38834	662 293-1405
Angel of Mercy Hospice	314 Court Street PO Box 10	Rosedale, MS 38769	662 759-0344
Continue Care Hospice II	810 East Sunflower Road, Suite 100F (PO Box 432)	Cleveland, MS 38732	662 846-6211
Delta Soul Medical, LLC	102 North Pearman Avenue, Suite 6	Cleveland, MS 38732	662 843-0006
Grace Community Hospice (was New-Era Hospice)	316 North Davis, Suite C	Cleveland, MS 38732	662 846-7600
Haven Hospice and Palliative Care, LLC	700 East Sunflower Rd, Suite 9 (PO Box 537)	Cleveland, MS 38732	662 846-0922
Mercy Hospice	901 Forest Street (PO Box 282)	Shelby, MS 38774	662 902-5183
Word of Deliverance Hospice	216 N. Chrisman (PO Box 239)	Cleveland, MS 38732	662 843-8797
Hospice Care Group	114 East Donald Street	Quitman, MS 39355	601 776-8880
Celestial Comfort Care Hospice	110 Yazoo Avenue; Suite 212 (PO Box 1052)	Clarksdale, MS 38614	662 592-4213
Healing Hands Hospice	1742 North State Street (PO Box 2174)	Clarksdale, MS 38614	662 621-9850
Lion Hospice & Palliative Care, LLC	2001 Hospital Drive	Clarksdale, MS 38614	662 621-1171
Memorial Hospice, Inc.	600 Ohio Street (PO Box 1726)	Clarksdale, MS 38614	662 624-2872
Revelation Hospice & Palliative Care, LLC	809 Delta Avenue, # 104 (PO Box 1626)	Clarksdale, MS 39614	662 621-1370
Zion Hospice, Inc.	112 East Second Street (PO Box 1985)	Clarksdale, MS 38614	662 624-6089
A&E Hospice	6810 Crumpler Blvd. Suite 101	Olive Branch, MS 38654	662 890-4646
Comprehensive Hospice Services	7155 Kerr Street. Suite 13	Olive Branch, MS 38654	662 890-6939
Hospice Advantage of Southaven	919 Ferncliff Cove, Suite 1	Southaven, MS 38671	662 393-4033
New-Era Hospice	8869 Centre Street, Suite 3 & 4 (PO Box 1229)	Southaven, MS 38671	662342-7023

North Delta Hospice & Palliative Services	123 Stateline Road East, Suite C (PO Box 1798)	Southaven, MS 38671	662 393-0170
Specialty Hospice, LLC	5600 Goodman Road, Suite D	Olive Branch, MS 38654	662 420-7157
Spring Valley Hospice	7139 Commerce Drive, Bldg B2	Olive Branch, MS 38654	662 890-5554
Unity Hospice Care - Southaven	7203 Goodman Road	Olive Branch, MS 38654	662 893-5662
Camellia Hospice	133 Mayfair Road	Hattiesburg, MS 39402	601 264-8691
Community Hospice, Inc.	6242 Highway 98; Suite 60	Hattiesburg MS 39401	601 336-5832
Deaconess Hospice - Hattiesburg	108 Lundy Lane (PO Box 15788 Zip 39404-5788)	Hattiesburg, MS 39401	601 261-4010
Forrest General Home Care Hospice	1414 South 28th Avenue	Hattiesburg, MS 39402	601 288-2500
Hospice Advantage of Hattiesburg (was Providence Hospice South)	811 Rebecca Avenue	Hattiesburg, MS 39401	601 705-0360
Kare-In-Home Hospice	6222 Hwy 98; Suite 200	Hattiesburg, MS 39402	601 336-7855
Southern Care Hattiesburg	4700 Hardy Street, Suite Y	Hattiesburg, MS 39402	601 579-9493
St. Luke Missionary Hospice	1715 Hardy Street, Suites 30-40	Hattiesburg, MS 39401	601 796-7993
Legacy Hospice of the South – Grenada (was Unity Hospice – Grenada)	1300 Sunset Drive, Suite Z	Grenada, MS 38901	662 226-4246
Milestone Hospice	347 Tatum Street	Grenada, MS 38901	662 226-8878
Camellia Hospice of the Gulf Coast	999 Howard Avenue, Suite 3	Biloxi, MS 39530-3679	228 374-4434
Canon Hospice (Inpatient & Outpatient – Gulfport)	1520 Broad Avenue, Suite 500	Gulfport, MS 39501	228 575-6251
Deaconess Hospice - Biloxi	951 Howard Avenue	Biloxi, MS 39531	228 435-2265
Hospice Advantage of Biloxi (was Gulf Coast Hospice)	4107 Popps Ferry Road	D'Iberville, MS 39540	228 354-9636
Odyssey Hospice -Biloxi	962 Tommy Munro Drive Cedar Lake Medical Park	Biloxi, MS 38532	228 385-7845
Saad's Hospice Services of MS	10598 D'Iberville Blvd, Suite B	D'Iberville, MS 39450	228 432-8855
Southern Care Biloxi	8195A Woolmarket Road	Biloxi, MS 39532	228 396-4756
Alpha Healthcare	407 Briarwood Drive, Suite 207C (PO Box 59446, Zip 39284)	Jackson, MS 39206	601 977-1198
Compassionate Hospice Care	5935 Highway 18 West, Suite A1 (PO Box 59503 Zip 39284)	Jackson, MS 39209	601 923-8070
Holistic Care Hospice Jackson	1757 Terry Road Bldg 3	Jackson, MS 39204	601 346-7737

Hospice Advantage of Jackson	13 Northtown Drive, Suite 130	Jackson, MS 39211	601 956-9755
Hospice Care At Home	6531 Dogwood View Pkwy (PO Box 9303 Zip 39286)	Jackson, MS 39213	601 713-0061
Miracle Care Hospice Care at Home	330 Edgewood Terrace Drive, Suite B	Jackson, MS 39206	601 982-1909
North Lion Hospice & Palliative Care, LLC	135 Bounds Street, Suite C	Jackson, MS 39206	601 321-8812
Our Family Home Hospice, Inc.	2570 Bailey Avenue, Suite 10 (PO Box 11233 Zip 39283)	Jackson, MS 39213	601 362-1712
Physician Hospice Care	953 North Street, Suite M	Jackson, MS 39202	601 949-8900
South West Hospice	5638 Terry Road	Byram, MS 39272	662 919-6131
Southern Care Jackson	322 Hwy 80 West, Suite 3	Clinton, MS 39056	601 924-8285
Sta-Home Hospice of MS, Inc.	406 Briarwood Drive, Suite 200	Jackson, MS 39206	601 991-1933
Mid Delta Hospice	405 Hayden Street (PO Box 373)	Belzoni, MS 39038	662 247-1254
Hospice of Light	2012 Highway 90 Suite 29	Gautier, MS 39553	228 497-2400
Pinnacle Hospice	10532 Automall Parkway, Suite D	D'Iberville, MS 39540	228 207-0390
We Care Hospice	3725 Main Street	Moss Point, MS 39563	228 474-2030
At Home Hospice Care	5 Harrison Street (PO Box 9)	Fayette, MS 39069	601 786-9494
Medgar Evers Hospice	210 Gilchrist Street (PO Box 699)	Fayette, MS 39069	601 786-3955
CCI Professional Healthcare Hospice	316 Central Avenue	Laurel, MS 39440	601 425-3047
Comfortcare Hospice	2260 Highway 15 N (PO Box 607)	Laurel, MS 39441	601 422-0054
Camellia Hospice of North MS	2166 South Lamar	Oxford, MS 38655	662 238-7771
Gentiva Hospice - Oxford	104 Skyline Drive	Oxford, MS 38655	662 234-0140
Harper's Hospice Care, Inc.	1703 24th Avenue	Meridian, MS 39301	601 483-4134
Hometown Hospice, Inc.	8366 Hwy 19 North	Collinsville, MS 39325	601 626-7277
Hospice Advantage of Meridian	1300-C 14th Street	Meridian, MS 39301	601 483-9990
Divinity Hospice	207 East Broadway Street (PO Box 523)	Monticello, MS 39654	601 341-0290
Homecare Hospice, Inc.	16482 Highway 21 (PO Box 365 , Sebastopol, MS 39359)	Walnut Grove, MS 39189	601 625-7840
Infinity Hospice, LLC	206 North Van Buren Street	Carthage, MS 39051	601 298-0060
Camellia Hospice of Northeast MS (was Unity Hospice Care – Tupelo)	1413 West Main Street, Suite B	Tupelo, MS 38801	662 844-2870
Gentiva Hospice - Tupelo	144 S. Thomas Street,	Tupelo, MS 38801	662 620-1050

	Suite 105		
Hospice Advantage of Tupelo/Corinth (was AseraCare Hospice –Tupelo)	280 S. Thomas, Suite 101 (PO Box 3478 Zip 38803)	Tupelo, MS 38801	662 840-3434
North MS Medical Center Hospice	422-A East President Street	Tupelo, MS 38801	662 377-3612
Sanctuary Hospice House, The	5159 West Main Street (PO Box 2177 Zip 38803)	Tupelo, MS 38802	662 844-2111
Southern Care Tupelo	1682 Wilson Street	Tupelo, MS 38804	662 841-0085
Angelic Hospice & Palliative Care Services	307 Lamar Street	Greenwood, MS 38930	662 453-5348
Divine Hospice & Palliative Care	415 Carrollton Avenue	Greenwood, MS 38930	662 454-6668
Serenity Hospice	703 Sycamore, Suite A	Greenwood, MS 38930	662 455-2606
Deaconess Hospice - Brookhaven	130 S. First Street	Brookhaven, MS 39601	601 823-5990
Hospice Advantage of Brookhaven (Hospice Advantage of Magee)	102 N. First Street	Brookhaven, MS 39601	601 849-5903
Southern Care Brookhaven	706 Highway 51 N	Brookhaven, MS 39601	601 823-4812
Hospice Ministries	450 Towne Center Blvd.	Ridgeland, MS 39157	601 898-1053
Mid-Delta Hospice of Canton	1150 East Peace Street	Canton, MS 39046	601 855-2400
Pax Hospice (Life Source Services, LLC)	359 Towne Center Blvd, Suite 500	Ridgeland, MS 39157	601 991-3840
Baptist Hospice-Golden Triangle	2623 5th Street North	Columbus, MS 39705	662 243-1173
Physician Hospice Care	617A Highway 7 South (PO Box 5083 Zip 38635)	Holly Springs, MS 38634	662 252-5052
Hospice Advantage of Aberdeen	117 South Meridian Street	Aberdeen, MS 39730	662 369-5777
Legacy Hospice of the South	231 North Main Street	Amory, MS 38821	662 257-9811
Covenant Pallative & Hospice	250 West Beacon Street	Philadelphia, MS 39350	601 656-7411
Quality Hospice Care, Inc.	340 Byrd Avenue (PO Box 517)	Philadelphia, MS 39350	601 656-5252
Southern Care Newton	191 Northside Drive	Newton, MS 39345	601 683-7500
Gentiva Hospice – Starkville (was Heart To Heart Hospice)	115-A Highway 12 West	Starkville, MS 39759	662 615-1519
Legacy of the South – Starkville (was Unity Hospice – Starkville)	1085 Stark Road, Suite F	Starkville, MS 39759	662 338-0078
Baptist Memorial Home Care and Hospice – North MS	126 Highway 51 N (PO Box 1429)	Batesville, MS 38606	662 578-8402
Community Hospice, Inc.	564 Highway 6 East	Batesville, MS 38606	662 561-0902
Legacy Hospice of the South (was Complete Hospice & Palliative	108 Woodland Road, Suite 70	Batesville, MS 38606	662 578-8177

Care)			
Mid-Delta Hospice of Batesville	112 Highway 51 North	Batesville, MS 38606	662 563-1021
Southern Care Batesville	465 Hwy 6 East	Batesville, MS 38606	662 578-4072
Camellia Hospice Southwest MS	620 Delaware Avenue	McComb, MS 39648	601 684-5033
Hospice Compassus - MS	140 North 5th Street, Suite B	McComb, MS 39648	601 250-0884
Gentiva Hospice (was Gilbert's Hospice Care – Tupelo, MS)	301 W. College Street	Booneville, MS 38829	662 728-7404
Camellia Hospice of Central MS	225 Katherine Drive	Flowood, MS 39232	601 932-9066
Gentiva Hospice (was Gilbert's Hospice Care of MS)	106 Riverview Drive	Flowood, MS 39232	601 983-6193
Gentiva Hospice (was Odyssey Hospice – Jackson)	106 Riverview Drive	Flowood, MS 39232	601 983-6193
Guardian Angel Hospice, Inc.	3269 Highway 80	Morton, MS 39117	601 732-8473
St. Joseph Hospice of Southern Mississippi	115 West College Avenue	Wiggins, MS 39577	601 928-2925
Carol's Hospice & Palliative Services of Shelby MS	163 North Main St (PO Box 23)	Drew, MS 38737	662 745-6100
Genesis Hospice Care - Indianola	201 Highway 82 West	Indianola, MS 38751	662 887-1274
Grenada North Delta Hospice & Palliative Services	141 North Main Street	Drew, MS 38737	662 745-0587
Saint Anthony's Hospice & Palliative Care	108 North Ruby Avenue	Ruleville, MS 38771	662 466-0330
Saint John Hospice & Palliative Care	106 North Ruby Avenue	Ruleville, MS 38771	662 756-0928
Sandanna Hospice	105 East Floyce Street (PO Box 365)	Ruleville, MS 38771	662 756-9999
Destiny Hospice, Palliative Care & Specialty Services, Inc.	202 Second Street (PO Box 190)	Tutwiler, MS 38963-0190	662 345-0077
AseraCare Hospice -Senatobia	144 Norfleet Drive	Senatobia, MS 38668	662 562-7607
Legacy Hospice of the South (Was Mercy Hospice – Ripley)	111-B Hospital Street	Ripley, MS 38663	662 837-9990
Hope Hospice	2073 Old Highway 61 N, Suite 1 (PO Box 458)	Tunica, MS 38676	622 357-0461
North Delta Hospice New Albany	212 Starlyn Avenue	New Albany, MS 39652	662 393-0170
L & L Hospice	423-A Beulah Avenue	Tylertown, MS 39667	601 876-6169
Patient's Choice Hospice & Palliative Care	1911-A Mission 66	Vicksburg, MS 39180	601 638-8308
Serenity Premier Hospice	1905B Mission 66, Suite 1	Vicksburg, MS	601 661-9752

	(PO Box 820472 Zip 39182)	39180	
Delta Area Hospice Care, LTD	522 Arnold Avenue (PO Box 5915 Zip 38704-5915)	Greenville, MS 38701	662 335-7040
Delta Regional Medical Center Hospice Agency	300 S. Washington (PO Box 5247 Zip 38704-5247)	Greenville, MS 38701	662 378-1200
North Haven Hospice & Palliative Care, LLC	1696 South Colorado Street, Suite 4	Greenville, MS 38701	662 335-1788
Wayne General Hospital Hospice Agency	920 Matthew Drive (PO Box 1249)	Waynesboro, MS 39367	601 735-7133
Continue Care Hospice	15359 Highway 49 South; Suite 3	Yazoo City, MS 39194	662 746-5815
Cornerstone Palliative & Hospice	125 South Main Street	Yazoo City, MS 39194	601 746-5153
Magnolia Hospice of Mississippi	405 East Fifteenth Street	Yazoo City, MS 39194	662 751-1888

Hospitals in Mississippi

Hospital Name	Address	Place	Phone Number
Natchez Community Hospital	129 Jefferson Davis Boulevard (PO Box 1203)	Natchez, MS 39120	601 445-6205
Natchez Regional Medical Center	54 Seargent Prentiss Drive (PO Box 1488)	Natchez, MS 39121	601 443-2100
Magnolia Regional Health Center	611 Alcorn Drive	Corinth, MS 38834	662 293-1000
Montford Jones Memorial Hospital	220 Highway 12 West (PO Box 887)	Kosciusko, MS 39090	662 289-4311
Bolivar Medical Center	901 E. Sunflower Road (PO Box 1380)	Cleveland, MS 38732	662 846-0061
Calhoun Health Services	140 Burke/Calhoun City Road	Calhoun City, MS 38916	662 628-6611
Trace Regional Hospital	1004 East Madison Street (PO Box 626)	Houston, MS 38851	662 456-3700
Pioneer Community Hospital of Choctaw	311 West Cherry Street (PO Box 417)	Ackerman, MS 39735	662 285-6235
Patients' Choice Medical Center of Claiborne County	123 McComb Avenue (PO Box 1004)	Port Gibson, MS 39150	601 437-5141
H. C. Watkins Memorial Hospital	605 South Archusa Avenue	Quitman, MS 39355	601 776-6925
Clay County Medical Corporation (*was NMMC)*	835 Medical Center Drive	West Point, MS 39773	662 495-2300
Northwest Mississippi Regional Medical Center	1970 Hospital Drive (PO Box 1218)	Clarksdale, MS 38614	662 627-3211
Hardy Wilson Memorial Hospital	233 Magnolia Street (PO Box 889)	Hazlehurst, MS 39083	601 894-4541
Covington County Hospital	701 South Holly Street (PO Box 1149)	Collins, MS 39428	601 765-6711
Baptist Memorial Hospital DeSoto	7601 Southcrest Parkway	Southaven, MS 38671	662 772-4000
Parkwood Behavioral Health	8135 Goodman Road	Olive Branch, MS 38654	662 895-4900
Forrest General Hospital	6051 U. S. Highway 49 (PO Box 16389)	Hattiesburg, MS 39404	601 288-7000
Regency Hospital of Southern Mississippi	6051 U.S. Hwy 49, 5th Floor	Hattiesburg, MS 39401	601 288-8510
Franklin County Memorial Hospital	40 Union Church Road (PO Box 636)	Meadville, MS 39653	601 384-5801
George County Hospital	859 Winter Street (PO Box 607)	Lucedale, MS 39452	601 947-3161
Greene County Hospital	1017 Jackson Street (PO Box 819)	Leakesville, MS 39451	601 394-4135
Grenada Lake Medical Center	960 Avent Drive	Grenada, MS 38901	662 227-7000
Hancock Medical Center	149 Drinkwater Boulevard (PO Box 2790)	Bay St. Louis, MS 39521-2790	228 467-8744
Biloxi Regional Medical Center	150 Reynoir Street (PO Box 128)	Biloxi, MS 39530	228 432-1571
Garden Park Medical Center	15200 Community Road	Gulfport, MS 39503	228 575-7000

Memorial Hospital of Gulfport	4500 13th Street (PO Box 1810)	Gulfport, MS 39501	228 867-4000
Select Specialty Hospital Gulf Coast	1520 Broad Avenue	Gulfport, MS 39501	228 575-7500
Central MS Medical Center	1850 Chadwick Drive	Jackson, MS 39204	601 376-1000
Mississippi Baptist Medical Center	1225 North State Street (PO Box 23668)	Jackson, MS 39202	601 968-1000
Mississippi Hospital for Restorative Care	1225 North State Street (PO Box 23695)	Jackson, MS 39202	601 973-1661
Mississippi Methodist Rehabilitation Center	1350 East Woodrow Wilson Drive	Jackson, MS 39216	601 364-3360
Regency Hospital of Jackson	969 Lakeland Drive 6th floor	Jackson, MS 39216	601 364-6200
Saint Dominic-Jackson Memorial Hospital	969 Lakeland Drive	Jackson, MS 39216	601 200-2000
Select Specialty Hospital Jackson	5903 Ridgewood Road	Jackson, MS 39211	601 899-3011
Univerity Hospitals & Health Systems University of Mississippi Medical Center	2500 North State Street	Jackson, MS 39216	601 984-4100
Holmes County Hospital & Clinics	239 Bowling Green Road	Lexington, MS 39095	662 834-1321
Patients' Choice Medical Center of Humphreys County	500 CCC Road (PO Box 510)	Belzoni, MS 39038	662 247-3831
Ocean Springs Hospital	3109 Bienville Boulevard	Ocean Springs, MS 39564	228 818-1111
Singing River Hospital	2809 Denny Avenue	Pascagoula, MS 39581	228 809-5000
Jasper General Hospital	15-A South 6th Street (PO Box 527)	Bay Springs, MS 39422	601 764-2101
Jefferson County Hospital	870 South Main Street (PO Box 577)	Fayette, MS 39069	601 786-3401
Jefferson Davis Community Hospital	1102 Rose Street (PO Box 1288)	Prentiss, MS 39474	601 792-4276
South Central Regional Medical Center	1220 Jefferson Street (PO Box 607)	Laurel, MS 39441	601 426-4500
John C. Stennis Memorial Hospital	14365 Highway 61 West	DeKalb, MS 39328	769 486-1000
Baptist Memorial Hospital North MS	2301 South Lamar Boulevard (PO Box 946)	Oxford, MS 38655	662 232-8100
South Mississippi State Hospital	823 Highway 589	Purvis, MS 39475	601 794-0100
Wesley Medical Center	5001 Hardy Street (PO Box 16509)	Hattiesburg, MS 39402	601 268-8000
Alliance Health Center	5000 Highway 39 North	Meridian, MS 39301	601 483-6211
Anderson Regional Medical Center	2124 14th Street	Meridian, MS 39301	601 553-6000
Anderson Regional Medical Center South	1102 Constitution Avenue (PO Box 1810)	Meridian, MS 39301	601 693-2511
East Mississippi State Hospital	4555 Highland Park Drive (PO Box 4128)	Meridian, MS 39304	601 581-7600
Regency Hospital of Meridian	1102 Constitution Ave., 2nd Floor	Meridian, MS 39301	601 484-7900
Rush Foundation Hospital	1314 19th Avenue	Meridian, MS 39301	601 703-1443
Specialty Hospital of Meridian, The	1314 19th Avenue	Meridian, MS 39301	601 486-4211

Lawrence County Hospital	1065 East Broad Street (PO Box 788)	Monticello, Ms 39654	601 587-4057
Baptist Medical Center Leake	310 Ellis Street (PO Box 909)	Carthage, MS 39051	601 267-1100
North Mississippi Medical Center	830 South Gloster	Tupelo, MS 38801	662 377-3000
North Mississippi State Hospital	1937 Briar Ridge Road	Tupelo, MS 38804	662 690-4200
Greenwood Leflore Hospital	1401 River Road (PO Box 1410)	Greenwood, MS 38935-1410	662 459-7000
LTAC Hospital of Greenwood	1401 River Road – 2nd Floor	Greenwood, MS 38930	662 451-5172
King's Daughters Medical Center	427 Highway 51 North (PO Box 948)	Brookhaven, MS 39601	601 835-9488
Madison River Oaks Medical Center	161 River Oaks Drive (PO Box 1607)	Canton, MS 39046	601 855-4000
Marion General Hospital	1560 Sumrall Road (PO Box 630)	Columbia, MS 39429	601 736-6303
Alliance HealthCare System	1430 Highway 4 East (PO Box 6000)	Holly Springs, MS 38635	662 252-1212
Gilmore Memorial Regional Medical Center	1105 Earl Frye Boulevard (PO Box 459)	Amory, MS 38821	662 256-7111
Pioneer Community Hospital of Aberdeen - Critical Access Hospital	400 South Chestnut Street (PO Box 548)	Aberdeen, MS 39730	662 369-2455
Kilmichael Hospital	301 Lamar Street (PO Box 188)	Kilmichael, MS 39747	662 262-4311
Tyler Holmes Memorial Hospital	409 Tyler Holmes Drive	Winona, MS 38967	662 283-4114
Neshoba County General Hospital	1120 East Main Street (PO Box 648)	Philadelphia, MS 39350	601 663-1200
Laird Hospital	25117 Highway 15	Union, MS 39365	601 774-8214
Pioneer Community Hospital of Newton	9421 Eastside Drive Extension (PO Box 299)	Newton, MS 39345	601 683-2031
Noxubee General Critical Access Hospital	606 North Jefferson Street (PO Box 480)	Macon, MS 39341	662 726-4231
OCH Regional County Hospital	400 Hospital Road (PO Drawer 1506)	Starkville, MS 39759	662 323-4320
Tri-Lakes Medical Center	303 Medical Center	Batesville, MS 38606	662 563-5611
Highland Community Hospital	130 Highlands Parkway (PO Box 909)	Picayune, MS 39466	601 798-4711
Pearl River County Hospital	305 West Moody Street (PO Box 392)	Poplarville, MS 39470	601 795-4543
Perry County General Hospial	206 Bay Avenue (PO Box 1665)	Richton, MS 39476	601 788-6316
Beacham Memorial Hospital	205 North Cherry Street (PO Box 351)	Magnolia, MS 39652	601 783-2353
Southwest MS Regional Medical Center	215 Marion Avenue (PO Box 1307)	McComb, MS 39649	601 249-5500
Pontotoc Critical Access Hospital	176 South Main Street (PO Box 790)	Pontotoc, MS 38863	662 489-7640
Baptist Memorial Hospital - Booneville	100 Hospital Street	Booneville, MS 38829	662 720-5004
Quitman County Hospital	340 Getwell Drive	Marks, MS 38646	662 326-8031

Brentwood Behavioral Healthcare of MS	3531 Lakeland Drive	Jackson, MS 39232	601 936-2024
Crossgates River Oaks Hospital	350 Crossgates Boulevard	Brandon, MS 39042	601 824-8501
Mississippi State Hospital	3550 Highway 468 West (PO Box 157-A)	Whitfield, MS 39193	601 351-8000
Oak Circle Center	3550 Highway 468 West Bldg 23, MS State Hospital	Whitfield, MS 39193	601 351-8000
River Oaks Hospital	1030 River Oaks Drive	Flowood, MS 39296	601 936-2390
Whitfield Medical Surgical Hospital	Building 60, Oak Circle	Whitfield, MS 39193	601 351-8023
Woman's Hospital at River Oaks	1026 North Flowood Drive	Flowood, MS 39232	601 932-1000
S. E. Lackey Critical Access Hospital & Swingbed	330 North Broad Street (PO Box 428)	Forest, MS 39074	601 469-4151
Scott Regional Hospital	317 Highway 13 South (PO Box 259)	Morton, MS 39117	601 732-6301
Sharkey-Issaquena Community Hospital	47 South Fourth Street (PO Box 339)	Rolling Fork, MS 39159	662 873-5150
Magee General Hospital	300 S. E. Third Avenue	Magee, MS 39111	601 849-5070
Simpson General Hospital	1842 Simpson, Highway 149 (PO Box 457)	Mendenhall, MS 39114	601 847-2221
Patients Choice Medical Center	327 Magnolia Drive	Raleigh, MS 39153	601 782-9997
Stone County Hospital	1434 East Central Avenue (PO Drawer 97)	Wiggins, MS 39577	601 928-6600
Medical/Dental Facility at Parchman	Highway 49 West (PO Box E)	Parchman, MS 38738	662 745-6611
North Sunflower Medical Center	840 North Oak Avenue (PO Box 369)	Ruleville, MS 38771	662 756-2711
South Sunflower County Hospital	121 East Baker Street	Indianola, MS 38751	662 877-5235
Tallahatchie General Hospital	201 South Market Street (PO Box 230)	Charleston, MS 38921	662 647-5535
North Oak Regional Medical Center	401 Getwell Drive (PO Box 648)	Senatobia, MS 38668	662 562-3100
Tippah County Hospital	1005 City Avenue North (PO Box 499)	Ripley, MS 38663	662 837-9221
Tishomingo Health Services, Inc.	1777 Curtis Drive (PO Box 860)	Iuka, MS 38852	662 423-6051
Baptist Memorial Hospital Union County	200 Highway 30 West	New Albany, MS 38652	662 538-7631
Walthall General Hospital	100 Hospital Drive	Tylertown, MS 39667	601 876-0400
Promise Hospital of Vicksburg	1111 Frontage Road 2nd Floor	Vicksburg, MS 39180	601 883-3439
River Region Health System	2100 Highway 61 North (PO Box 590)	Vicksburg, MS 39183	601 883-5000
Allegiance Specialty Hospital of Greenville	300 South Washington Avenue 3rd Floor	Greenville, MS 38701	662 332-7925
Delta Regional Med Ctr - West Campus	300 South Washington Avenue	Greenville, MS 38701	662 334-2169
Delta Regional Medical Center	1400 East Union Street (PO Box 5247)	Greenville, MS 38703	662 334-2169
Wayne General Hospital	950 Matthew Drive (PO Box 1249)	Waynesboro, MS 39367	601 735-5151

Webster Health Services, Inc.	70 Medical Plaza	Eupora, MS 39744	662 258-6221
Field Memorial Community Hospital	270 West Main Street (PO Box 639)	Centreville, MS 39631	601 645-5221
Diamond Grove Center	2311 Highway 15 South	Louisville, MS 39339	662 779-0119
Winston Medical Center	562 East Main (PO Box 967)	Louisville, MS 39339	662 773-6211
Yalobusha General Hospital	630 S. Main Street (PO Box 728)	Water Valley, MS 38965	662 473-1411
King's Daughters Hospital of Yazoo County	823 Grand Avenue	Yazoo City, MS 39194	662 746-2261

ICF/MR Facilities and Community Homes in Mississippi

Facility Name	Address	Place	Phone Number
Alexander Milne Home for Women	616 East 19th Street	Laurel, MS 39440	601 399-0700
Boswell Regional Center/W. L. Jaquith ICF/MR	(PO Box 128)	Magee, MS 39111	601 867-5000
Canton Manor	1145 Tisdale Avenue	Canton, MS 39046	601 859-6712
Clover Circle – Ellisville State School	1101 Hwy 11	Ellisville, MS 39437	601 477-9384
Delta Manor	701 U.S. Hwy, 322 West	Clarksdale, MS 38614	662 627-2212
Hillside ICF/MR -- Ellisville State School	1101 Highway 11 South	Ellisville, MS 39437	601 477-9384
Hudspeth Regional Center	(PO Box 127-B)	Whitfield, MS 39193	601 664-6000
Lincoln Residential Center	524 Brookman Drive	Brookhaven, MS 39601	601 835-1884
Millcreek ICF/MR	900 First Avenue, N.E.	Magee, MS 39111	601 849-4221
Mississippi Adolescent Center	760 Brookman Drive Extension	Brookhaven, MS 39601	601 823-5700
North MS Regional Center	967 Regional Center Drive (PO Box 967)	Oxford, MS 38655	662 234-1476
Paul D Cotton ICF/MR-- Ellisville State School	1101 Hwy 11 South	Ellisville, MS 39437	601 477-9384
Pecan Grove – Ellisville State School	1101 Highway 11 South	Ellisville, MS 39437	601 477-9384
Rolling Hills Development Center	200 Womack Road	Starkville, MS 39759	662 323-9183
Son Valley	461 Goodloe Road (PO Box 406)	Canton, MS 39046	601 859-2100
South MS Regional Center	1170 West Railroad Street	Long Beach, MS 39560	228 868-2923

Nursing Homes in Mississippi

Name	Address	Place	Phone Number
Adams County Nursing Center	587 John R. Junkin Drive	Natchez, MS 39120	601 446-8426
Crown Health & Rehab of Natchez	344 Arlington Avenue	Natchez, MS 39120	601 443-2344
Glenburney Nursing Home	555 John R. Junkin Drive	Natchez, MS 39120	601 442-4395
Cornerstone Health & Rehabilitation of Corinth	302 Alcorn Drive	Corinth, MS 38834	662 286-2286
MS Care Center of Alcorn County	3701 Joanne Drive	Corinth, MS 38834	662 287-8071
Whitfield Nursing Home	2101 East Proper Street (PO Box 1425)	Corinth, MS 38834	662 286-3331
Liberty Community Living Center	323 Industrial Park Drive (PO Box 676)	Liberty, MS 39645	601 657-1000
Attala County Nursing Center	326 Highway 12 West	Kosciusko, MS 39090	662 289-1200
MS State Veterans Home-Kosciusko	310 Autumn Ridge Drive	Kosciusko, MS 39090	662 289-7809
Ashland Health & Rehabilitation	16056 Boundary Drive (PO Box 490)	Ashland, MS 38603	662 224-6196
Bolivar Medical Center - LTC Facility	901 E. Sunflower Road (PO Box 1380)	Cleveland, MS 38732	662 846-2508
Cleveland Nursing & Rehabilitation Center	4036 Hwy 8 East (PO Box 1688)	Cleveland, MS 38732	662 843-4014
Joy Health & Rehabilitation Center of Cleveland, LLC	200 Dr. Martin Luther King Drive	Cleveland, MS 38732	662 843-5347
Oak Grove Retirement Home, Inc.	209 Oak Circle (PO Box 198)	Duncan, MS 38740	662 395-2577
Shelby Nursing & Rehab Center	1108 Church Street	Shelby, MS 38774	662 398-5117
Bruce Community Living Center	176 Highway 9 South (PO Box 1280)	Bruce, MS 38915	662 412-5100
Calhoun County Nursing Home	152 Burke CC Road (PO Box 110)	Calhoun City, MS 38916	662 628-6651
Vaiden Community Living Center	868 Mulberry Street	Vaiden, MS 39176	662 464-7714
Floy Dyer Manor	Highway 8 East 1000 East Madison	Houston, MS 38851	662 456-1100
Shearer Richardson Memorial Nursing Home	512 Rockwell Drive (PO Box 420)	Okolona, MS 38860	662 447-5463
Choctaw County Nursing Center	311 West Cherry Street (PO Box 1039)	Ackerman, MS 39735	662285-3257
Claiborne County Senior Care	2124 Old Highway 61 South (PO Box 1018)	Port Gibson, MS 39150	601 437-8737

Lakeside Living Center	191 Highway 511 East (PO Drawer 10)	Quitman, MS 39355	601 776-2141
Dugan Memorial Home	804 East Main Street (PO Box 698)	West Point, MS 39773	662 494-3640
West Point Community Living Center	1122 N. Eshman Avenue (PO Box 817)	West Point, MS 39773	662 494-6011
Clarksdale Nursing Center	1120 Ritchie Street (PO Box 1304)	Clarksdale, MS 38614	662 627-2591
Greenbough Nursing Center	340 Desoto Avenue, Extended	Clarksdale, MS 38614	662 627-3486
Copiah Living Center	806 West Georgetown Road	Crystal Springs, MS 39039	601 892-1880
Pine Crest Guest Home, Inc.	133 Pine Street	Hazlehurst, MS 39083	601 894-1411
Arrington Living Center	701 South Holly Street	Collins, MS 39428	601 765-6711
Covington County Nursing Center	1207 South Fir Street (PO Box 1089)	Collins, MS 39428	601 765-8262
MS State Veterans Home-Collins	3261 Highway 49 South	Collins, MS 39428	601 765-0403
DeSoto Healthcare Center	7805 Southcrest Parkway	Southaven, MS 38671	662 349-7500
Golden Living Center- Southaven	1730 Dorchester Drive	Southaven, MS 38671	662 393-0050
Landmark of DeSoto	3068 Nail Road West	Horn Lake, MS 38637	662 280-1219
Bedford Alzheimer's Care Center	300 Cahal Street	Hattiesburg, MS 39401	601 544-5300
Bedford Care Center of Hattiesburg	10 Medical Boulevard	Hattiesburg, MS 39401	601 264-3709
Bedford Care Center of Petal	908 S. George Street	Petal, MS 39465	601 544-7441
Bedford Care Center-Monroe Hall	300 Cahal Street	Hattiesburg, MS 39401	601 544-5300
Hattiesburg Health & Rehab Center	514 Bay Street	Hattiesburg, MS 39401	601 544-4230
Meadville Convalescent Home, Inc.	300 Highway 556	Meadville, MS 39653	601 384-5861
George Regional Health & Rehab Center	859 Winter Street (PO Box 607)	Lucedale, MS 39452	601 947-9101
Glen Oaks Nursing Center	55 Suzanne Street	Lucedale, MS 39452	601 947-2783
Greene Rural Health Center	1017 Jackson Street (PO Box 819)	Leakesville, MS 39451	601 394-2371
Leakesville Rehabilitation & Nursing Center, Inc.	1300 Melody Lane (PO Box 640)	Leakesville, MS 39451	601 394-2331
Grace Health & Rehab of Grenada	1966 Hill Drive	Grenada, MS 38901	662 226-2442
Grenada Lake Transitional Care Center	960 Avent Drive	Grenada, MS 38901	662 227-7000
Grenada Living Center	1950 Grandview Drive	Grenada, MS	662 226-9554

		38901	
Dunbar Village	725 Dunbar Avenue	Bay St. Louis, MS 39520	228 466-3099
Woodland Village Nursing Center	5427 Gex Road	Diamondhead, MS 39525	228 255-4832
Biloxi Community Living Center	2279 Atkinson Road	Biloxi, MS 39531	228 388-1805
Boyington Health Care Facility	1530 Broad Avenue	Gulfport, MS 39501	228 864-6544
Dixie White House Nursing Home	538 Menge Avenue	Pass Christian, MS 39571	228 452-4344
Driftwood Nursing Center	1500 Broad Avenue	Gulfport, MS 39501	228 868-1314
Greenbriar Nursing Center	4347 West Gay Road	D'Iberville, MS 39540	228 392-8484
Lakeview Nursing Center	16411 Robinson Road	Gulfport, MS 39503	228 831-3001
Belhaven Senior Care	1004 North Street	Jackson, MS 39202	601 355-0763
Chadwick Nursing & Rehabilitation Center	1900 Chadwick Drive	Jackson, MS 39204	601 372-0231
Clinton Healthcare	1251 Pinehaven Road	Clinton, MS 39056	601 924-2996
Community Place	1129 Langley Avenue	Jackson, MS 39204	601 355-0617
Compere's Nursing Home	865 North Street	Jackson, MS 39202	601 948-6531
Cottage Grove Nursing Home	1116 Forest Avenue	Jackson, MS 39206	601 366-6461
Forest Hill Nursing Center	927 Cooper Road	Jackson, MS 39212	601 372-0141
Hinds County Nursing & Rehabilitation Center	3454 Albemarle Road	Jackson, MS 39213	601 362-5394
Lakeland Nursing & Rehabilitation Center	3680 Lakeland Lane	Jackson, MS 39216	601 982-5505
Magnolia Senior Care	3701 Peter Quinn Drive	Jackson, MS 39213	601 366-1712
Manhattan Nursing & Rehabilitation Center	4540 Manhattan Road	Jackson, MS 39206	601 982-7421
MS State Veteran's Home	4607 Lindbergh Drive	Jackson, MS 39209	601 353-6143
Pleasant Hills Community Living Center	1600 Raymond Road	Jackson, MS 39204	601 371-1700
Trinity Mission Health & Rehab of Clinton	102 Woodchase Park Drive	Clinton, MS 39056	601 924-7043
Willow Creek Retirement Center	49 Willow Creek Lane	Jackson, MS 39272	601 863-4201
Holmes County LTC Center - Durant	15481 Bowling Green Road	Durant, MS 39063	662 653-4106
Lexington Manor Senior Care	56 Rockport Road	Lexington, MS 39095	662 834 3021
Humphreys County Nursing Center	500 CCC Road	Belzoni, MS 39038	662 247-1821
Courtyards Community Living Center	907 East Walker Street (PO Box 69)	Fulton, MS 38843	662 862-6140
The Meadows (Daniel Health Care d/b/a)	1905 South Adams Street (PO Box 127)	Fulton, MS 38843	662 862-2165
Ocean Springs Nursing Center	1199 Ocean Springs Road	Ocean Springs, MS 39564	228 875-9363
Plaza Community Living Center	4403 Hospital Road	Pascagoula, MS 39581-5335	228 762-8960

River Chase Village	5090 Gautier Vancleave Road	Gautier, MS 39533	228 522-6700
Singing River Rehab. & Nursing Center	3401 Main Street	Moss Point, MS 39563	228 762-7451
Sunplex Subacute Center	6520 Sun Scope Drive	Ocean Springs, MS 39565	228 875-1177
Jasper County Nursing Home	15 South Sixth Street (PO Box 527)	Bay Springs, MS 39422	601 764-2101
Jefferson County Nursing Home	910 Main Street (PO Box 1089)	Fayette, MS 39069	601 786-3888
Jefferson Davis Community Hospital ECF	1320 Winfiled Street (PO Box 1288)	Prentiss, MS 39474	601 792-1172
Care Center of Laurel	935 West Drive	Laurel, MS 39440	601 649-8006
Comfort Care Nursing Center	1100 West Drive	Laurel, MS 39440	601 422-0022
Jones County Rest Home	683 County Home Road	Ellisville, MS 39437	601 477-3334
Laurelwood Community Living Center	1036 West Drive	Laurel, MS 39440	601 425-3191
MS Care Center of Dekalb	220 Willow Avenue (PO Box 577)	DeKalb, MS 39328	601 743-5888
Graceland Care Center of Oxford	1301 Belk Blvd.	Oxford, MS 38655	662 234-7821
MS State Veterans Home – Oxford	120 Veterans Drive	Oxford, MS 38655	662 236-1218
Lamar Healthcare & Rehabilitation Center	6428 U. S. Highway 11	Lumberton, MS 39455	601 794-8566
The Windham House of Hattiesburg	37 Hillcrest Drive	Hattiesburg, MS 39402	601 264-0058
Wesley Medical Center TCU	5001 Hardy Street	Hattiesburg, MS 39402	601 268-8000
Bedford Care Center of Marion	6434-A Dale Drive	Marion, MS 39342	601 294-3515
Golden Living Center - Meridian	4728 Highway 39 North (PO Box 3604)	Meridian, MS 39301	601 482-8151
James T. Champion Nursing Facility	1455 North Lakeland Drive	Meridian, MS 39307	601 581-8450
Meridian Community Living Center	517 33rd Street	Meridian, MS 39305	601 483-3916
Poplar Springs Nursing Center	6615 Poplar Springs Drive (PO Box 3623)	Meridian, MS 39305	601 483-5256
Queen City Nursing Center	1201 28th Avenue	Meridian, MS 39301	601 483-1467
Reginald P. White Nursing Facility	1451 North Lakeland Drive (PO Box 4128)	Meridian, MS 39307	601 581-8500
The Oaks Rehabilitation & Healthcare Center	3716 Highway 39 North	Meridian, MS 39301	601 482-7164
Lawrence County Nursing Center	700 South Jefferson (PO Box 398)	Monticello, MS 39654	601 587-2593
Golden LivingCenter - Carthage	1101 East Franklin Street	Carthage, MS 39051	601 267-4551
Leake Memorial Extended Care Unit	310 Ellis Street (PO Box 909)	Carthage, MS 39051	601 267-1356

Cedars Health Center	2800 West Main Street	Tupelo, MS 38801	662 844-1441
Golden LivingCenter- Eason Boulevard	2273 S. Eason Boulevard	Tupelo, MS 38804	662 842-2461
North MS Medical Center / Baldwyn Nursing Facility	739 4th Street South	Baldwyn, MS 38824	662 365-4091
North MS Medical Center SNF	830 South Gloster	Tupelo, MS 38801	662 377-3000
Tupelo Nursing & Rehabilitation Center	1901 Briar Ridge Road	Tupelo, MS 38804	662 844-0675
Crystal Health & Rehabilitation of Greenwood	902 Sgt John A. Pittman Drive	Greenwood, MS 38930-7343	662 453-9173
Golden Age	2901 Highway 82 East	Greenwood, MS 38930	662 453-6323
Greenwood Leflore Sub acute Unit	1401 River Road (PO Box 1410)	Greenwood, MS 38930	662 459-2698
Riverview Nursing & Rehabilitation Center	1600 West Claiborne Avenue Extended	Greenwood, MS 38930	662 453-8140
Country brook Living Center	525 Brookman Drive (PO Box 3369)	Brookhaven, MS 39601	601 833-2330
Golden Living Center-Brook Manor	519 Brookman Drive	Brookhaven, MS 39601	601 833-2881
Haven Hall Healthcare Center	101 Mills Street	Brookhaven, MS 39601	601 833-5608
Silver Cross Home	503 Silver Cross Drive (PO Box 617)	Brookhaven, MS 39601	601 833-2361
Aurora Health and Rehabilitation	310 Emerald Drive	Columbus, MS 39702	662 327-8021
Baptist Memorial Hospital	2520 5th Street North	Columbus, MS 39701	662 244 1500
The Windsor Place Nursing & Rehab Center	81 Windsor Blvd.	Columbus, MS 39702	662 241-5518
Trinity Healthcare Center	230 Airline Road	Columbus, MS 39702	662 327-9404
Vineyard Court Nursing Center	2002 5th Street North	Columbus, MS 39705	662 328-1133
Highland Home	638 Highland Colony Parkway	Ridgeland, MS 39157-8724	601 853-0415
Madison County Nursing Home	1421 E. Peace Street (PO Box 488)	Canton, MS 39046	601 855-5760
St. Catherine's Village-Siena Center	200 Dominican Drive	Madison, MS 39110	601 856-0100
The Arbor	600 S. Pear Orchard Road	Ridgeland, MS 39157	601 856-2205
The Nichols Center	1308 Highway 51 North	Madison, MS 39110	601 853-4343
Song Health & Rehab of Columbia, LLC	1506 North Main Street	Columbia, MS 39429	601 736-9557
The Grove	11 Pecan Drive	Columbia, MS 39429	601 736-4747
The Myrtles Nursing Center	1018 Alberta Avenue	Columbia, MS 39429	601 731-1745

Trinity Mission Health & Rehab of Great Oaks	111 Chase Street	Byhalia, MS 38611	662 838-3670	
Trinity Mission Health & Rehab of Holly Springs	1315 Highway 4 East	Holly Springs, MS 38635	662 252-1141	
Care Center of Aberdeen	505 Jackson Street (PO Box 211)	Aberdeen, MS 39730	662 369-6431	
Golden LivingCenter - Amory	1215 Earl Frye Boulevard	Amory, MS 38821	662 256-9344	
River Place Nursing Center	1126 Earl Frye Boulevard	Amory, MS 38821	662 257-9919	
Winona Manor	627 Middleton Road	Winona, MS 38967	662 283-1260	
Choctaw Residential Center	135 Hospital Circle	Philadelphia, MS 39350-6780	601 656-2582	
Hilltop Manor Nursing Center	101 Kirkland Street	Union, MS 39365	601 774-8233	
Neshoba County Nursing Home	1001 Holland Avenue (PO Box 648)	Philadelphia, MS 39350	601 663-1200	
Bedford Care Center of Newton	1009 South Main Street	Newton, MS 39345	601 683-6601	
J. G. Alexander Nursing Center	25112 Highway 15	Union, MS 39365	601 774-5065	
Noxubee County Nursing Home	606 North Jefferson Street (PO Box 480)	Macon, MS 39341	662 726-2097	
Starkville Manor Nursing Home	1001 Hospital Road (PO Box 1466)	Starkville, MS 39760	662 323-6360	
The Carrington	307 Reed Road	Starkville, MS 39759	662 323-2202	
Golden LivingCenter - Batesville	154 Woodland Road	Batesville, MS 38606	662 563-5636	
Sardis Community Nursing Home	613 East Lee Street	Sardis, MS 38666	662 487-2720	
Covenant Health & Rehabilitation Center	1620 Read Road (PO Box 937)	Picayune, MS 39466	601 798-1811	
Pearl River County Nursing Home	305 West Moody Street (PO Box 392)	Poplarville, MS 39470	601 795-4543	
Perry County Nursing Center	202 Bay Avenue West	Richton, MS 39476	601 788-2490	
Camellia Estates	1714 White Street	McComb, MS 39648	601 250-0066	
Courtyard Rehabilitation and Healthcare Street	501 South Locust	McComb MS, 39648	601 684 8111	
McComb Nursing & Rehabilitation Center	415 Marion Avenue	McComb, MS, 39648-1169	601 684-8700	
Graceland Care Center of Pontotoc	278 W. Eight Street (PO Box 547)	Pontotoc, MS 38863	662 489-6411	
NMMC - Pontotoc Nursing Home	176 South Main Street (PO Box 790)	Pontotoc, MS 38863	662 489-5510	
Sunshine Health Care	1677 Highway 9 North	Pontotoc, MS 38863	662 489-1189	
Landmark Nursing & Rehab. Center, The	100 Lauren Drive	Booneville, MS 38829	662 720-0972	
Longwood Community Living Center	200 Long Street	Booneville, MS	662 728-6234	

	(PO Box 326)	38829		
Quitman County Nursing Home	350 Getwell Drive	Marks, MS 38646	662 326-3690	
Brandon Court	100 Burnham Road	Brandon, MS 39042	601 664-2259	
Brandon Nursing & Rehabilitation Center	355 Crossgate Boulevard	Brandon, MS 39042	601 825-3192	
Briar Hill Rest Home	1201 Gunter Road	Florence, MS 39073	601 939-6371	
Jaquith Nursing Home - Adams Inn	3550 Highway 468 West - Bldg. 31 & 48 (PO Box 207)	Whitfield, MS 39193-0207	601 351-8081	
Jaquith Nursing Home – Jaquith Inn	3550 Highway 468 West - Bldg. 69 & 78 (PO Box 207)	Whitfield, MS 39193-0207	601 351-8411	
Jaquith Nursing Home - Madison Inn	3550 Highway 468 West - Bldg. 28 & 34 (PO Box 207)	Whitfield, MS 39193-0207	601 351-8151	
Jaquith Nursing Home - Monroe Inn	3550 Highway 468 West - Bldg. 40 & 41 (PO Box 207)	Whitfield, MS 39193-0207	601 351-8136	
Jaquith Nursing Home -Jefferson Inn	3550 Highway 468 West - Bldg 29 & 33 (PO Box 207)	Whitfield, MS 39193-0207	601 351-8132	
Methodist Specialty Care Center	1 Layfair Drive, Suite 500	Flowood, MS 39232	601 420-7760	
Wisteria Gardens	5420 Highway 80 East	Pearl, MS 39208	601 420-7760	
Lackey Convalescent Home	266 First Avenue (PO Box 428)	Forest, MS 39074	601 469-4151	
MS Care Center of Morton	96 Old Highway 80 East (PO Box 459)	Morton, MS 39117	601 732-6361	
Heritage Manor of Rolling Fork	431 West Race Street (PO Box 279)	Rolling Fork, MS 39159	662 873-6218	
Bedford Care Center of Mendenhall	925 West Mangum Avenue	Mendenhall, MS 39114	601 847-1311	
Hillcrest Nursing Center	1401 First Avenue, N.E. (PO Box 398)	Magee, MS 39111	601 849-0384	
Rolling Acres Retirement Center	309 Magnolia Drive (PO Box 128)	Raleigh, MS 39153	601 782-4244	
Azalea Gardens Nursing Center	530 Hall Street	Wiggins, MS 39577	601 928-5281	
Stone County Nursing & Rehabilitation Center	1436 East Central Avenue	Wiggins, MS 39577	601 928-1889	
Liberty Health & Rehab of Indianola, LLC	401 Highway 82 West	Indianola, MS 38751	662 887-2682	
Ruleville Nursing & Rehabilitation Center	800 Stansel Drive (PO Box 368)	Ruleville, MS 38771	662 756-4361	
Tallahatchie General Hospital Extended Care Facility 201 South Market Street		Charleston, MS 38921	662 647 553	

Walter B. Crook Nursing Facility	840 North Oak Avenue (PO Box 369)	Ruleville, MS 38771	662 756-2711	
Senatobia Convalescent Center & Rehab	402 Getwell Drive	Senatobia, MS 38668	662 562-5664	
Golden Living Center - Ripley	101 Cunningham Drive	Ripley, MS 38663	662 837-9940	
Rest Haven Health & Rehabilitation	103 Cunningham Drive	Ripley, MS 38663	662 837-3062	
Tippah County Nursing Home	1005 City Avenue North (PO Box 499)	Ripley, MS 38663	662 837-9221	
Tishomingo Community Living Center	1410 West Quitman (PO Box 562)	Iuka, MS 38852	662 423-3422	
Tishomingo Manor	230 Khaki Avenue	Iuka, MS 38852	662 423-9112	
Tunica Nursing Home, LLC	1024 Highway 61 South	Tunica, MS 38676	662 363-3164	
Graceland Care Center of New Albany	118 South Glenfield Road	New Albany, MS 38652	662 534-9506	
Union County Health & Rehabilitation Center	1111 Bratton Road	New Albany, MS 38654	662 539-0502	
Billdora Senior Care	314 Enoch Street	Tylertown, MS 39667	601 876-2173	
Golden LivingCenter - Tylertown	200 Medical Circle (PO Box 112)	Tylertown, MS 39667	601 876-2107	
Covenant Health & Rehab of Vicksburg	2850 Porter's Chapel Road	Vicksburg, MS 39180	601 638-9211	
Heritage House Nursing Center	3103 Wisconsin Avenue (PO Box 820485)	Vicksburg, MS 39180	601 638-1514	
Shady Lawn Health & Rehabilitation	60 Shady Lawn Place	Vicksburg, MS 39180	601 636-1448	
Vicksburg Convalescent Center	1708 Cherry Street	Vicksburg, MS 39180	601 638-3632	
Arbor Walk Healthcare Center	570 North Solomon	Greenville, MS 38701	662 335 5863	
Legacy Manor Nursing & Rehab Center	1935 North Theobald Extended	Greenville, MS 38704	662 334-4501	
MS Care Center of Greenville	1221 East Union Street (PO Box 4767)	Greenville, MS 38701	662 335-5811	
River Height Healthcare Center	402 Arnold Avenue	Greenville, MS 38701	662 332 0318	
Washington Care Center	1920 Lisa Drive Extension	Greenville, MS 38703	662 335-2897	
Pine View Healthcare Center	1304 Walnut Street (PO Box 512)	Waynesboro, MS 39367	601 735-9025	
Golden Living Center – Eupora	200 Walnut Avenue	Eupora, MS 39744	662 258-8293	
Webster Health Services, Inc.	70 Medical Plaza	Eupora, MS 39744	662 258-9310	
Wilkinson County Senior Care	166 South Lafayette Street (PO Box 310)	Centreville, MS 39631	601 645-5253	

Louisville Healthcare, LLC	543 East Main Street (PO Box 452)	Louisville, MS 39339	662 773-8047	
Winston County Nursing Home	562 East Main Street (PO Box 967)	Louisville, MS 39339	662 773-6211	
Yalobusha County Nursing Home	630 S. Main Street (PO Box 728)	Water Valley, MS 38965	662 473-1411	
Martha Coker Green House Homes	2041 Grand Avenue	Yazoo City, MS 39194	662 746-4621	
Adams County Nursing Center	588 John R. Junkin Drive	Natchez, MS 39120	601 446-8427	

Personal Care Homes in Mississippi

Center Name	Address	Place	Phone Number
Arnold's Personal Care	722 N. Rankin Street	Natchez, MS 39120	(601) 442-7519
Magnolia House	311 Highland Boulevard	Natchez, MS 39120	(601) 446-5097
Country Cottage-Corinth	3002 N. Polk	Corinth, MS 38834	(662) 287-7811
Dogwood Plantation of Corinth	1101 Levee Road (PO Box 958 zip 38835)	Corinth, MS 38834	(662) 286-7021
Atwood Personal Care Home #2	328 Goodman Street	Kosciusko, MS 39090	(662) 289-2547
Cleveland Personal Care Community	800 3rd Street	Cleveland, MS 38732	(662) 846-6160
Indywood Estate	218 Ronaldman Road	Cleveland, MS 38732	(662) 843-7885
Lawson Personal Care Home	407 Old Hwy 61	Shaw, MS 38773	(662) 754-3278
Bruce Community Living Center	176 Highway 9 South (PO Box 1280)	Bruce, MS 38915-1280	(662) 412-5100
Fernbrooke Personal Care Home	127 CR 31	Houston, MS 38851	(662) 456-9400
Brooklyn Hall Personal Care Home	23283 Highway 15 South	Mathiston, MS 39752	(662) 263-4685
Paradise Personal Care Home	20114 Highway 18	Hermanville, MS 39086	(601) 535-2495
Golden Meadow	4043 County Road 110	Shubuta, MS 39360	(601) 776-2019
Wisteria Manor	421 CR 280	Shubuta, MS 39360	(601) 776-0426
Waverly Care Home	315 W. Broad (PO Box 797)	West Point, MS 39773	(662) 494-0074
Flowers Glen Personal Care	1251 Lee Drive	Clarksdale, MS 38614	(662) 627-2222
Covington Ridge	100 Covington Ridge Place	Collins, MS 39428	(601) 765-1100
Hermitage Gardens of Southaven	108 Clarington Avenue	Southaven, MS 38671	(662) 349-9043
Lifepointe Village	2782 Star Landing Road E	Southaven, MS 38672	(662) 429-7672
Olive Grove Terrace	9684 Goodman Road	Olive Branch, MS 38654	(662) 895-7609
Silvercreek Retirement Communities	6630 Crumpler Boulevard	Olive Branch, MS 38654	(662) 895-8352
Wesley Meadows Personal Care	1325 McIngvale Road (PO Box 487)	Hernando, MS 38632	(662) 429-2070
Country Comfort	330 Pop Runnels Road	Petal, MS 39465	(601) 545-7603
Emeritus at Forrest Park	103 Fox Chase	Hattiesburg, MS 39402	(601) 271-2177

Emeritus at Pine Meadow	107 Fox Chase Drive	Hattiesburg, MS 39402	(601) 271-8480
Hartford Place	705 Hall Avenue	Hattiesburg, MS 39401	(601) 545-8333
Amelia's Garden Assisted Living	145 Mill Street Ext	Lucedale, MS 39452	(601) 508-2493
Eva's Place	258 Vestry Road	Perkinston, MS 39573	(601) 945-5053
Smith Manor	15255 Hwy 613	Lucedale, MS 39452	(601) 947-7796
Sparrow Hills	4285 Beaver Dam Road	Lucedale, MS 39452	(601) 947-6586
Graceland of Grenada	1855 Hill Drive	Grenada, MS 38901	(662) 226-8556
Dunbar Village Courtyard	725 Dunbar Avenue	Bay St. Louis, MS 39520	(228) 466-3099
Woodland Village Nursing Center	5427 Gex Road	Diamondhead, MS 39525	(228) 255-4832
Chapman Oaks	210 Roberts Street	Long Beach, MS 39560	(228) 868-7199
Emeritus of Biloxi	2120 Enterprise Drive	Biloxi, MS 39531	(228) 388-0946
Lakeview Nursing Center Personal Care Home	16411 Robinson Road	Gulfport, MS 39503	(228) 831-3001
LeMoyne Place	14306 Lemoyne Blvd.	Biloxi, MS 39532	(228) 396-3336
One Magnolia	16391 Robinson Road	Gulfport, MS 39503	(601) 206-3293
Alpha & Omega Personal Care Home	131 S. Prentiss Street	Jackson, MS 39203	(601) 354-0783
Autumn Light Care Home	1458 Moncure Marble Road	Terry, MS 39170	(601) 878-9684
Central Mississippi Personal Care Home	915 Hunt Street (PO Box 3301 Zip 39207)	Jackson, MS 39203	(601) 355-8076
Community Welfare & Health Center	1208 Wiggins Street	Jackson, MS 39203	(601) 352-4721
Creation Elite Residential Home	350 Adelle Street	Jackson, MS 39202	(601) 421-4332
Eldercare	232 Moss Avenue	Jackson, MS 39209	(601) 969-2273
Emeritus At Trace Pointe	501 East Northside Drive	Clinton, MS 39056	(601) 926-1224
Erie Personal Care Home	606 Erie Street	Jackson, MS 39203	(601) 948-6862
Gena's Four Seasons	103 Jackson Street	Edwards, MS 39066	(601) 354-9186
Genesis Personal Care	2227 Old Vicksburg Road (PO Box 8842 Jackson, 39284)	Clinton, MS 39056	(601) 925-1348
Horizon Personal Care	438 Clifton St (PO Box 6774 Zip 39282)	Jackson, MS 39203	(601) 592-6174
House of Faith, LLC	928 Hunt Street (120 Falcon Ridge Dr,	Jackson, MS 39203	(601) 720-6692

	Raymond 39154)		
Love & Gracious Center Services	1625 Westhaven Boulevard	Jackson, MS 39209	(601) 487-8875
McAllister's Personal Care Home	425 Earle Street	Jackson, MS 39203	(601) 352-4271
Mississippi Cares Residential Home	432 Clifton Street (PO Box 10840 Zip 39289)	Jackson, MS 39203	(601) 948-8923
Myles Retreat Home for the Golden Age	5911 Holmes Circle	Jackson, MS 39213	(601) 956-4185
North Grove Assisted Living	641 Flag Chapel Road	Jackson, MS 39206	(601) 922-2008
Paradise Cove Personal Care Home	1874 Longwood Drive	Jackson, MS 39212	(601) 201-2985
Parker's Personal Care Home	532 Earle Street	Jackson, MS 39203	(601) 355-2030
Pope Personal Care Home	4314 Sunset Drive	Jackson, MS 39213	(601) 982-4012
Riggs Manor Retirement Community	2300 Seven Springs Road	Raymond, MS 39154	(601) 857-5011
Shanell's Assisted Living	4022 California Avenue (PO Box 83040, Jackson, MS 39283)	Jackson, MS 39213	(601) 362-4549
St. David's Personal Care Home	714 Rose Street (PO Box 6944 Zip 39282)	Jackson, MS 39203	(601) 354-0690
T's Personal Care Home, Inc.	3651 Mosley Avenue	Jackson, MS 39206	(601) 366-9147
Miss Bernice's House	50 Patton Place	Lexington, MS 39095	(662) 834-3344
Charleston Place	804 South Adams (PO Drawer 127)	Fulton, MS 38843	(662) 862-2465
Countrywood Manor Assisted Living	145 Watson Drive	Mantachie, MS 38855	(662) 282-7808
Dogwood Plantation of Fulton	201 W. Pierce Town Road (PO Box 620, Tupleo, MS 38802)	Fulton, MS 38843	(662) 862-6120
Phillips Personal Care Home	1207 Sandlin Road	Fulton, MS 38843	(662) 862-2665
Alternative Personal Care Home	6816 Washington Avenue	Ocean Springs, MS 39564	(228) 872-7022
Kare Med Assisted Living	2701 Catherine Drive (PO Box 506)	Ocean Springs, MS 39564	(228) 806-3283
Residence at Bay Cove	1753 Brasher Road (PO Box 775, Ocean Springs 39566)	Biloxi, MS 39532	(228) 702-0142
Seashore Oaks Assisted Living	1450 Beach Boulevard (PO Box 8056, Biloxi, MS 39535)	Biloxi, MS 39530	(228) 207-5225
Serenity Springs Personal Care Home	9405 Tucker Road	Biloxi, MS 39532	(228) 872-3373
Settler's Pointe	13625 Wilfred Seymour Road (PO Box 1804 Zip 39566)	Ocean Springs, MS 39564	(228) 872-1746

The Gardens	1260 Ocean Springs Road	Ocean Springs, MS 39564	(228) 818-0650
Summerland Manor	12 Summerland Road (PO Box 527)	Bay Springs, MS 39422	(601) 764-2130
Comfortcare Nursing Center	1100 West Drive	Laurel, MS 39440	(601) 422-0022
Cottonwood Manor	3147 Old Amy Road (41 South Hall Rd; Morton, MS 39117)	Laurel, MS 39443	(601) 425-0095
Extra Care Personal Care Home	354 Trace Road	Laurel, MS 39443	(601) 425-5599
Lynnwood Senior Care	12 Victory Road	Laurel, MS 39443	(601) 342-2502
Magnolia Gardens Assisted Living of Ellisville	303 East Ivy Street (PO Box 430)	Ellisville, MS 39437	(601) 477-9041
Magnolia Gardens Assisted Living of Laurel	845 West Drive	Laurel, MS 39440	(601)477-9041
Northview Villa Personal Care	625 Northview Drive	Laurel, MS 39440	(601) 426-3782
Emeritus at Oxford	100 Azalea Drive	Oxford, MS 38655	(662) 234-9600
Hermitage Gardens of Oxford	1488 Belk Avenue	Oxford, MS 38655	(662) 234-8244
Alden Pointe	2 Courtland Drive	Hattiesburg, MS 39402	(601) 296-9711
Magnolia Place	4901 Hwy 589 (PO Box 93)	Sumrall, MS 39482	(601) 758-0600
Provision Living at Hattiesburg	217 Methodist Blvd.	Hattiesburg, MS 39402	(601) 329-2030
The Windham House of Hattiesburg	37 Hillcrest Drive	Hattiesburg, MS 39402	(601) 264-0058
Aldersgate Personal Care Home	6600 Poplar Springs Drive (PO Box 3846)	Meridian, MS 39305	(601) 485-9484
Bee Hive Homes of Marion	5750 Dale Drive	Marion, MS 39342	(601) 482-8800
Emeritus at Silverleaf Manor	4555 35th Avenue	Meridian, MS 39305	(601) 483-4566
Fisher Care	5207 Zero Road (5035 Fisher Rd, Meridian, MS 39301)	Meridian, MS 39301	(601) 693-9619
Magnolia Home	1900 24th Avenue (PO Box 3064 Zip 39305)	Meridian, MS 39301	(601) 938-2435
McCoy Personal Care Home	919 35th Avenue	Meridian, MS 39301	(601) 693-4104
Heritage Manor Personal Care Home	2051 Fergerson Mill Road	Silver Creek, MS 39663	(601) 886-7251
Bee Hive Homes of Carthage	704 Highway 16 East	Carthage, MS 39051	(601) 267-8222
Avonlea Assisted Living & Retirement Community	2429 Lawndale Drive	Tupelo, MS 38801	(662) 840-6163
Magnolia Manor at Tupelo	1514 CR 41	Tupelo, MS 38801	(662) 842-6776
Mitchell Center	2800 West Main Street	Tupelo, MS 38801	(662) 844-1441
Riverbirch Residence	2554 Main Street (PO Box 159)	Plantersville, MS 38862	(662) 844-3451

Rosewood Residence	2441 McCollough Boulevard (PO Box 491)	Belden, MS 38826	(662)844-5856
Saltillo Assisted Living	200 Knight Drive	Saltillo, MS 38866	(662) 869-7009
Samaritan Gardens	2603 South Gloster Street	Tupelo, MS 38803	(662) 566-9974
Country Meadow Personal Care Home	4100 CR 164	Greenwood, MS 38930	(662) 453-4533
Indywood Glen	1416 Erie Street	Greenwood, MS 38930	(662) 455-3878
Haven Hall Assisted Living	101 Mills Street	Brookhaven, MS 39601	(601) 833-5608
New Dawn Retirement Center	3987 Silverton Trail, S.E.	Ruth, MS 39662	(601) 823-4020
Collegeview Personal Care Home	1323 College Street	Columbus, MS 39701	(662) 327-9463
Home Place of Columbus	2082 Yorkville Road East	Columbus, MS 39702	(662) 329-2772
New Horizon Residential Living Facility	200 Molly Lane (PO Box 5573 Zip 39704)	Columbus, MS 39702	(662) 241-4756
The Arrington	234 Windsor Blvd.	Columbus, MS 39702	(662) 241-0001
Trinity Place Personal Care Center	250 Airline Road	Columbus, MS 39702	(662) 327-6795
Emeritus at Ridgeland Pointe	410 Orchard Park	Ridgeland, MS 39157	(601) 957-0727
Old Ladies Home	7521 Old Canton Road (PO Drawer 720 Zip 39130)	Madison, MS 39110	(601) 856-1085
Seasons	1421-A East Peace Street Hwy 16 East (PO Box 488)	Canton, MS 39046	(601) 855-5760
St. Catherine's Village/Campbell Cove	200 Dominican Drive	Madison, MS 39110	(601) 856-0100
St. Catherine's Village/Marian Hall	200 Dominican Drive	Madison, MS 39110	(601) 856-0100
The Blake at Township	608 Steed Road	Ridgeland, MS 39157	(601) 500-7955
The Orchard Personal Care Home	600 South Pear Orchard Road	Ridgeland, MS 39157	(601) 856-2205
Willard F. Bond Home	7521 Old Canton Road (PO Drawer 720 Zip 39130)	Madison, MS 39110	(601) 856-8041
The Grove Personal Care Home	11 Pecan Drive	Columbia, MS 39429	(601) 736-4747
Christopher's Personal Care Home	885 Highway 178 East	Holly Springs, MS 38635	(662) 551-1122
Garden Suites Assisted Living	400 South Chestnut (PO Box 747)	Aberdeen, MS 39730	(662) 319-2085
Oak Tree Plantation	60139 Cotton Gin Port Road	Amory, MS 38821	(662) 256-8406
Pressicare Personal Care Home	20434 Old Houston Road	Aberdeen, MS	(662) 369-0070

		39730	
Atwood Personal Care Home of Philadelphia	101 Pilot Street	Philadelphia, MS 39350	(601) 656-7394
Bee Hive Homes of Philadelphia	708 Columbus Avenue	Philadelphia, MS 39350	(601) 656-0220
Bee Hive Homes of Newton	601 South Main Street	Newton, MS 39345	(601) 479-1267
River Birch Estate Assisted Living	606 East Jackson Road	Union, MS 39365	(601) 562-3402
Woodland Court	260 Northside Drive	Newton, MS 39345	(601) 683-2330
Elderly Care Center	496 Magnolia Drive	Macon, MS 39341	(662) 726-2630
Oakwood Manor	355 North Pine St. (PO Box 325)	Brooksville, MS 39739	(662) 738-4866
Montgomery Gardens Senior Care	4348 Old Highway 12 West	Starkville, MS 39759	(662) 323-4663
Vickers Personal Care Homes	114 North Montgomery Street	Starkville, MS 39759	(662) 323-4617
Fairfield of Batesville	640 Keating Road (PO Box 1632)	Batesville, MS 38606	(662) 563-2345
Blueberry Hill	1005 South Shivers	Poplarville, MS 39470	(601) 403-8177
New Country Living Personal Care Home	258 George Ford Road	Carriere, MS 39426	(601) 798-5673
Our Place Personal Care Home	27 Alsobrooks Road	Picayune, MS 39466	(601) 799-3303
Shady Oaks of Carriere	83 White Chapel Road	Carriere, MS 39426	(601) 798-0207
Southern Oaks Personal Care Home	174 Cliff Mitchell Road	Picayune, MS 39466	(601) 898-9998
Southern Pines Personal Care Home	340 Sycamore Road	Carriere, MS 39426	(601) 798-9969
Community Development Assisted Living	109 Elm Street (PO Box 689)	Richton, MS 39476	(601) 788-5865
Community Development Assisted Living II	200 North Front Street (PO Box 689)	Richton, MS 39476	
Friend's Personal Care Home	367 Corinth Church Road	Petal, MS 39465	(601) 545-9545
Aston Court Retirement Community	222 Aston Avenue	McComb, MS 39648	(601) 249-0023
Camellia Estates	1714 White Street	McComb, MS 39648	(601) 250-0066
Church Street Personal Care Home of Ecru	36 Elm Lane	Ecru, MS 38841	(662) 489-6462
Scott Del Cottage	1101 West Chambers Drive	Booneville, MS 38829	(662) 720-9593
The Landmark Community Personal Care Home	701 W. Church Street	Booneville, MS 38829	(662) 728-3539
Brandon Court	100 Burnham Road	Brandon, MS 39042	(601) 664-2259
Briar Hill Rest Home Personal Care	1201 Gunter Road	Florence, MS	(601) 939-6371

Home		39073	
Emeritus at Heritage House	140 Castlewoods Boulevard	Brandon, MS 39047	(601) 919-1208
Peach Tree Village	6100 Old Brandon Road	Brandon, MS 39042	(601) 933-1100
Villa South Assisted Living	271 College Street	Florence, MS 39073	(601) 845-1888
Wisteria Gardens	5420 Highway 80 East	Pearl, MS 39208	(601) 988-6800
Bee Hive Homes of Forest	410 Townsend Road	Forest, MS 39074	(601) 469-9476
Magnolia Manor	410 First Street	Forest, MS 39074	(601) 469-4389
Bryant's Residential Care Facility, Inc.	1310 Highway 541 South (PO Box 972)	Magee, MS 39111	(601) 849-6160
Chateau Jorja	219 Bass Road	Florence, MS 39073	(601) 845-5619
Christopher's Court	219 Bass Road	Florence, MS 39073	(601) 845-5619
Lakeview Place	1116 Frances Avenue	Magee, MS 39111	(601)849-1920
Stone County Nursing & Rehabilitation Center	1436 East Central Ave.	Wiggins, MS 39577	(601) 928-1889
Indywood	541 Dorsett Drive	Indianola, MS 38751	(662) 887-3005
Ruleville Nursing & Rehabilitation Center	800 Stansel Drive	Ruleville, MS 38771	(662) 756-4361
Shirley's Personal Care Home	100 BB King Drive	Indianola, MS 38751	(662) 887-3000
Providence PCC of Senatobia	700 Moore's Crossing	Senatobia, MS 38668	(662) 562-9229
Carrington House	1670 Whitehouse Road	Iuka, MS 38852	(662) 423-3307
Southern Magnolia Estates I	1308 North Pearl Street	Iuka, MS 38852	(662) 424-0023
Southern Magnolia Estates II	1308 North Pearl St.	Iuka, MS 38852	(662) 424-0023
Southern Magnolia Estates II of Belmont-Golden	267 Front Street Hwy 366	Golden, MS 38847	(662) 454-0544
Dogwood Plantation of New Albany	250 Fairfield Drive	New Albany, MS 38652	(662) 534-7331
Magnolia Place of New Albany	1515 Munsford Road (PO Box 620 Zip 38802)	New Albany, MS 38652	(662) 534-0046
Sunshine Inn, Inc.	1645 Hwy 178 (PO Box 378)	Myrtle, MS 38650	(662) 988-3959
Belmont Gardens	3102 Wisconsin Avenue (PO Box 820874)	Vicksburg, MS 39182	(601) 636-8006
Heritage House Assisted Living Center	3103 Wisconsin Avenue (PO Box 820485)	Vicksburg, MS 39182	(601) 638-1514
Destiny Manor	104 Martin Luther King Drive (PO Box 383)	Arcola, MS 38722	(662) 378-7962
Magnolia Gardens of Greenville	1644 South Colorado Street	Greenville, MS 38703	(662) 335-9699
Restorations Assisted Living Facility	1766 Old Leland Road (308 Camellia Ln,Indianola, MS 38751)	Greenville, MS 38703	(662) 537-4976

Wellington Place of Greenville	1880 Fairground Road	Greenville, MS 38703	(662) 334-4646
Brookwood Villa	915 Wayne Street	Waynesboro, MS 39367	(601) 735-0264
Southern Living Specialty Care	511 Gray Drive	Waynesboro, MS 39367	(601) 735-0120
Vickers Personal Care Home	238 Spring Valley Road	Mathiston, MS 39752	(662) 263-4685
Bee Hive Homes of Louisville	541 A East Main Street (PO Box 1883 Zip 39302)	Louisville, MS 39339	(662) 773-0000
Lakeside Village	8862 Highway 15 South (PO Box 401 Noxapater, MS 39346)		(662) 724-2500

Portable X-RAY Mobile Services in Mississippi

Mobile Service Name	Address	Place	Phone Numbers
American Mobile Imaging	2818 22nd Avenue, Suite 11 *Alabama's Address: 5766 Carmichael Pkwy*	Gulfport, MS 39501 *Montgomery, AL 36117*	(228) 248-0090 MS (334) 269-3322, Ext 221 AL
Mobile Care	200 Rawls Drive, Suite 300 (100 Laurel Street)	McComb, MS 39648	(601) 249-0911 MS 1-800-960-3238 LA
Mobile Services, Inc.	1807 24th Avenue	Meridian, MS 39301	(601) 693-4893
Portable Medical Diagnostics	1855 Lakeland Drive, #G10	Jackson, MS 39216	(601) 987-9729
Southern Radiology Services	304 South Spring Street, Suite E	Tupelo, MS 38801	(662) 841-2820 1-800-845-8183

Psychiatric Residential Treatment Center in Mississippi

Name of Center	Address	Place	Phone Number
Cares Center	402 Wesley Avenue	Jackson, MS 39202	(601) 360-0583
Diamond Grove Center, Psychiatric Residential Treatment Center	2311 Highway 15 South (PO Box 848)	Louisville, MS 39339	(662) 779-0119
Millcreek of Pontotoc	1814 Highway 15 North (PO Box 619)	Pontotoc, MS 38863	(662) 488-8878
Millcreek Psychiatric Residential Treatment Facility	900 First Avenue, N.E. (PO Box 1160)	Magee, MS 39111	(601) 849-4221
Parkwood Behavioral Health System	8135 Goodman Road (PO Box 766)	Olive Branch, MS 38654	(662) 895-4900
Specialized Treatment Facility	14426 James Bond Road	Gulfport, MS 39503	(228) 328-6000
The Crossings	5000 Highway 39 North	Meridian, MS 39301	(601) 483-5452

Rehabilitation Centers in Mississippi

Name	Address	Place	Phone Number
360 Total Rehab	2625 Courthouse Circle	Flowood, MS 39208	(601) 932-8555
Ability Rehab Group	14306 LeMoyne Blvd.	Biloxi, MS 39350	(228) 396-3330
Batson Physical Therapy	711 Hall Street	Wiggins, MS 39577	(601) 928-5511
Coastal Rehabilitation of South MS	15190 Community Road, Suite 110	Gulfport, MS 39503	(228) 831-4646
Cross Creek Physical Therapy Limited Partnership	7501 Goodman Road, Suite I	Olive Branch, MS 38654	(662) 890-3382
Crossroads Rehabilitation Services, Inc.	206B Oxford Road	New Albany, MS 38652	(662) 534-4445
Encore Rehabilitation, Inc.	2210B Mill Street Extension	Lucedale, MS 39452	(601) 947-9005
First Place Physical Therapy, LLC	433 Broad Street	Columbia, MS 39429	(601) 444-0030
Fulton Rehabilitation Services, PA	1110 South Adams Street (PO Box 455)	Fulton, MS 38843	(662) 862-4104
Genesis Physical Therapy and Rehab Services	227 Highway 51	Ridgeland, MS 39157	(601) 898-4324
Grenada Rehab Works	1300 Sunset Drive, Suite G	Grenada, MS 38901	(662) 227-9748
GT Physical Therapy, Inc.	501 East Main Street	Louisville, MS 39339	(662) 773-3700
Innovative Therapies, Inc.	12303 Hwy 49	Gulfport, MS 39503	(228) 832-6221
Key Rehab Associates, Inc.	123 Jefferson Davis Blvd.	Natchez, MS 39120	(601) 445-0005
Magnolia Outpatient Rehabilitation	2205 5th Street North	Columbus, MS 39701	(662) 243-1097
Medicomp Physical Therapy - Brandon	2015 Highpointe Drive	Brandon, MS 39042	(601) 824-8814
Outpatient Rehab Center of Fulton	204 Wheeler Drive	Fulton, MS 38843	(662) 862-3070
Performance Rehab, Inc.	7213 Siwell Road	Byram, MS 39272	(601) 346-9191
Physiotherapy Associates, Inc. Brookhaven	631 Brookway Blvd.	Brookhaven, MS 39601	(601) 833-7317
Physiotherapy Associates, Inc. Monticello	314 Main Street, Suite C	Monticello, MS 39654	(601) 587-2563
Physiotherapy Associates, Southaven	7900 Airways Blvd.	Southaven, MS 38671	(662) 536-4096
Quality Rehabilitation, Inc.	1137 Main Street	Fayette, MS 39069	(601) 786-9133
Quest Rehab, Inc.	1440 East Central Avenue	Wiggins, MS 39577	(601) 928-6740

Restore Physical Therapy	10 Melody Lane	Collins, MS 39428	(601) 765-2900
River City Rehabilitation, LLC	1707 South Colorado Street, Suite A	Greenville, MS 38703	(662) 335-8332
Sunbelt Rehab Systems, Inc.	3688 Veterans Memorial Drive, Suite 300	Hattiesburg, MS 39401	(601) 543-0221
Sunplex Subacute Center	6520 Suncope Drive	Ocean Springs, MS 39564	(228) 875-1177
The Summit Health & Rehab Services, Inc.	4109 Hwy 98 West	Summit, MS 39666	(601) 276-3900
The Therapy Group	126 W. Gallatin Street	Hazlehurst, MS 39083	(601) 894-5929
Total Rehab Plus, Inc.	4387 Leisure Time Drive	Dimondhead, MS 39525	(228) 255-3536
Tri Vista Rehab, Inc. (Shiloh Ridge Athletic Club)	3303 Shiloh Ridge Road	Corinth, MS 38834	(662) 287-5662
Trinity Rehabilitation	133 Executive Drive, Suite D	Madison, MS 39110	(601) 956-0607

Rural Health Centers in Mississippi

Name	Address	Place	Phone Number
Natchez Rural Health Clinic	500 Martin Luther King Street	Natchez, MS 39120	(601) 446-7332
Pediatric & Adolescent Clinic	308 Highland Blvd.	Natchez, MS 39120	(601) 442-7676
Family Acute Care Center	2045 East Shiloh Road	Corinth, MS 38834	(662) 286-5112
Family Clinic of Rienzi, The	82 Main Street (PO Box 194)	Rienzi, MS 38865	(662) 462-8600
Medi-Stat Clinic	703 Alcorn Drive, Suite 109	Corinth, MS 38834	(662) 286-1499
Tri State Rural Health Clinic	502 Alcorn Drive	Corinth, MS 38835	(662) 287-5216
FMCH Gloster Clinic	434 North Captain Gloster Dr.	Gloster, MS 39638	(601) 225-4711
Hickory Flat Association Clinic	407 Oak Street	Hickory Flat, MS 38633	(601) 333-6387
Hickory Flat Family Clinic	250 Oak Avenue	Hickory Flat, MS 38633	(662) 333-6387
Rosedale Family Medical Clinic	512 Levee Street	Rosedale, MS 38769	(662) 759-6806
Family Medical Clinic of Vardaman	310 West Sweet Potato Street	Vardaman, MS 38878	(662) 682-7555
Family Medical Clinic of Houston	105 Hillcrest Drive	Houston, MS 38851	(662) 456-5008
Family Medical Clinic of Okolona	521 West Drive	Okolona, MS 38860	(662) 4471405
Trace Family Health & Internal Medicine	1002 East Madison	Houston, MS 38851	(662) 456-2800
Woodland Clinic	120 Market Street (PO Box 186)	Woodland, MS 39776	(662) 456-0111
CCMC Rural Health Clinic Weir	547 Front Street	Weir, MS 39772	(662) 547-9677
Patient's Choice Clinic of Port Gibson	123 McComb Avene	Port Gibson, MS 39150	(601) 437-5141
Medical Group of Quitman, The	305 South Archusa	Quitman, MS 39355	(601) 776-2123
The Woman's Clinic	2000 North State Street	Clarksdale, MS 38614	(662) 627-7361
Crystal Springs Clinic	123 Bobo Drive	Crystal Springs, MS 39086	(601) 892-2225
Hazlehurst Clinic	213 Caldwell Drive	Hazlehurst, MS 39083	(601) 894-661
Collins Family Practice Clinic	704 Fifth Street	Collins, MS 39428	(601) 765-4414
Family Clinic of Seminary	215 Bobby Beasley Street	Seminary, MS 39479	(601) 722-4300
Family Medical Associates of Covington County	701 South Holly Avenue	Collins, MS 39428	(601) 765-3180
Green Tree Family Medical Center	603 Main Street (PO Box 1107)	Mount Olive, MS 39119	(601) 797-3405
Runnelstown Clinic	5034 Hwy 29	Petal, MS 39465	(601) 583-1553
Family Medical Group of Meadville	Highway 84 & Union Church Road (PO Box 636)	Meadville, MS 39653	(601) 384-5801

Community Medical Center	92 W. Ratcliff Street, Suite A	Lucedale, MS 39452	(601) 947-8181
Lucedale OB/GYN Center	92 Ratcliff Street, Suite B	Lucedale, MS 39452	(601) 947-6000
Greene County Family Medical Clinic	1017 Jackson Avenue	Leakesville, MS 39451	(601) 394-2820
Grenada Primary Care Clinic	965 JK Avent Drive, Suite 105	Grenada, MS 38901	(662) 227-7575
Women's Health Clinic of Grenada, Inc.	1401 Oak Street	Grenada, MS 38901	(662) 226-4010
Hancock Family Care Center	16230 Hwy 603, Suite G	Kiln, MS 39556	(228) 255-5200
Hancock Medical Services	3068 Port & Harbor Drive	Bay St. Louis, MS 39520	(228) 467-8688
Bolton Family Clinic	Corner of Madison and Depot (PO Box 217)	Bolton, MS 39041	(601) 866-7733
Minor Med Care, Raymond	120 West Main Street (PO Box 1223)	Raymond, MS 39154	(601) 857-2341
Charles W. Campbell Rural Health Clinic	102 Carrollton Street	Lexington, MS 39095	(662) 834-1721
Durant Primary Care Clinic	638 Northwest Avenue	Durant, MS 39063	(662) 653-1002
Holmes County Med Clinic - West	18295 Emory Road	West, MS 39192	(662) 967-2462
Holmes Family Med Clinic	239 Bowling Green Road	Lexington, MS 39095	(662) 834-1321
Internal Medicine Clinic of Lexington	115 West China Street	Lexington, MS 39095	(662) 834-3956
Lexington Primary Care Clinic	110 Tchula Street	Lexington, MS 39095	(662) 834-1855
Gorton Rural Health	107 Church Street	Belzoni, MS 39038	(662) 247-2105
East Central Medical Center-RHC	7001 Highway 614	Hurley, MS 39555	(228) 588-6622
Bay Springs After Hours Family Health Clinic	31 East 5th Avenue	Bay Springs, MS 39422	(601) 764-2143
Jefferson Davis Community Hospital Family Medicine of Prentiss	1014 Rose Street, Suite A	Prentiss, MS 39474	(601) 792-2200
Prentiss Family Practice Clinic	1014 Rose Street, Suite D	Prentiss, MS 39474	(601) 792-2072
Laurel Pediatric & Adolescent	234 South 12th Avenue	Laurel, MS 39440	(601) 649-3520
South Central Ellisville Medical Clinic	103 Avenue B	Ellisville, MS 39437	(601) 477-8553
Children's Clinic Oxford	2888 South Lamar	Oxford, MS 38655	(662) 234-8286
Toccopola Family Medical Clinic	7908 Hwy 334 (PO Box 389)	Toccopola, MS 38874	(662) 488-0270
Family Clinic of Purvis	101 Weems Street	Purvis, MS 39475	(601) 794-2224
Purvis Family Practice Clinic	102 Shelby Speights Drive	Purvis, MS 39475	(601) 794-8065
Central MS Family Health Clinic	905-C South Frontage Road	Meridian, MS 39301	(601) 486-4210
East Mississippi Medical Clinic	4711 Poplar Springs Drive	Meridian, MS 39305-2666	(601) 485-7777
Immediate Care Family Clinic	1710 14th Street	Meridian, MS 39301	(601) 482-9211
North Hill Family Medical Clinic	5009 Highway 493	Meridian, MS 39305	(601) 626-8874

Rush Medical Clinic – Collinsville	9097 Collinsville Road	Collinsville, MS 39325	(601) 626-8374
Lawrence County Family Practice	1135 East Broad Street	Monticello, MS 39654	(601) 249-2701
Leake Memorial RHC	302 Ellis Street	Carthage, MS 39051	(601) 267-1385
Adults & Children Medical Clinic	733 North 4th Street	Baldwyn, MS 38824	(662) 365-3431
Family Care Medical Clinic	109 Parkgate Ext.	Tupelo, MS 38801	(662) 840-4175
Nurse Med, Inc.	1031 Northridge Road	Baldwyn, MS 38824	(662) 365-9305
Plantersville Family Clinic	2464 Main Street (PO Box 219)	Plantersville, MS 38862	(662) 842-4877
Shannon Family Medical Clinic, LLC	219 Broad Street	Shannon, MS 38868	(662) 767-8840
Twin Care Family Clinic, LLC	2686 Hwy 124 South, Suite B	Saltillo, MS 38866	(662)-869-8693
EMS Clinic	1509 Strong Avenue	Greenwood, MS 38930	(662) 455-4411
GLH-Magnolia Medical Clinic	1413 Strong Avenue	Greenwood, MS 38930	(662) 459-1207
Golden Age Clinic	2901 Highway 82 East	Greenwood, MS 38930	(662) 374-2185
Greenwood Leflore After Hours Clinic	1601 Strong Avenue	Greenwood, MS 38930	(662) 451-7565
Itta Bena Clinic	103 Basket Street	Itta Bena, MS 38941	(662) 254-7717
Canton Family Clinic	120 East Academy Street	Canton, MS 39046	(601) 859-2611
Canton Physicians Group	1421 East Peace Street, Suite A	Canton, MS 39046	(601) 855-5261
L.C. Tennin, Jr. MD, PA	1883 Hwy 43 South, Suite D (PO Box 647)	Canton, MS 39046	(601) 859-8992
Madison Canton Medical Clinic	1317 East Peace Street, Suite A	Canton, MS 39046	(601) 859-9888
Patient's Choice Clinic of Canton	1360 East Peace Street	Canton, MS 39046	(601) 859-9544
Columbia Family Clinic	502 Broad Street	Columbia, MS 39429	(601) 736-8282
Internal Medicine Clinic of Columbia	914 Sumrall Road	Columbia, MS 39429	(601) 731-1470
Woman's Pavilion of South MS, PLLC	1212 Broad Street	Columbia, MS 39429	(601) 736-6137
Health 1st Family Medical Clinic	2422 Church Street	Byhalia, MS 38611	(662) 838-5565
Williams Medical Clinic	538 Access Road (PO Box 5040)	Holly Springs, MS 38635	(662) 252-1599
Williams Medical Clinic of Potts Camp	39 Center Street (PO Box 40)	Potts Camp, MS 38659	(662) 333-6933
Aberdeen Health Clinic	501 Chestnut Street	Aberdeen, MS 39730	(662) 369-6131
Chestnut Medical Clinic	502 South Chestnut Street	Aberdeen, MS 39730	(662) 369-9525
Evergreen Clinic	Route 3, Box 379-M	Nettleton, MS	(662) 963-9154

		38858	
Nina Jurney's Family Medical Practice	502 Jackson Street, Suite 4	Aberdeen, MS 39730	(662) 369-9945
Pioneer Family Medical	502 Jackson Street, Suite 5	Aberdeen, MS 39730	(662) 369-9500
Pioneer Family Medical of Amory	1506 Hwy 278 East, Suite A	Amory, MS 38821	(662) 304-4027
Pioneer Family Medical of Caledonia	771 Main Street	Caledonia, MS 39740	(662) 356-4621
Pioneer Family Medical of Hamilton	40128 Hamilton Road	Hamilton, MS 39746	(662) 343-5129
Kilmichael Clinic	301 Lamar	Kilmichael, MS 39747	(662) 262-4284
Fairchild-Clearman Medical Ass., RHC	1122 East Main St reet, Suite 4	Philadelphia, MS 39350	(601) 656-1002
Alliance-Laird Clinic	25155 Hwy 15	Union, MS 39365	(601) 774-1513
Decatur Medical Clinic	68 4th Avenue	Decatur, MS 39327	(601) 635-2258
East MS Medical Clinic	9425 Eastside Drive Ext., Suite A	Newton, MS 39345	(601) 635-3333
Family Medical Group of Union	24345 Highway 15	Union, MS 39365	(601) 774-8211
Newton Family & Specialty Clinic	208 South Main Street	Newton, MS 39345	(601) 683-6041
Newton Family Practice Clinic	252 Northside Drive	Newton, MS 39345	(601) 683-3117
Brooksville Primary Care Clinic, Inc.	139 North Oliver Street	Brooksville, MS 39739	(662) 738-4424
Macon Primary Clinic	606 North Jefferson Street	Macon, MS 39341	(662) 726-4231
Golden Triangle Rural Family Health Center (Clayton Village Community)	1237 Highway 182 East	Starkville, MS 39759	(662) 320-7001
Sardis Family Medical Clinic Inc.	111 West Lee Street	Sardis, MS 38666	(662) 487-1064
Tri Lakes Pediatric Clinic	435 Highway 6 East	Batesville, MS 38606	(662) 712-2367
Tri Lakes Women's Clinic	303 Medical Center Drive	Batesville, MS 38606	(662) 712-2220
Pearl River Family Clinic	302 Hwy 11 South	Poplarville, MS 39470	(601) 403-8284
Picayune Health Services	711 Sixth Avenue	Picayune, MS 39466	(601) 798-5798
Poplarville Family Med Clinic, The	1407 South Main Street	Poplarville, MS 39470	(601) 795-0659
Doctors Clinic	210 Bay Avenue West	Richton, MS 39476	(601) 788-9222
Anazia Medical Clinic	120 5th Avenue	McComb, MS 39648	(601) 249-0013
Family Practice Clinic McComb	1506 Harrison Avenue	McComb, MS 39648	(601) 249-2142
Osyka Family Clinic	1081 Second Avenue	Osyka, MS 39657	(601) 542-3300
Pinnacle Medical Clinic	7900 MS Hwy 570 West	Summit, MS 39666	(601) 684-7771
Southwest Family Medicine	1510 Harrison Avenue	McComb, MS 39648	(601) 684-6891
Southwest Internal Medicine RHC	215 Marion Avenue	McComb, MS 39648	(601) 249-0706
Family Medical 101, Inc.	101 Mimosa Street	Booneville, MS	(662) 720-4919

		38829	
Lower Crossing Medical Clinic	670 Hwy 178, Suites 2 & 3	Sherman, MS 38869	(662) 844-7999
Sherman Family Clinic	608 Highway 178	Sherman, MS 38869	(662) 840-8978
Wheeler Family Medical Clinic	618 CR 5031	Wheeler, MS 38880	(662) 365-0200
Deporres Health Center	411 Poplar Street	Marks, MS 38646	(662) 326-9232
Florence Family Clinic	204 East Main Street	Florence, MS 39073	(601) 845-6602
Harrisville Medical Clinic	1865 Hwy 469	Florence, MS 39073	(601) 847-7784
Clark Clinic	36 Church Street	Morton, MS 39117	(601) 732-8612
Community Health Clinic	330 North Broad Street	Forest, MS 39074	(601) 463-4771
Forest Family Practice Clinic	#1 Medical Lane	Forest, MS 39074	(601) 469-4861
Morton Family Medical Clinic	317 Hwy 13 South	Morton, MS 39117	(601) 732-7114
Rush Family Practice	24489 Hwy 80	Lake, MS 39092	(601) 775-3264
Total Care Clinic	526 Deerfield Lane, Suite C	Forest, MS 39074	(601) 469-0291
Andrew George, M.D.	25 South 4th Street	Rolling Fork, MS 39159	(662) 873-0477
Jackson Clinic	102 South Fourth	Rolling Fork, MS 39159	(662) 873-4361
Magee After Hours Clinic	376A Simpson Hwy 49	Magee, MS 39111	(601) 849-5321
Stone County Family Medical Center	144 Eat Central Avenue	Wiggins, MS 39577	(601) 928-6700
Wiggins Clinic	303 South 1st Street	Wiggins, MS 39577	(601)928-4412
Wiggins Primary Care Clinic	200 Coastal Paper Drive	Wiggins, MS 38751	(601) 528-9119
Indianola Family Medical Group	122 East Baker Street	Indianola, MS 38751	(662) 887-2212
Indianola Medical Clinic	401 Catchings Avenue	Indianola, MS 38751	(662) 887-2494
Sunflower Rural Health Clinic	840 North Oak Avenue	Ruleville, MS 38771	(662) 756-2711
Charleston Clinic	401 Church Street Post Office Box 27	Charleston, MS 38921	(662) 647-5816
Glendora Clinic	Corner Gibson Avenue & Westbrook Street	Glendora, MS 38928	(662) 375-5578
Sumner Clinic	100 North Court Square	Sumner, MS 38957	(662) 375-9989
Tutwiler Clinic	205 Alma Street	Tutwiler, MS 38963	(662) 345-8334
Wolfe Family Medical Clinic	204 E. Walnut Street	Charleston, MS 38921	(662) 647-0900
Cotton Plant Family Clinic	100 CR 714	Blue Mountain, MS 38610	(662) 538-4111
Family Nurse Clinic	1305 City Avenue	Ripley, MS 38663	(662) 512-8590
Nurse Med. Inc.	716 South Main Street	Ripley, MS 38663	(662) 837-1534
Segars Clinic	1507 West Quitman	Iuka, MS 38852	(662) 423-1000
Preventive Care Health Service	2073 Old Highway 61	Tunica, MS 38676	(662) 357-7602
Family Clinic of New Albany	474 W. Bankhead Stree	New Albany, MS 38652	(662) 534-7777
Internal Medicine Rural Health Clinic of New Albany	300 Oxford Road	New Albany, MS 38652	(662) 534-8166
River Region Rural Health – Mob1	2100 Hwy 61 North	Vicksburg, MS 39183	(601) 883-5000
River Region Rural Health-FM	1907 Mission 66	Vicksburg, MS 39180	(601) 636-1173

Delta Regional Health Clinic	129 East Starling Street	Greenville, MS 38701	(662) 378-2020
Greenville Primary Care Clinic	2363 Hwy 1 South	Greenville, MS 38701	(662) 334-1253
Hollandale Primary Care	1257 Highway 61 South	Hollandale, MS 38748	(662) 827-2214
Leland Medical Clinic	201 Baker Boulevard	Leland, MS 38756	(662) 686-4121
Arthur E. Wood Medical Clinic	920 Matthew Drive	Waynesboro, MS 39367	(601) 735-7101
Waynesboro Family Medicine and Obstetrics	920 Matthew Drive, Suite A	Waynesboro, MS 39367	(601) 735-2401
FMCH Catching Clinic	451 Bank Street	Woodville, MS 39669	(601) 888-3421
FMCH Field Clinic	206 Main Street	Centreville, MS 39631	(601) 645-5361
Louisville Medical Associates, LTC	564 East Main Street	Louisville, MS 39339	(662) 773-7500
Odom Rural Health Clinic	604 South Main Street	Water Valley, MS 38965	(662) 473-1311
Yazoo Family Healthcare PLLC	307 East 15th Street	Yazoo City, MS 39194	(662) 746-2113

ISPRS Int. J. Geo-Inf. **2013**, *2*, 110-134; doi:10.3390/ijgi2010110

OPEN ACCESS
ISPRS International Journal of **Geo-Information**
ISSN 2220-9964
www.mdpi.com/journal/ijgi/

Article

Potential Impact of Climate Changes on the Inundation Risk Levels in a Dam Break Scenario

Sudha Yerramilli

National Center for Biodefense Communications, Jackson State University, 1230 Raymond Road, Jackson, MS 39204, USA; E-Mail: sudha.yerramilli@jsums.edu; Tel.: +1-601-519-5252

Received: 10 December 2012; in revised form: 6 February 2013 / Accepted: 15 February 2013 / Published: 4 March 2013

Abstract: The overall objective of the study is to generate information for an enhanced land use planning with respect to flood hazards. The study assesses the potential impact of climate change by simulating a dam break scenario in a high intensity rainfall event and evaluates the vulnerability risk in the downstream region by integrating ArcGIS and Hydrologic Engineering Centers River Analysis System (HEC-RAS) technologies. In the past century, the evidence of climate changes are observed in terms of increase in high intensity rainfall events. These events are of high concern, as increased inflow rates may increase the probability of a dam failure, leading to higher magnitude flooding events involving multiple consequences. The 100 year historical rainfall data for the central Mississippi region reveals an increased trend in the intensity of rainfall rates after the 1970s. With more than 10% of high hazard dams in the central region, the damage can be far accumulative. The study determines occurrence of the high intensity rainfall event in the past 100 years for central Mississippi and simulates a Ross Barnett Reservoir dam break scenario and evaluates the vulnerability risks due to inundation in the immediate downstream region, which happens to be the State Capital. The results indicate that the inundation due to a Ross Barnett Reservoir failure under high intensity rainfall event is comparable to a catastrophic flood event experienced by the region in 1979, which almost equals a 200-year flood magnitude. The results indicate that the extent and depth of flood waters poses a significant destructive threat to the state capital, inundating various infrastructural and transportation networks.

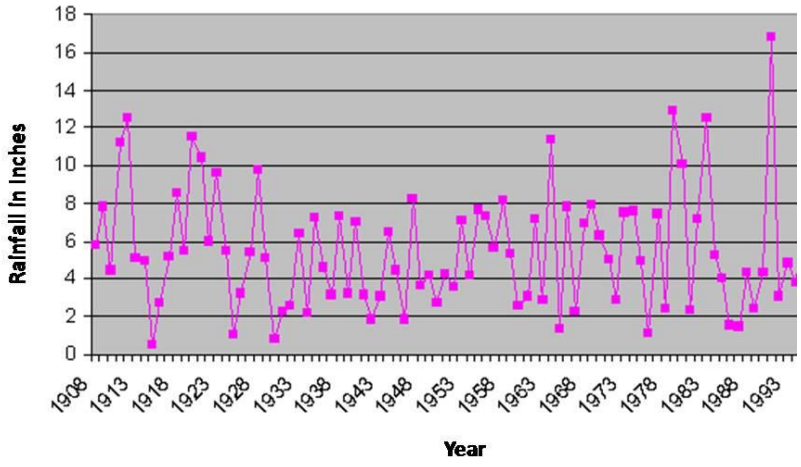
Keywords: GIS; flood simulation; dam break; climate change; HEC-RAS

1. Introduction

According to the Intergovernmental Panel on Climate Change (IPCC), over the 20th century, the observed temperature and precipitation changes in the United States was quite higher than the rest of the world [1]. The predicted temperature changes in central North America are higher than the global mean values [2], because of higher latitudes. The central USA witnessed significant changes in temperature and rainfall, and according to the United States Environmental Protection Agency (USEPA) report [3], over the last century, the precipitation levels in Mississippi have increased by about 20% over the mean annual rainfall and are predicted to rise by 5–25% in the coming century throughout the state. The evidence of the climate changes is observed not only in terms of increase in the average rainfall, but also in terms of increase in the intensity (The greatest depth of precipitation for a given duration that is physically possible over a given area at a particular geographical location at a certain time of year) of precipitation events, which is of high concern towards flooding problems [1].

The climatic changes have disastrous consequences that impact physical systems, infrastructure and social organization in many ways. Floods can also be categorized as high-impact events, as they involve multiple consequences, such as disruptions in the transportation and communication sectors, property damage and prolonged submergence of agricultural lands, wetlands, *etc*. The disaster effects are greatest for floods than any other calamity. Of all the natural calamities, floods can be described as catastrophic events whose impact lasts for a long period of time. The 100 year historical rainfall data for the central Mississippi region reveals an increased trend in the intensity of rainfall rates after 1970s (Figure 1). These events are of high concern, as increased inflow rates may increase the probability of a dam failure, leading to higher magnitude flooding events involving multiple consequences. In a report prepared by the city of Roseville, CA, climate change effects trigger the probability of dam failure, as dams are designed based on the assumptions about the river flow behavior or hydrographs. The changes in the weather patterns due to climate change effects can bring significant effects on hydrographs used for the design of the dam [4]. In case of a dam failure in these rainfall events, the damage can be far more severe and accumulative to the State of Mississippi, with 277 high hazard dams.

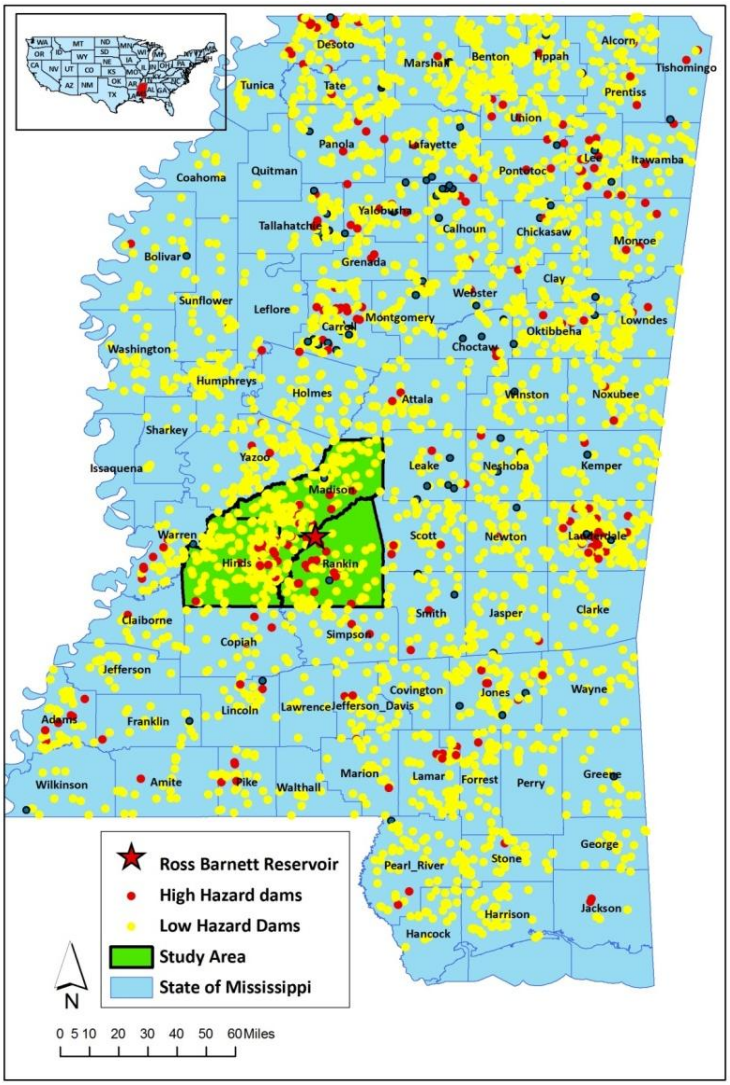
Figure 1. Intensity of the rainfall events after 1970 in the central Mississippi region.



1.1. Background Information

In the United States, flash floods are the leading cause of weather related mortality [5], and the gulf coast region is especially subject to extreme floods. The state of Mississippi ranks eighth [6] and stands among the worst flood-hit states in the nation, with a repetitive number of flash floods occurring from short and intense rainfall events. Extreme inflow rates resulting from these intense rainfall events increase the probability of an eventual dam failure and subsequent flash floods [6]. Many of the existing dams in the US do not have adequate capacity to handle extreme inflows from the intense rainfall events [7]. One such incident happened when a 70-year rainfall event (13.54 cm of rainfall within 24-h) triggered a catastrophic failure of the Ivex dam on the Chagrin River in northeastern Ohio in 1994 [8]. Living in the downstream regions of a high hazard dam in such intense rainfall events doubles the hazard potential, posing severe threat to the lives and damaging infrastructure and property. In budget dollars per dam, Mississippi ranks 47th, leaving thousands of Mississippians to live below high and significant-hazard dams (Figure 2), which, without timely maintenance and inspection, become potentially lethal time bombs [9]. For central Mississippi, with the state capital situated downstream of a high hazard dam (Ross Barnett Reservoir dam), the outbreak of this dam in an event of high intense rainfall can prove to be devastating.

Figure 2. Spatial location of high hazard dams in the State of Mississippi.



In its history, the City of Jackson had experienced one of the most catastrophic floods, the Pearl River flood of 1979, also known as the "Easter Flood of 1979". According to a US Army Corps of Engineers report [10] the 1979 flood flows surpassed the records of past flood events, causing more than $200 million damage. The extent of damage was severe, as serious disruptions occurred to transportation and communications that blocked the State capital for a number of weeks. The 1979 flood was estimated as a 200-year flood event that left the City of Jackson with devastating damages.

Rutherford [11] describes the damages in the City of Jackson resulting from a flood hazard as a public policy disaster, which had occurred due to continuous developmental activities in the floodplain after the 1961 flood. Failure to integrate the vulnerable locations into the land use planning resulted in devastating conditions in the state capital.

1.2. History of Ross Barnett Reservoir Dam

The Ross Barnett Reservoir was built in 1960, located in the midst of the three most densely populated counties and economic centers of Mississippi, is one of the high hazard dams that lacks an Emergency Action Plan (MDEQ, Division of Dam safety, 2000). Further, the design lifetime of the earth-fill dam posing more risk to downstream region is of serious concern [12]. Despite flood control improvements of levees and clearing, the Jackson, MS, metropolitan area below the Ross Barnett Reservoir Dam suffers annual flood damages from the Pearl River of about $10 M [10].

A review of the literature suggests that a number of studies have been done for simulating dam break floods. Katopodes [13] examines the flooding patterns when a finite element method (Galerkin formulation) model is applied on a discontinuous channel flow. By comparing the results with the analytical solutions, the study rated the performance of the model to be poor. In another study, Hromadka [14] analyzes a two-dimensional dam-break model developed for a flood plain, with flow equations solving a diffusion model coupled to the equation of continuity. The study concludes this approach can better predict at a two-dimensional dam-break flood plain over a broad, flat plain more accurately than a one-dimensional model. Similarly, Akanbi [15] presented a model to indicate the changes in the behavior of flood waters propagating on a dry bed, and Zhao *et al*. [16] investigated the effects of changes in bed elevations on the flooding patterns using three defined solvers and concluded that these models are useful for studying levee failure or dam break due to extreme flood events. The capabilities of numerical models that simulate flooding due to failure of a dam/reservoir in a natural river were presented by Sharma [17]; Zoppou and Robert [18]. Preliminary review of these articles indicates that these studies focused on analyzing the differences between the flood model equations that were solved and estimated the level of stability and accuracy of the equations. These studies lack the application of results on the communities, which can facilitate in identifying vulnerable locations.

On the other hand, a MEMA report [12] presented examples of two dam break studies that have been executed for Oak Lake Dam (Rankin County) and Acacia Woods Lake Dam (Rankin County). In both cases, the dam break simulations were done at the normal water level failure and have not considered future possible dam failure due to extreme inflows caused by intense rainfall events. Another study on the Ross Barnett Reservoir by Davies [19] was done from a hydraulic analysis dimension. In spite of numerous studies, there exists a knowledge gap in identifying vulnerable locations due to dam failure and applying the results to enhance activities in planning and developmental fields.

1.3. Purpose of the Research

The purpose of the research is to analyze the vulnerability risks posed by potential flood hazard due to a dam failure in an event of intense rainfall event. In this context, the current study, by integrating ArcGIS and HEC-RAS technologies, simulated a dam break analysis triggered by extreme inflow resulting from an intense rainfall event. The study conducted a vulnerability risk assessment for the downstream regions from a cultural, socio-economic and infrastructural impact perspective.

1.4. Description of the Simulation Scenario

The 100 year historical precipitation data (1908–2007) for the central Mississippi region is collected from the National Climatic Data Center (http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp). A rainfall event with 16.75 inches/24 h (April 1991) was determined as the high intensity event in the past 100 years. The actual inflow data into the Ross Barnett Reservoir dam for this rainfall event is collected from the 02486000 USGS (United States Geological Survey) station, which happens to be 16,026 cubic feet per second. As the rainfall event has been experienced by the region in the past, a dam break simulation under this event provides the inundation risks on a typical scale (base flooding scenario) that have a high possibility of occurrence on any given day in a year [20].

1.5. Suitability of HEC-RAS for Dam Break Scenario Simulations

As the current study focuses on the spatial identification and vulnerability index development for the flood hazard, simulation of the flood scenarios are required to be conducted in a geospatial environment. The models should facilitate the capability of applying a geographic information system (GIS) to the flood simulations and assist in analyzing the flood levels or extents spatially. GIS helps in visualizing flood simulations in an interactive setting, where the spatial impact of various scenarios can be viewed along with the location of critical facilities and, thus, helps in assessing the region's vulnerability towards a flood event efficiently [21]. In this connection, Bajwa and Tim [22] describes the 1-dimensional HEC-Geo models (HEC-GeoRAS developed by the US Army Corps of Engineers) as a geospatial hydrology toolkit that recognizes the power of the Arc-GIS environment in generating and visualizing flood simulations. The GeoRAS product developed by the US Army Corps of Engineers enables these flood simulation models to be compatible with the ArcGIS environment and provide valuable tools to evaluate impacts associated with flood plains [23]. Hicks and Peacock [24] strengthen the suitability and capability of HEC-RAS models in simulating floods by stating that the HEC-RAS flood simulations, examined through an application, shows accuracy comparable to more sophisticated hydraulic models.

Numerous studies have used HEC-RAS tools to evaluate dam break failures and subsequent flood hazards in a geospatial environment. Presently, many of the State and Federal agencies are using HEC-RAS for simulating dam failure scenarios to visualize the results in the GIS environment. The Oregon Department of Fish and Wildlife assessed Canyon Creek Meadows Dam failure affects using HEC-RAS 4.0. The Oregon State department has developed a scenario where a dam failure occurs due to high intensity rainfall events determined from 100 year historical data. A hydraulic model was developed for Canyon Creek to simulate a dam breach of the Canyon Meadows Dam. The geometric

profile of the region, needed for the HEC-RAS dam breach model, was constructed from USGS 30-m digital elevation models (DEMs), inline structures and land use. The external boundary conditions (the slope) were measured from the DEM and the model was simulated by setting the dam breach parameters. The Oregon State Department declared that the HEC-RAS dam break simulation run for a high intensity rainfall event was comparable to the dam breach analysis results conducted as part of the Emergency Action Plan (EAP) for the Canyon Creek Meadows Dam. The success of HEC-RAS in simulating dam break scenarios can be established, as the results provided similar results to that of the EAP dam breach analysis.

USGS through the Water Information coordination program has conducted numerous dam break studies. In one of its publications, the agency states that HEC-RAS solves Saint-Venant equations, which is well suited for computing the flood wave propagation resulting from a dam failure scenario [25]. The agency asserts that the integration of the HEC-RAS tool with ArcGIS allows flood plain managers and emergency managers to visualize the resulting hazards and assists in enhancing the protection and mitigation activities [23]. On similar lines, the Quebrada Beatriz Reservoir Dam break scenario was simulated under three high intensity rainfall events using the HEC-RAS tool to estimate the maximum flood waters in the study area [26].

1.6. Limitations of Study

Visualizing the HEC-RAS 1D model results in a 2D environment and involves uncertainties, as the results are complemented with information from various sources, such as topography (DEMs), infrastructure, land use and modeling input/outputs. The study is bound to the limitations of HEC-RAS and carries the uncertainties that come along with the model.

2. Methodology

The study integrates geospatial technologies with the HEC-RAS model to simulate the dam break flood inundation for the determined rainfall scenario. HEC-RAS 4.0, a flood simulation model developed by the US Army Corps of Engineers, computes steady flow and unsteady flow (dam break) simulations. The preprocessing of the geometric data (extraction of the physical characteristics of the study region) and the post-processing of the outputs (to visualize the flooding impact) that are required by the HEC-RAS dam break model are done by using HEC-GeoRAS (Figure 3). HEC-GeoRAS, an extension in ArcGIS, facilitates integration and visualization of HEC-RAS inputs and outputs.

2.1. Simulation of Ross Barnett Reservoir Dam Failure Scenario

The DEM data for the study area is collected from Mississippi Automated Resource Information System (MARIS) and is converted to Triangular Irregular Networks (TIN) format using ArcGIS functionalities (Figure 4).

From the 100 year historical rainfall data for central Mississippi, the high intensity rainfall event is determined. The actual inflow data for this event at Ross Barnett reservoir is obtained from the USGS station. Using HEC-GeoRAS, the geometric data for the study region (TIN, land use, physical properties of the dam, levees, bridges, cross sectional lines, *etc.*) is prepared in ArcGIS for the study region

(Figure 5). The preprocessed geospatial RAS layers are then exported to the HEC-RAS model. With the geometric data input obtained from HEC-GeoRAS, processing of a dam break-induced flood simulation is performed using the HEC-RAS unsteady flow method. To check for any errors before entering the dam breach information, the study performed a steady flow simulation.

Upon successful simulation of the steady flow run, the dam breach parameters are fed to the model. The dam breach parameters are set by following the information provided by the USACE HEC-RAS user manual in estimating the breach parameters for an earthen dam. The manual suggests that the breach width should be 1/2 to 3-times the height of the dam, side slope of the breach to be 0:1 to 1:1 and failure time as 0.5 to 4 h. For the Ross Barnett Reservoir failure, the parameters are set with breach width as 192 ft (height of the dam is 64 ft), side slope of 2:2 and failure time set to 4 h. The failure mode type is set as piping, with an initial piping elevation set to 150 ft. An initial condition of 4,000 cfs, and the downstream boundary condition as normal depth (the slope of the river computed from upstream to downstream −0.001) are determined. A simulation time of 7 days was given so as to capture the dam break flood till it reaches the end of Hinds County. Upon successful implementation of the simulation, the HEC-RAS output is exported to HEC-GeoRAS for post-processing of the output in ArcGIS. In the ArcGIS environment, using the HEC-GeoRAS extension, the imported results are processed with TIN (topographic data) to generate the flood water surface extents and the flood water depth files for the simulated scenario.

Figure 3. Processing of data flow.

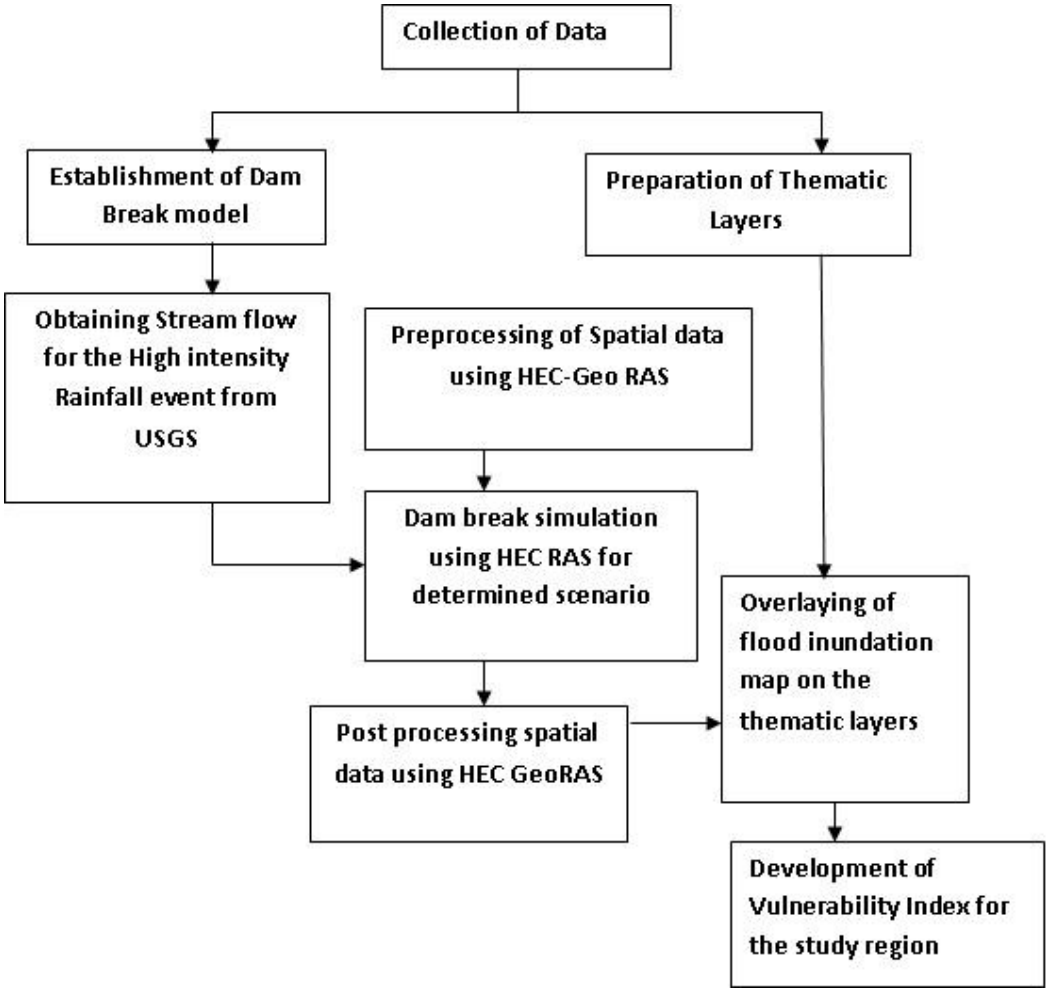


Figure 4. Triangular Irregular Networks (TIN) of the study region.

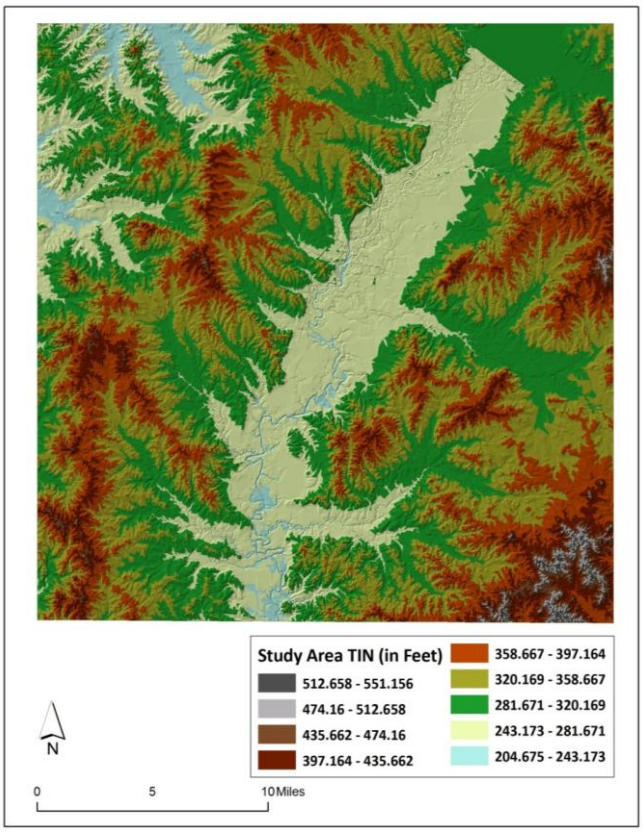
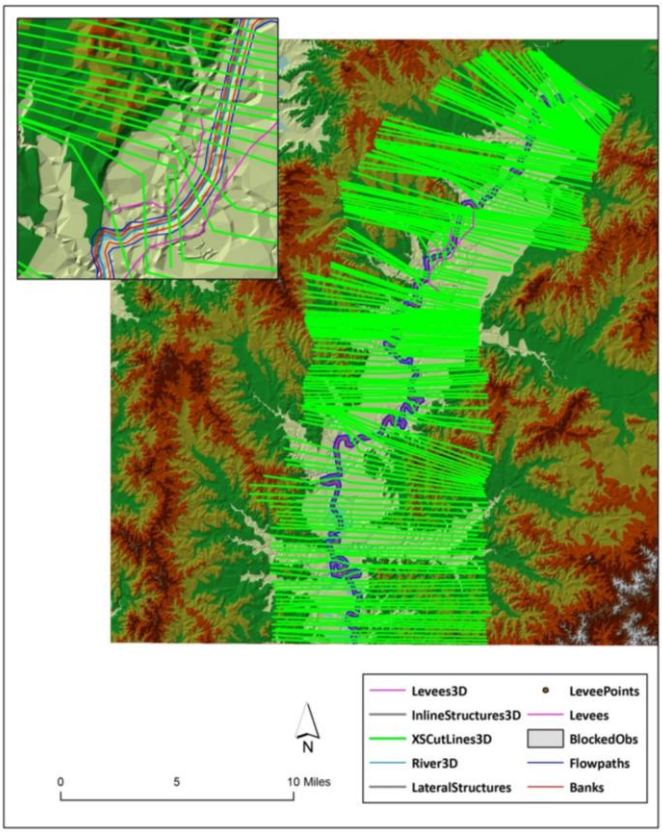


Figure 5. Preprocessing of geospatial RAS layers.



2.2. Inundation Risk Level Assessment

Using ArcGIS functionalities, the thematic layers for the study region are mapped for key facilities, historic districts, road network, population and housing units (Figures 6–10). The data is collected from various sources, such as Census Bureau, Mississippi Automated Resource Information System (MARIS) and City of Jackson. The maps of density of population and housing units are generated by dividing the total population/housing units in each block by the area of that corresponding block. The spatial location of key facilities and transportation routes are downloaded from MARIS and clipped to the study region using data management tools in ArcGIS.

Overlaying of inundation data with the thematic layers of the study region in a geospatial environment not only provided a visual representation of the spatial extent of flood waters, but also facilitated the quantitative assessment of the vulnerability risk of the region towards facing the hazard.

Assessment of the inundation risk hazard on the downstream region due to Ross Barnett reservoir dam failure in an high intensity rainfall event is discussed under (1) hazard identification and (2) vulnerability assessment sections. While the hazard identification involves estimating the spatial extent and depth of the flood waters, vulnerability risk assessment analyzes the impact of inundation on the cultural, socio-economic, transformational and infrastructural networks in the downstream region

Figure 6. Population at block level in Hinds, Rankin and Madison County, MS.

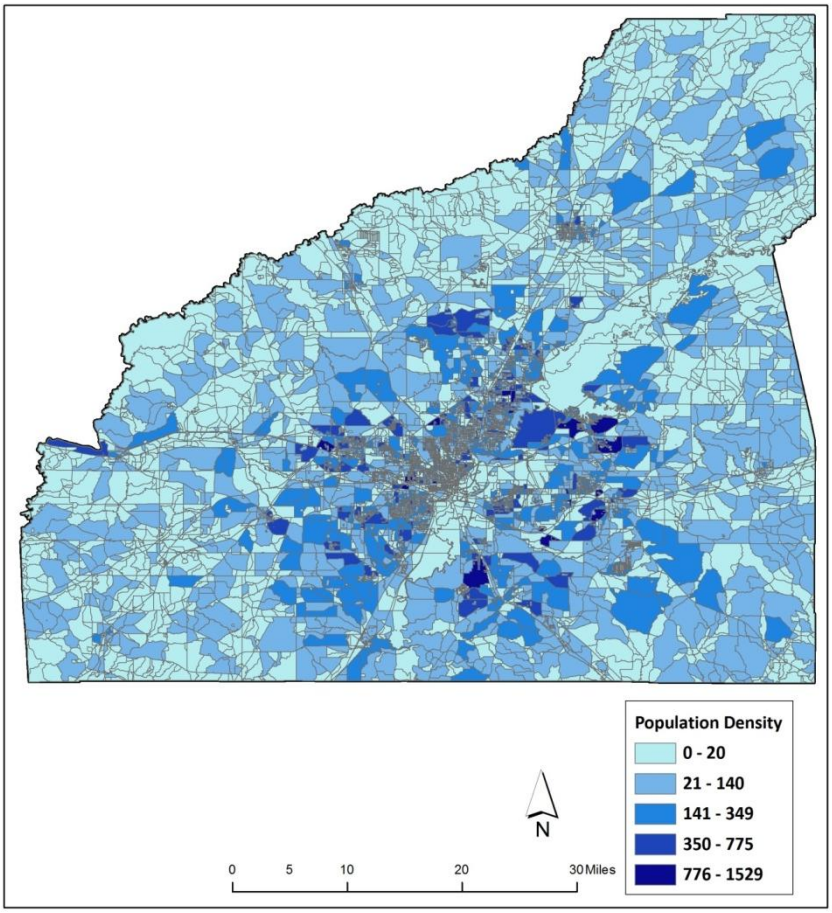


Figure 7. Housing Units at block level in Hinds, Rankin and Madison County, MS.

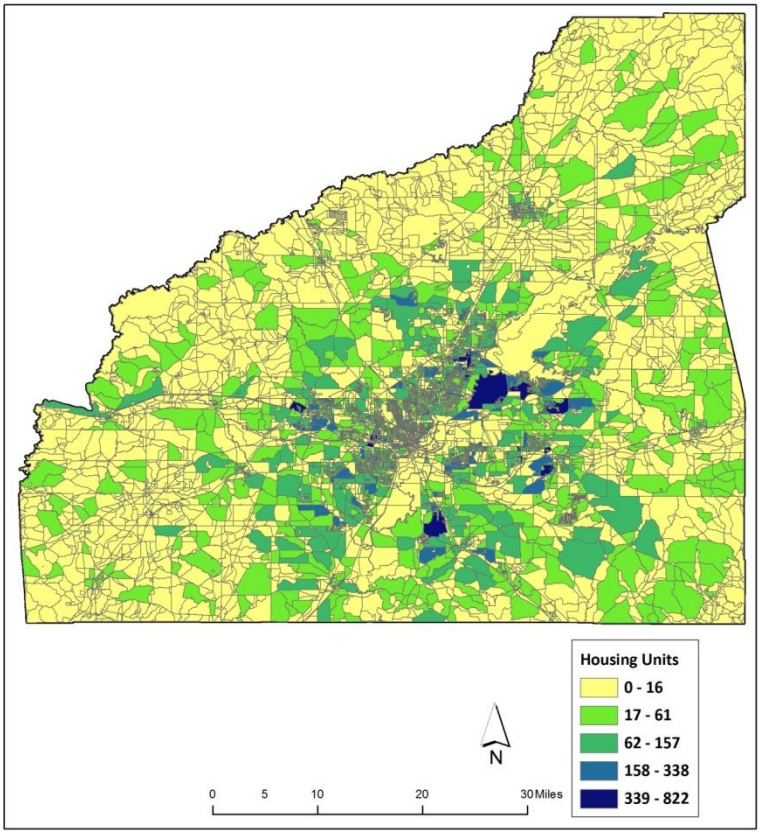


Figure 8. Major transportation routes through Hinds, Rankin and Madison County, MS.

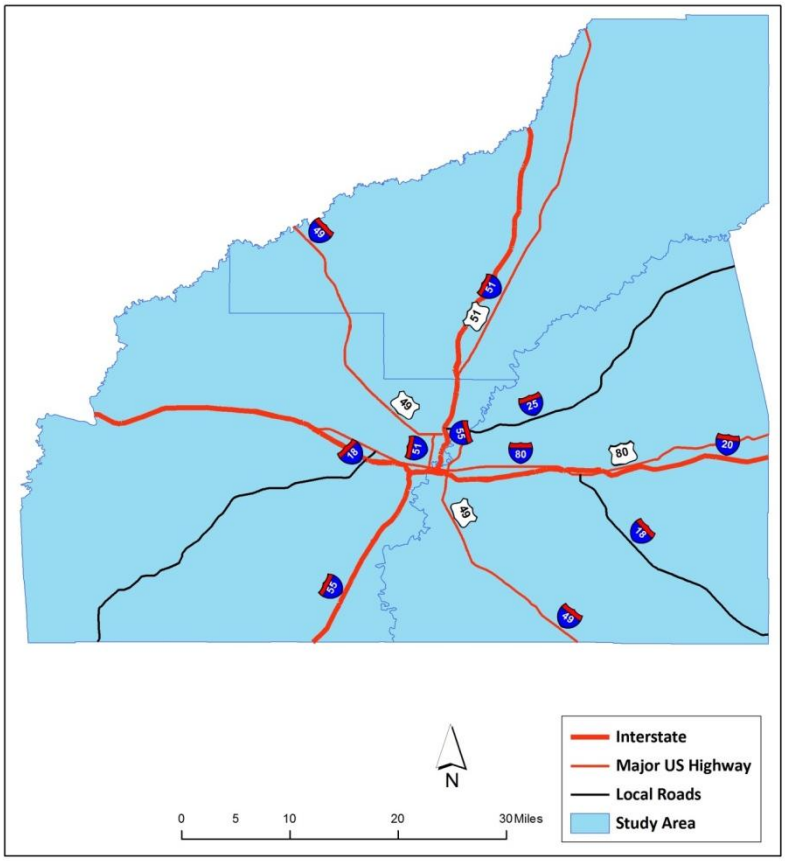


Figure 9. Key facilities in Hinds, Rankin and Madison County, MS.

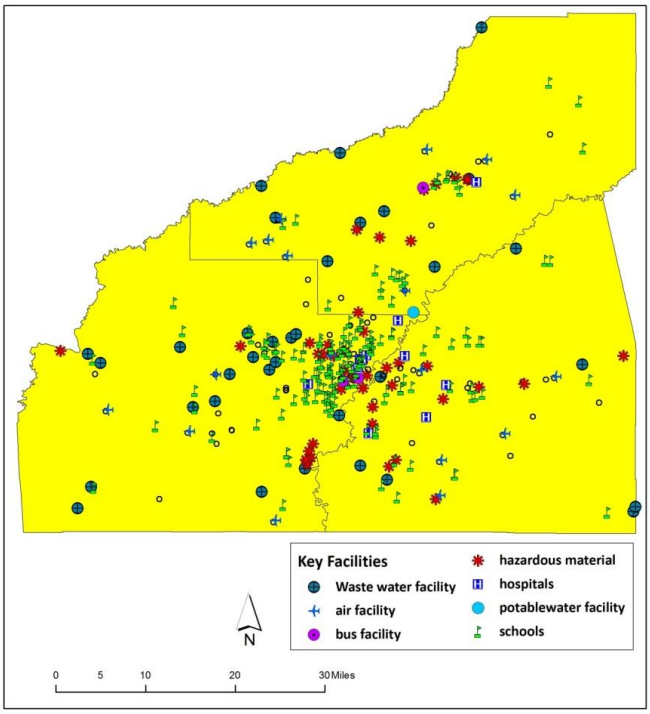
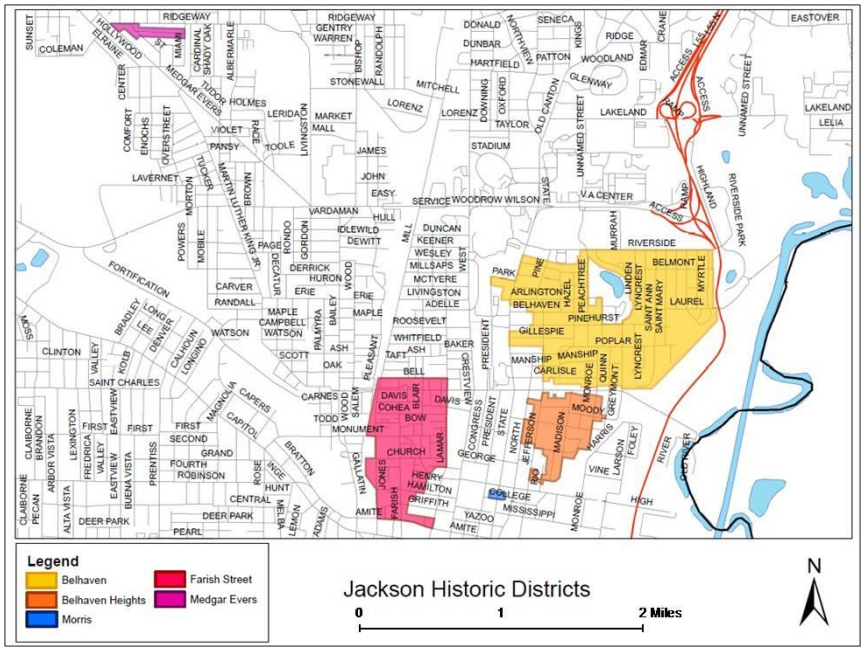


Figure 10. Historic districts in the City of Jackson (Source: www.jacksonms.gov [27]).



3. Results and Discussion

Integration of the scientifically generated information in the planning process helps in exposing the long-term threats posed by flood hazards. It facilitates the present planning activities to be designed in a futuristic manner and assists towards retaining the social, economical and environmental functionalities viable in the face of hazard. In order to achieve safe and sustainable communities, urban planners should view and understand the physical characteristics of hazard as an important indicator for identifying

vulnerable areas [28]. The interaction of the hazard with an urban area can have a potential impact on its cultural, historical, social, infrastructural and economical functionalities. Analyzing the vulnerability of these functionalities in the face of a flood event will not only present an insight about the disturbances that might occur in these inter-related functionalities, but also expose the risk factor upon which the current developmental activities are being planned. The analysis from this vulnerability assessment, as said by Geertman [29], can assist in advancing the inclusive nature of climate change factors in the planning process, thereby strengthening the focus of the sustainable approach in the future developmental activities downstream of the Ross Barnett Reservoir Dam.

As the downstream region (Hinds County) is taking part in the FEMA National Flood Insurance Program (NFIP), all the developmental activities are bound to a 100-year base flood magnitude. With the presence of a high hazard dam in the upstream and with climate change effects in place, planning developmental activities considering a 100-year flood hazard level may not hold good. In this context, the spatial locations of the blocks with potential flood threat are assessed by calculating a vulnerability index from an integrated risk factor obtained from the spatial extent of the identified hazard and the vulnerability risk of the region.

The study discusses the dam break simulation results under two sections: (1) hazard identification and (2) vulnerability assessment. While the hazard identification section discusses the spatial extent and depth of the flood waters, vulnerability assessment is described under the cultural-historical and social-infrastructural (population, housing units, transportation, key facilities) impacts due to inundation. Finally, the study calculates the integrated risk factor and presents the spatial location of high/medium/low vulnerable blocks due to a Ross Barnett Reservoir Dam failure under a high intensity rainfall event in Hinds County.

3.1. Hazard Identification

Hazard identification involves defining the spatial extent and magnitude of a hazard that may be possible for a specific geographical area. For the simulated scenario, the hazard is identified in terms of area of acres inundated and the flood water depth in the downstream region.

3.1.1. Spatial Extent

The spatial extent of inundation occurring due to Ross Barnett Reservoir Dam failure continues till the end of Hinds County (Figure 11). Using ArcGIS functionalities, it is estimated that an area of 50,331 acres of land is under inundation. Interestingly, the amount of area inundated by the simulation almost equals the 50,955 acres of spatial extent in the flood waters of the 1979 flood [30]. Correspondingly, through a visual comparison made by overlaying the spatial flood extents of the dam break simulation and 1979 flood waters (obtained from National Oceanic and Atmospheric Administration (NOAA)), the inundation pattern in both the scenarios almost matches and appears to follow the same path (Figure 12). As the 1979 flood is considered as a catastrophic 200-year flood, it can be inferred that the Ross Barnett Reservoir Dam failure under a high intensity rainfall event can cause a flood equivalent to a magnitude of a 200-year flood level.

Figure 11. Spatial extent of inundation downstream of Ross Barnett Reservoir.

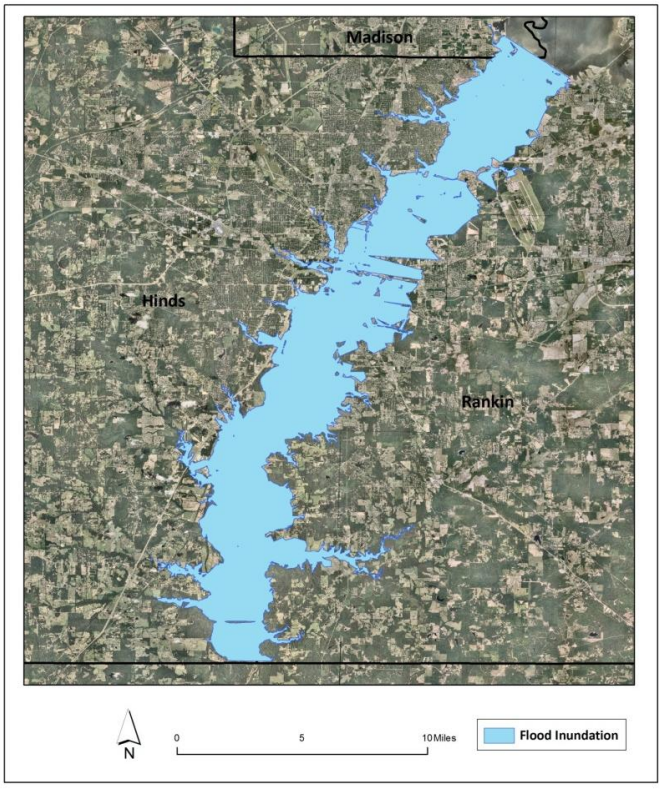
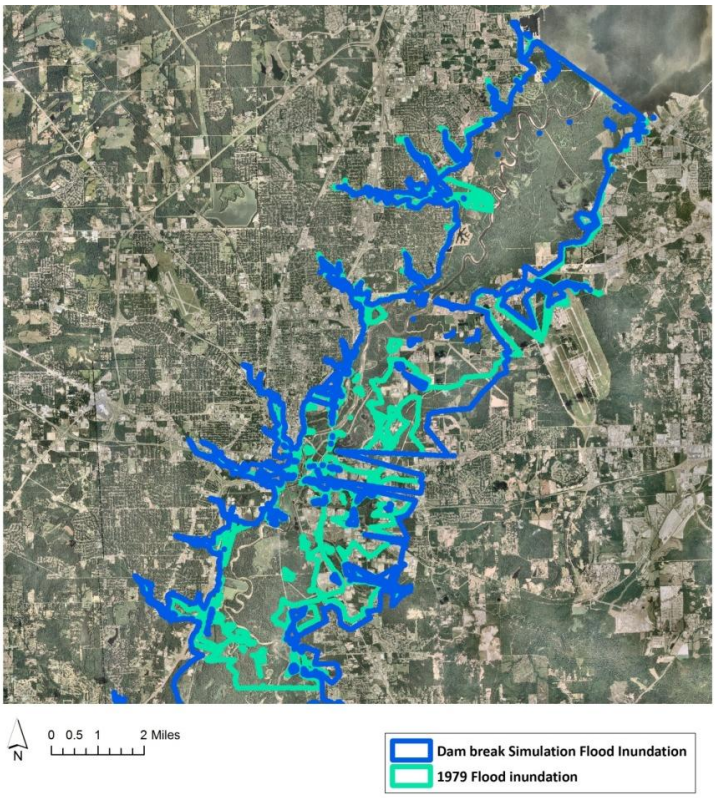


Figure 12. Comparison of inundation patterns of dam break simulation and 1979 flood inundation.



The possibility of a recurrence of the flood threat under the high intensity rainfall event is quite possible. The results reveal alarming information, as the current developmental activities in the

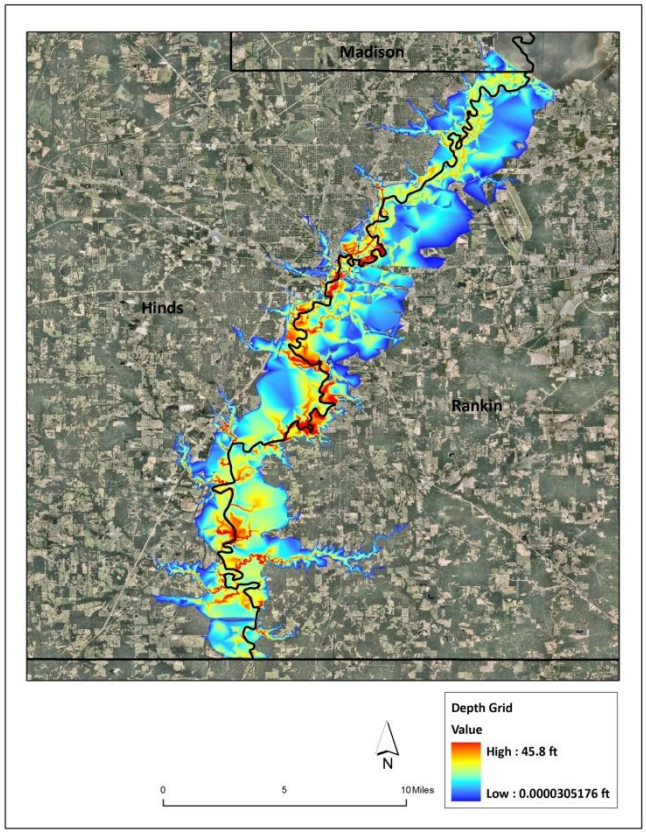
downstream counties are bound to a 100-year flood magnitude, while situated under a quite possible 200-year flood threat.

3.1.2. Depth Grids

The depth of the flood waters due to the Ross Barnett Reservoir Dam failure are presented in Figure 13, with range of water depths illustrated (Red–Blue: High–Low).

The maximum depths of the flood waters go up to 45.8 feet for the simulated scenario. The flood water depth at the same location in a 1979 flood situation was measured as 42.5 ft. The flood water depth obtained from dam failure is observed to be more than the 1979 flood scenario. The water depth values of Ross Barnett Reservoir Dam failure in a high intensity rainfall event can be compared to a 1979 flood scenario, leading to almost similar flooding depths. The comparison is done on the assumption that the difference in the depth values are caused due to the calculation of the depth grids from the corresponding DEMs used in their respective simulations.

Figure 13. Depth grid of inundation downstream of Ross Barnett Reservoir.



3.2. Vulnerability Assessment

Assessing the vulnerability helps in identifying the risk factor and exposing the disturbance that might affect the economical, social/infrastructural and environmental activities of a community. According to a UNEP report (2001), information from vulnerability assessment helps in building flood-resistant communities by reducing or enhancing the coping capacity of the community with the flood hazard by sensibly planning developmental activities. Using ArcGIS functionalities, the study

discusses the vulnerability of the downstream region of Ross Barnett Reservoir in the face of its failure (under determined rainfall events) under two factors: (1) impact on cultural and historical values and (2) Impact on social and Infrastructural facilities.

3.2.1. Impact on Cultural and Historical Values

The Mississippi state capital is considered as a cultural hub, with around 700 events taking place every year in its vicinity. The majority of these events are hosted in the downtown region of the City of Jackson at the Fairgrounds and Coliseum. People all across the state and nation are culturally connected to the South's biggest fair, the Mississippi State Fair, rodeos, trade mart, live stock shows and many more events. Apart from these events, numerous museums that portray the historical and cultural richness of the region (Old Capitol Museum, Oaks House Museum, Municipal Art Gallery, *etc*.) in downtown Jackson and the Natural Science Museum and the Agricultural Museum on Lake Land Drive are situated in the flood risk zone in the downstream region of the Ross Barnett Reservoir Dam. Visualizing results from the simulation indicate that the cultural functioning of the region gets impacted either directly by flood waters or indirectly through disrupted transportation networks leading to the facility (Figures 14 and 15).

Figure 14. Ross Barnett Reservoir Flood waters at the intersection of Lake Land drive and I-55.

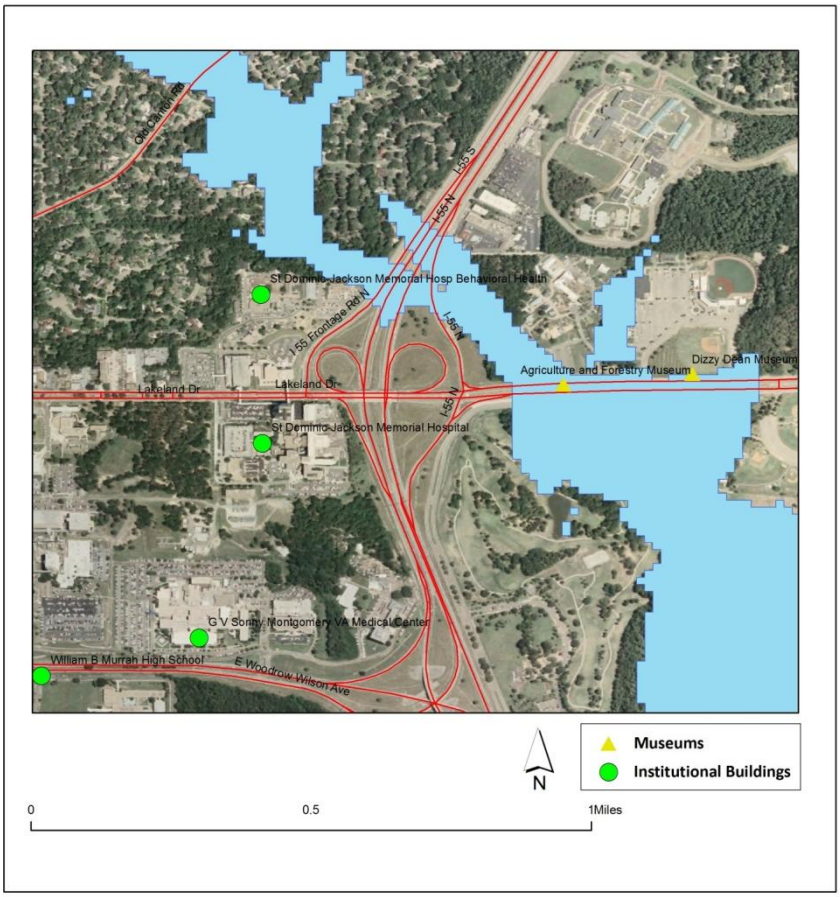
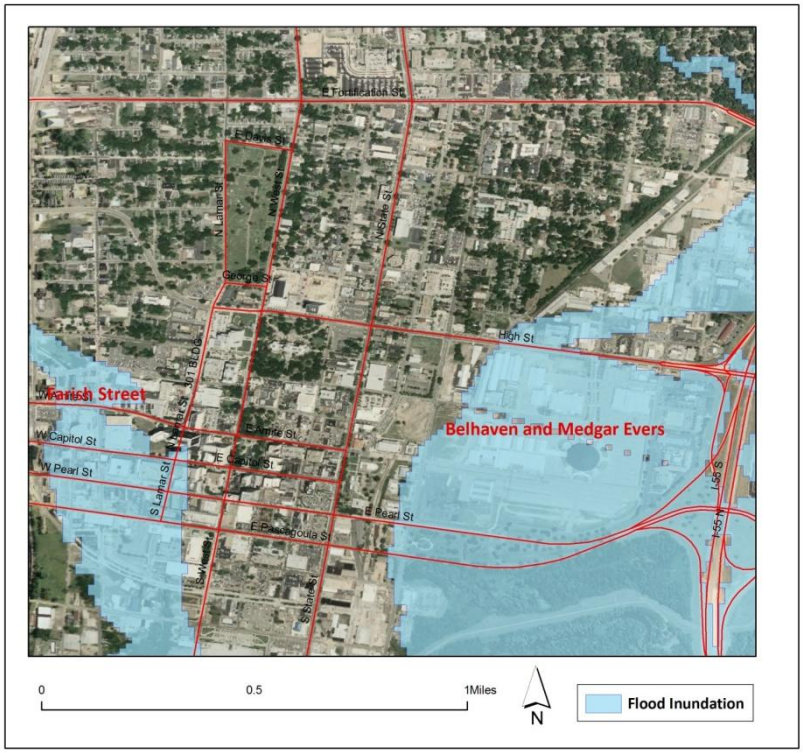


Figure 15. Inundation of the historic districts in downtown Jackson.



Inundation of the Mississippi Coliseum and Fairgrounds (from I-55 till Jefferson Street under the flood waters) can be seen from Figure 15, above. Hindrances to these events can impact the economic capability of businesses, such as hotels, restaurants and cab services, which depend on cultural events for their survival.

The effort of the department of planning and development to enhance tourism, preserve historic buildings and stabilize neighborhoods has resulted in identifying five historic districts in the City of Jackson, the majority of them situated in the downtown region (Figure 15). Millions of dollars are funded to these projects by local, state and national trust funds in order to support them to retain and stabilize. From the above figures, it can be evident that the Ross Barnett Reservoir Dam failure in a high intensity rainfall event puts the Belhaven, Medgar Evers and Farish Street historic districts at flood risk. The results identify the possible threat under which these areas are situated and, thereby, facilitate the planning authorities to integrate such information, so as to steer the programs to withstand such threats.

3.2.2. Impact on Social and Infrastructural Facilities

Population and Housing Units: Analyzing population and housing units is a vital factor in assessing the vulnerability of a region towards a flood event. Zheng [31] describes social vulnerability, determining the presence of population and housing units, as the key factor in assessing the resistance of any community towards a flood threat. Census block level data was used to estimate the potential number of housing units and population at risk (Figures 16 and 17).

Figure 16. Population at flood risk downstream of Ross Barnett Reservoir.

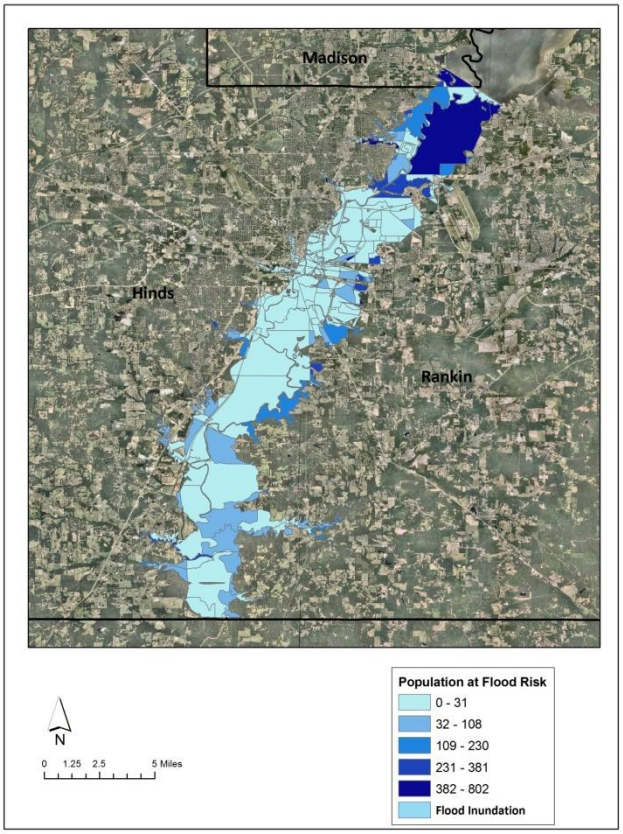


Figure 17. Number of housing units at flood risk downstream of Ross Barnett Reservoir.

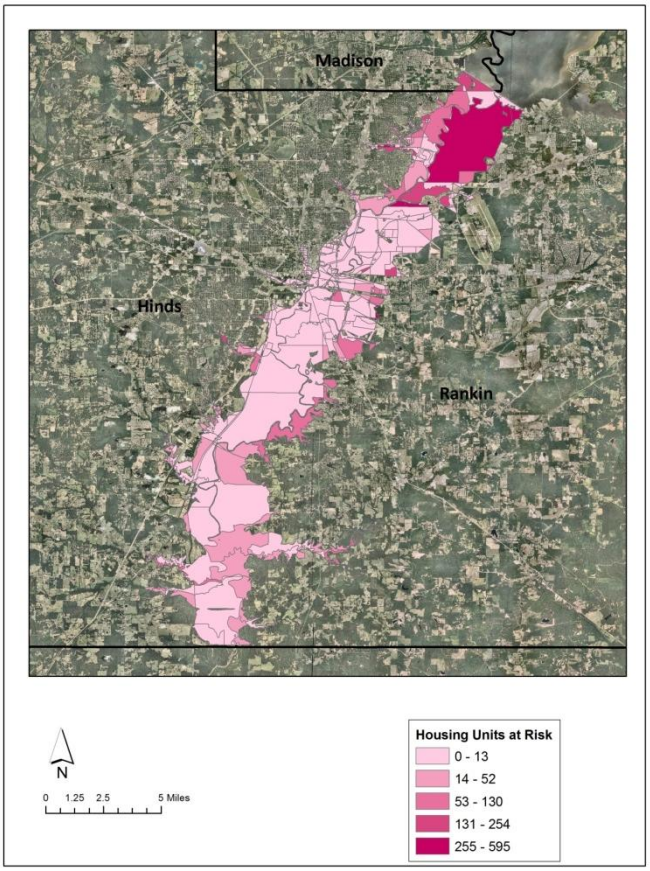


Table 1 shows the statistical numbers of housing units and population likely to be affected by inundation. The impact of Ross Barnett Reservoir Dam failure is majorly borne by Hinds County, as the county bears more than 60% of the population and housing units at flood risk. With the state capital, the City of Jackson, being in Hinds County, the impact of this flood hazard can be cumulative.

Table 1. Number of population at risk in the downstream and Hinds County region.

Total Population Affected Downstream	Hinds County Population Affected	Total Inundated in Downstream (Housing Units)	Total Inundated in Hinds County
34,100	21,660	16,279	10,384

Transportation: Disruption in the transportation system by flooding can paralyze the region's social and economic functionality from a local to regional level. Figures 18 and 19 depict the spatial location of the length of inundation on the major highways and interstates running through the downstream region of Ross Barnett Reservoir. Highway 49 and I-55 are the most affected transportation corridors.

Figure 18. Spatial location of possible disruptions in the major transportation corridors.

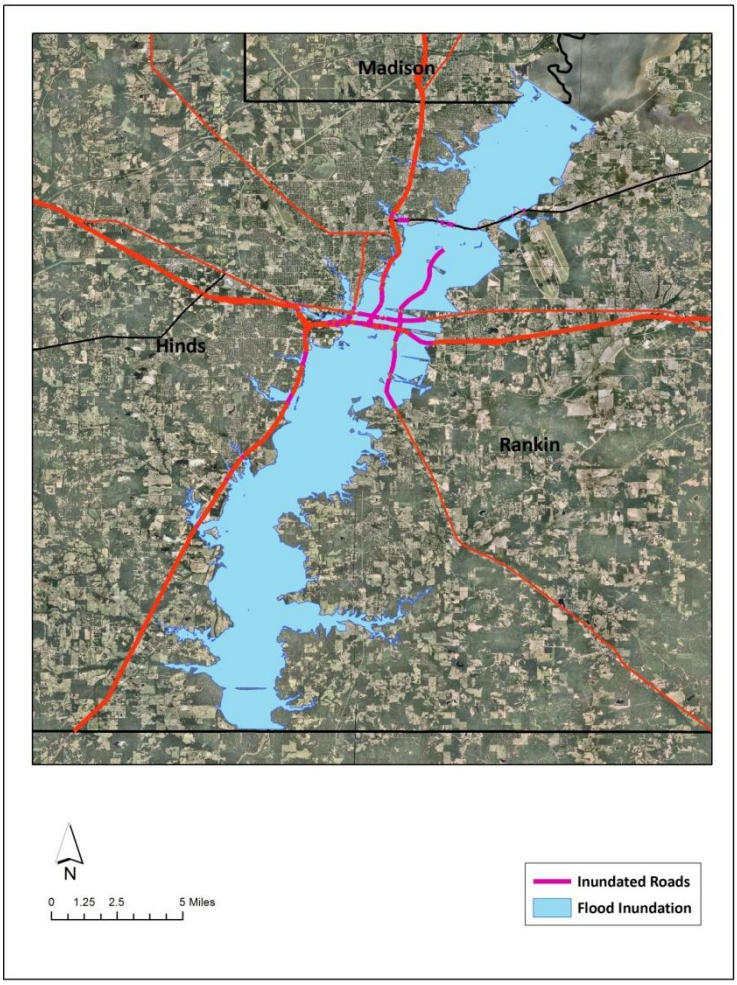
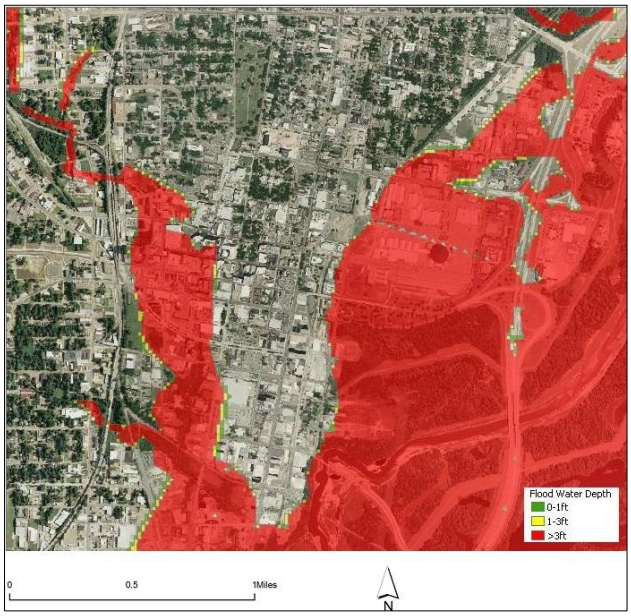


Figure 19. Depth of the flood waters on Lake Land Drive.

With the depth grid information obtained from HEC-RAS results, the study integrated the information with the road network and identified locations under various ranges of flood water by categorizing the water depths into 0–1 ft, 1–3 ft and >3 ft levels. Figure 16 indicates that these flood water depths on I-55 and Hwy 25 impact accessibility to the major healthcare facilities of the region, namely St. Dominic's Jackson Memorial Hospital, Mississippi Medical Center, and cultural facilities, such as the Agricultural Museum and the Natural Science Museum, that are located along these major transportation corridors of the region.

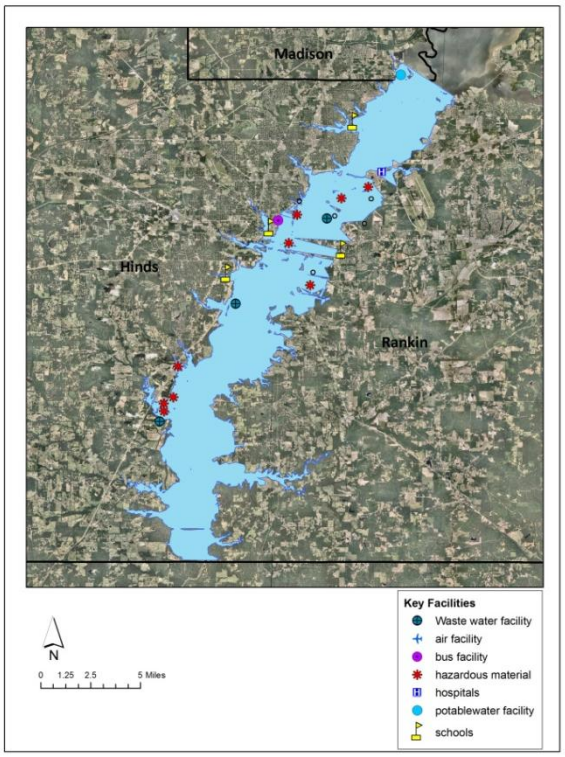
The water depth grids generated by HEC-RAS on I-55 (downtown region) indicate that an inundation of more than 3-ft-deep flood waters can impact the economic functionality of the region, as most of the commercial and administrative functionalities are accomplished from this region (Figure 20).

Figure 20. Inundation on I-55 near downtown region.



Key facilities: Inundation of key facilities impacts the operational efficiency of any region. Simulation results specify that around 33 key facilities get directly affected by the Ross Barnett Reservoir Dam failure under high intensity rainfall event (Figure 21).

Figure 21. Spatial locations of affected key facilities downstream of Ross Barnett Reservoir.



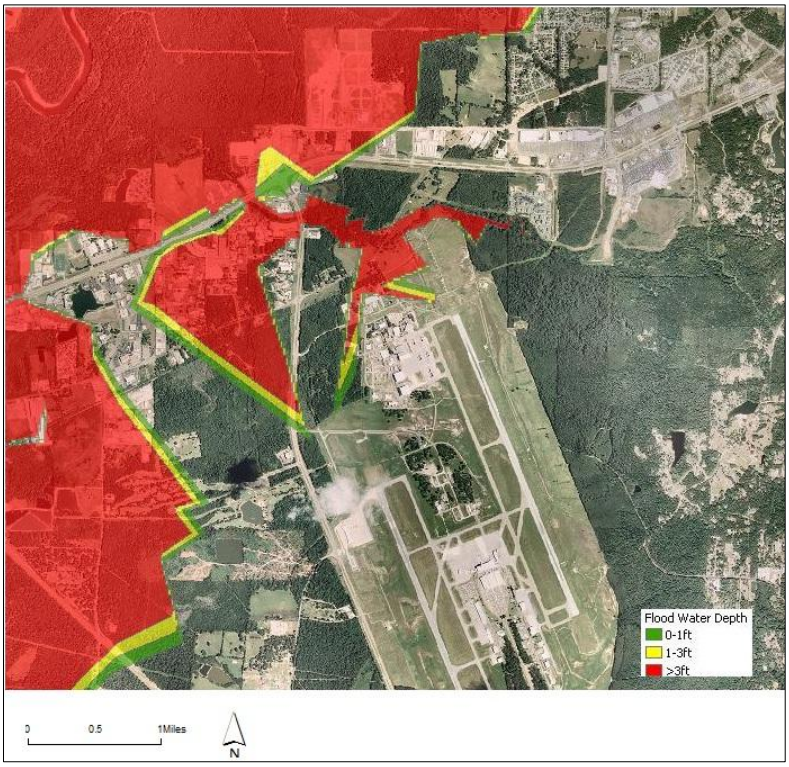
The only potable water facility to the downstream region of Ross Barnett Reservoir Dam will be under flood waters, putting the Hinds County population at risk, with long reaching health impacts (Table 2). The presence of 12 hazardous material plants in the inundation zone further aggravates the risk, as the outbreak/chemical accidents at these plants can pollute the natural resources (such as drinking water, wetlands, *etc.*), leaving the population and livestock in contaminated environments. The two major transit facilities of the state capital—City of Jackson-Transit Services (on S. President Street) and Jackson Amtrak station on W. Capitol Street—seem to be severely impacted in the simulated event, and shutting down of these may result in paralyzing the economic functionalities of the state capital.

Table 2. Affected key facilities.

Key Facilities	Simulation Event
Hospitals	1
Schools	8
Wastewater Facility	4
Potable Water Facility	1
Bus Facility	1
Communication Facilities	4
Hazardous Material Plants	12
Railway Facilities	2
Total	33

Another major impact the region could experience due to Ross Barnett Reservoir Dam failure is the impact on the only airport serving the state capital region (Jackson International Airport) being surrounded by the flood waters (Figure 22). The global commerce of the state capital with the rest of the world can come to a standstill, which may even have wide consequences in terms of economic expansion of the region.

Figure 22. Spatial representation of flood waters near Jackson International Airport.



3.2.3. Identification of Highly Vulnerable Blocks in Hinds County

While the discussed vulnerability indicators expose the risks on an individual basis, estimation of a composite index provides an overall risk associated with the region. As described by Timothy [32], this study considered two types of indicators (hazard and social vulnerability indicators) for the estimation of vulnerability index. Under the hazard indicator, the flood hazard map is scaled to a block level, and the maximum flood depth at each block is estimated. These flood depths were then standardized by dividing the flood depth of each block by the maximum flood depth value to create an integrated hazard index that ranges from 0–1.

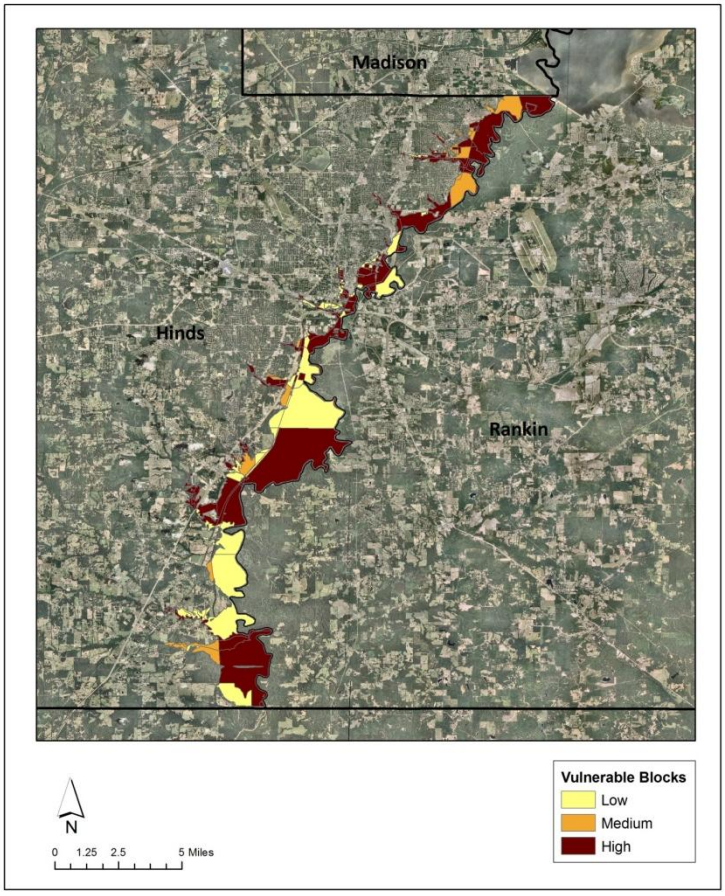
The presence of at-risk population, housing units and the level of accessibility to critical facilities in a hazard region determines the social vulnerability [32]. In this regard, the study considers four elements to represent the social vulnerability indicators. Density of population and the presence of housing units per each block in the incident area are calculated. The more the density of people and housing units, the greater the risk. The density values are then standardized by dividing each corresponding block value by the maximum density value, so as to create a population and housing units density index that ranges from 0–1. The other two elements that determine the social vulnerability are the accessibility to critical facilities and the functionality of roads that lead to them. Using ArcGIS analysis tools (extract and

overlay tools), the shape files of flood waters, location of critical facilities and the transportation network are overlaid, and the critical facilities coming under flood waters and the length of transportation routes affected (in miles) are estimated. Standardization across the blocks are done by dividing each block value with the maximum value to obtain an integrated index that ranges from 0–1.

Once the index values of all the four elements are computed, they are again summed to re-index on a new scale of 0–1 that is assigned to each block. The final vulnerability index across block level is visualized in ArcGIS.

For each block in the study region, the composite index is calculated by taking into consideration the level of flood depth, percentage of inundated roads, number of inundated population and housing units and number of key facilities impacted. Figure 23 illustrates the spatial location of the blocks under various vulnerable zones.

Figure 23. Levels of vulnerability at a census block level in Hinds County.



4. Conclusions

The objective of the study is to assess the potential impact of climate changes on the inundation risk levels in a Ross Barnett Reservoir Dam failure scenario. Most of the previous studies that were done on the impacts of dam failure mainly focused on the hydraulic analysis dimension, analyzing the design failure causes of the dam. Thus, filling the gap, the present study simulates the impact of Ross Barnett Reservoir Dam failure under a high intensity rainfall event on the downstream region and conducted a flood vulnerability index of the blocks.

The 100 year historical rainfall data for the central Mississippi region reveals an increased trend in the intensity of rainfall rates after the 1970s. These events are of high concern, as increased inflow rates may increase the probability of dam failure, leading to higher magnitude flooding events involving multiple consequences. A rainfall event with 16.75 inches/24 h (April 1991) was determined as the high intensity event in the past 100 years. The actual inflow data into the Ross Barnett Reservoir Dam for this rainfall event is collected from the 02486000 USGS (United States Geological Survey) station, which happens to be 16,026 cubic feet per second. As the rainfall event has been experienced by the region in the past, a dam break simulation under this event provides the inundation risks on a typical scale.

The study integrates geospatial technologies with the HEC-RAS model to simulate the dam break flood inundation for the determined rainfall scenario. The integration of geospatial technologies (ArcGIS) with the HEC-RAS 1-D flood simulation model indicates the capability of simulating flood events and spatially depicting the degree of exposure or vulnerability of the region towards a hazard event in terms of inundation extent and depth of water levels.

Simulation of Ross Barnett Reservoir Dam failure under a high intensity rainfall event yielded results with the flood hazard impact (spatial extent and depth grids) extending till the end of Hinds County. The results revealed alarming information indicating the spatial extent and depth grids of flood hazard to equal a 200-year magnitude flood. The numbers of acres coming under flood waters and the maximum depths almost match with the 1979 catastrophic flood. These results bear utmost significance, as the current developmental activities in the downstream counties are bound to 100-year flood magnitude, while situated under a quite possible 200-year flood threat. Vulnerability assessment in this event of a 200-year magnitude flood hazard exposed the possible disturbances that can occur to cultural, economic, transportation and infrastructural amenities, affecting their inter-connected functionalities. Finally, the study developed a composite index to identify the spatial location of vulnerable areas by standardizing individual indicators at a census block level.

The overall objective of the research is to generate information for an improved or enhanced land use planning with respect to flood hazards. By exposing the long-term flood threats, the study assists the planning authorities at the local or county level in identifying vulnerable zones and incorporating the essence of information in its future developmental activities. The basic intention of planning or developmental strategies is to build safer communities by locating developments away from the hazard-prone areas. Identifying vulnerable areas under various possible scenarios plays an important role in the decision-making process. The increase in the vulnerability levels that might occur due to climate change affects downstream of Ross Barnett Reservoir can help the local government to improve the inclusive nature of environmental factors to their focus on achieving sustainable development.

References

1. IPCC. Summary for Policymakers. In *Climate Change 2007: The Physical Science Basis*; Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., Eds.; Cambridge University Press: Cambridge, UK/New York, NY, USA, 2007; p. 996.
2. Easterling, Williams. Adapting North American agriculture to climate change. *Agr. Forest Meteorol.* **1996**, *80*, 1–53.

3. United States Environmental Protection Agency. *Climate Change and Mississippi*; US EPA: Washington, DC, USA, 1998.
4. *City of Roseville. Dam Failure*. Available online: http://www.roseville.ca.us/civica/filebank/blobdload.asp?BlobID=19067 (accessed on 19 November 2012).
5. Zhu, N.H.; Turner, E.R.; Doyle, T.; Abdollahi, K. *The Potential Consequences of Climate Variability and Change*; GCRCC and LSU Graphic Services: Baton Rouge, LA, USA, 2003.
6. FEMA. *Grants Target Mississippi Flood Risks.* Avaiable Online: http://www.fema.gov/news-release/grants-target-mississippi-flood-risks (accessed on 21 August 2012).
7. Idaho National Laboratory. *Assessment of Potential Flood Events and Impacts at INL's Proposed Remote-Handled Low-Level Waste Disposal Facility Sites*; US Department of Energy: Idaho Falls, ID, USA, 2010.
8. Jami, E.; Scuddkr, M.; Johan, G.; Wilfrid, G. Lessons from a dam failure. *J. Sci.* **2000**, *100*, 121–131.
9. *Proactive Measures Needed to Prevent Future Dam Disasters*; Association of State Dam Safety Officials: Lexington, KY, USA, 2004.
10. US Army Corps of Engineers. *Pearl River Watershed Study. 2004.* Avaiable Online: http://www.mvk.usace.army.mil/offices/pp/projects/prws/background.htm (accessed on 17 August 2012).
11. Rutherford, P.H. The Jackson Flood of 1979 a public policy disaster. *J. Amer. Plan. Assn*. **1982**, *48*, 219–231.
12. MSMEMA. *The State of Mississippi Standard Mitigation Plan*; MSMEMA: Jackson, MS, USA, 2012.
13. Katopodes, N.D. Two dimensional Surges and shocks in open channels. *J. Hydraul. Div.* **1984**, *110*, 794–812.
14. Hromadka, T.V.; Berenbrock, C.E.; Freckleton, J.R.; Guymon, G.L. A two dimensional dam-break flood plain model. *Adv. Water Resour*. **1985**, *8*, 7–14.
15. Akbani, A.A. Model for flood propagation on initially dry land. *J. Hydraul. Eng.* **1988**, *114*, 689–706.
16. Zhao, D.H. Approximate riemann solver in FVM for 2D hydraulic shock wave modeling. *J. Hydraul. Eng.* **1996**, *122*, 692–702.
17. Sharma, A.K. *A Study of Two-Dimensional Flow Propagating from an Opening in the River Dike*; Gauhati Univeristy: Gauhati, India, 1999.
18. Zoppou, C.; Robert, S. Numerical solution of the two dimensional unsteady dam break. *Appl. Math. Model* **2000**, *24*, 457–475.
19. Davies, T.R.; Scott, B.K. Dam break flood hazard from the Callery River. *J. Hydrol.* **1997**, *36*, 1–13.
20. New Zealand Institute of Water and Atmospheric Research. *A Methodology to Assess the Impacts of Climate Change on Flood Risk in New Zealand*; Ministry of Environment: Wellington, New Zealand, 2005.
21. Bernardo, R.; Ramos, I. *GIS in Flood Risk Management*. Available online: http://libraries.maine.edu/Spatial/gisweb/spatdb/egis/eg94056.html (accessed on 11 August 2009).

22. Bajwa, H.S.; Tim, U.S. Toward immersive virtual environments for GIS-based Floodplain modeling and Visualization. In *Proceedings of 22nd ESRI User Conference*, San Diego, TX, USA, 8–12 July 2002.
23. Cameron, A.T. Geo Spatial Capabilities of HEC RAS for Model Development and Mapping. In *Proceedings of 2nd Joint Federal Interagency Conference*, Las Vegas, NV, USA, 27 June–1 July 2010.
24. Hicks, F.E.; Peacock, T. Suitability of HEC-RAS for flood forecasting. *Can. Water Res. J.* **2005**, *30*, 159–174.
25. Cameron, A.T.; Gary, W.; Brunner, P.E. *Dam Failure Analysis Using Hec-Ras And Hec-Georas.* Available online: http://www.gcmrc.gov/library/reports/physical/Fine_Sed/8thFISC2006/3rdFIHMC/11F_Ackerman.pdf (accessed on 23 September 2008).
26. Morris, G.L. *Dam Break Flood Hazard Analysis for Quebrada Beatriz Reservoir*; Puerto Rico Infrastructure Financing Authority: Caguas, Puerto Rico, 2007.
27. *City of Jackson*. *Historic Districts*. Available online: http://www.jacksonms.gov/assets/planning/historic%20districts.pdf (accessed on 24 August, 2009).
28. Burby, R.J. *Cooperating with Nature: Confronting Natural Hazards with Land Use Planning for Sustainable Communities*; Joseph Henry/National Academy Press: Washington, DC, USA, 1998.
29. Geertman, S.; Stillwell, J. Planning Support Systems: An Introduction. In *Planning Support Systems in Practice*; Springer: Berlin, Germany, 2003; pp. 25–55.
30. Yerramilli, S. A hybrid approach of integrating HEC-RAS and GIS towards the identification and assessment of flood risk vulnerability in the city of Jackson, MS. *Am. J. Geogr. Inform. Syst.* **2012**, *1*, 7–16.
31. Zheng, N.; Takara, K.; Tachikawa, Y.; Kozan, O. Analysis of vulnerability to flood hazard based on land use and population distribution in the Huaihe River basin, China. *Ann. Disaster Prev. Restor.* **2008**, *20*, 83–91.
32. Timothy, C.W. Vulnerability to environmental hazards in the Ciudad Juárez (Mexico)–El Paso (USA) metropolis: A model for spatial risk assessment in transnational context. *Appl. Geogr.* **2009**, *29*, 448–461.

OPEN ACCESS Freely available online

PLOS ONE

A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network

Wei Luo[1]*, Peifeng Yin[2], Qian Di[3], Frank Hardisty[1], Alan M. MacEachren[1]

1 GeoVISTA Center, Department of Geography, Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** PDA Group, Department of Computer Science & Engineering, Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

Abstract

The world has become a complex set of geo-social systems interconnected by networks, including transportation networks, telecommunications, and the internet. Understanding the interactions between spatial and social relationships within such geo-social systems is a challenge. This research aims to address this challenge through the framework of geovisual analytics. We present the GeoSocialApp which implements traditional network analysis methods in the context of explicitly spatial and social representations. We then apply it to an exploration of international trade networks in terms of the complex interactions between spatial and social relationships. This exploration using the GeoSocialApp helps us develop a two-part hypothesis: international trade network clusters with structural equivalence are strongly 'balkanized' (fragmented) according to the geography of trading partners, and the geographical distance weighted by population within each network cluster has a positive relationship with the development level of countries. In addition to demonstrating the potential of visual analytics to provide insight concerning complex geo-social relationships at a global scale, the research also addresses the challenge of validating insights derived through interactive geovisual analytics. We develop two indicators to quantify the observed patterns, and then use a Monte-Carlo approach to support the hypothesis developed above.

Citation: Luo W, Yin P, Di Q, Hardisty F, MacEachren AM (2014) A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network. PLoS ONE 9(2): e88666. doi:10.1371/journal.pone.0088666

Editor: Bin Jiang, University of Gävle, Sweden

Received October 13, 2013; **Accepted** January 14, 2014; **Published** February 18, 2014

Funding: This material is based, in part, upon work supported by the United States Department of Homeland Security under Award #: 2009-ST-061-CI0001. Here is the website: http://www.dhs.gov/st-oup. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Department of Homeland Security; a grant from the Gates Foundation also provided partial support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that they have no competing interests exist. This material is based, in part, upon work supported by the United States Department of Homeland Security under Award #: 2009-ST-061-CI0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Department of Homeland Security; a grant from the Gates Foundation also provided partial support.

* E-mail: wul132@psu.edu

Introduction

The world has become an increasingly interconnected system with multi-scale geographically embedded networks (i.e., transportation, internet). Spatial analysis aims to understand such systems in terms of spatial patterns, relationships, processes, and change within and among geographical spaces [1]. Social network analysis has been used to understand how systems emerge through the interaction of individual agents (i.e., humans, companies). Both approaches have advantages and limitations as methods through which to understand the complex geo-social interconnected world. Many geo-social interconnected systems mainly grow from the bottom-up, but traditional spatial analysis is a top-down approach that cannot deal with the evolution of the systems over space and time [2,3]. Social network analysis, a bottom-up approach, can link individual-level behaviors and interactions to the emergence of social phenomena [4], but the approach typically ignores geographical constraints [5]. An effective integration of both approaches has the potential to aid understanding of geo-social systems from a more comprehensive perspective. For example, the integration of spatial consideration into a social network approach enables understanding of why and how an air-borne disease diffuses within an urban area in a manner that can generate disease hot spots as well as cold spots [6]. The integration of spatial analysis and social network analysis has the potential to link individual-level behaviors and interactions (i.e., human, vehicle, organization) to understand urban sprawl over space and time [4]. Although spatial analysis and social network analysis have the potential to complement each other, the formal integration of two approaches remains relatively underdeveloped in the literature [7].

This paper therefore integrates spatial analysis and social network analysis into a unified framework through a geovisual analytics approach. Geovisual analytics tools integrate computational methods with interactive visualization, in order to enable insights on large and complex geospatial datasets [8,9,10,11]. Specifically, we present and apply a geovisual analytics tool, GeoSocialApp [12], that consists of three major analytical "spaces" implemented as linked components: a geographic space, a network space, and an attribute space. Each performs a specific task and can coordinate with other components to facilitate a process through which insights are enabled. We illustrate how the GeoSocialApp facilitates development of hypotheses, with the

international trade network (ITN) as a case study. The explicit geographical and network representations in the GeoSocialApp facilitate and enable insight in terms of different roles that spatial and social relationships have in the ITN across geographical regions with network hierarchies at different scales. One major goal of geovisual analytics is to develop hypotheses on how space matters based on the patterns identified from geo-spatial data [13]; but the validation of geovisual analytics results is still regarded as a challenge [14]. Here, we propose a Monte-Carlo approach as a statistical validation to support the hypothesis developed through visual-computational exploration of spatial and social interaction in the ITN.

The paper begins below by reviewing the development of geo-social visual analytics methods in geography and network domains (Section 2). We then present an overview of the methods (Section 3) and the international trade network data used in this study (Section 4). The results obtained through applying the methods to the data (Section 4) provide insights on the different roles that spatial and social relationships play in relation to trade across geographical regions (Section 5). We next introduce the Monte-Carlo approach as a statistical validation to support the insights discussed in section 5 (Section 6). Finally, we present conclusions and an outlook for future research (Section 7).

Literature Review

Current geo-social visual analytics tools can be classified into two major groups: the first group, rooted in geography, focuses on geographical analysis with an implicitly network representation; the second group, rooted in social network science, has an explicitly network representation with geography as a background to visualize the results. This section reviews the geo-social visual analytics tools from geography and social network science domains, and argues for a more balanced approach that emphasizes spatial relationships and social networks simultaneously.

Spatial interactions/flows associated with topics such as human migration and disease transmission are major research domains for integrating network representation into geovisual analytics. For example, Andrienko and Andrienko [15] develop a spatial generalization method to transform trajectories with common origins and destinations into aggregated flows maintaining essential characteristics of the movement between areas. In complementary research, Guo [16] proposes an integrated interactive visualization framework that is applied to county-to-county migration data in the U.S. in order to visualize and discover network structures, multivariate relations, and their geographic patterns simultaneously. Additional relevant research can be found in recent papers by Andrienko et al. [17], Demšar and Virrantaus [18], Guo, Liu and Jin [19], and Wood, Dykes and Slingsby [20].

All of the above studies consider the geo-social processes from a primarily geographical perspective. Spatial interactions/flows in research taking this perspective are typically visualized on maps, which provide important information on spatial context. The observed spatial patterns can be related to the spatial context (e.g., big cities tend to be hotspots for human interaction). The methods for geo-social interaction discussed so far assume that geographic locations define the geo-social process, but new communication and transportation technologies clearly spread social networks beyond traditional geographical constraints (i.e., distance) [21]. Therefore, understanding the social meaning behind the geo-social processes is equally important.

Geo-social visual analytics from a social network science perspective tends to have an explicit network context with an implicitly geographical representation. Ahmed et al. [22] introduce new visual analysis methods with dynamic network views (e.g., wheel layout, radial layout, and hierarchical layout) to explore the 2006 International Federation of Association Football (FIFA) World Cup competition in which countries are clustered based on their geographical locations in the dynamic graph representation. The visual analysis methods allow users to analyze and compare each country's performance within the geo-social context. The explicit network representation and implicitly geographical representation require analysts to relate the explicit network representation to his or her unrepresented geographic background knowledge in the visually interactive process [8]. Thiemann [23] developed the SPaTo Visual Explorer, which implements multiple explicitly geographical and network representations. Using a case study focused on global air flight networks, he illustrates how SPaTo can allow users to develop hypotheses about the interaction between geographical distance and social network distance. For example, they derive evidence showing that geographical proximity of cities corresponds with short social distance among the cities. Beyond the above, four additional research efforts have focused on specific components of methods to involve explicitly geographical representations into a traditional social network approach: 1) spatial point pattern exploration approach (e.g., kernel density) can be used to understand spatial impacts on the development of social networks [24]; 2) spatial autocorrelation coefficient (e.g., Moran's I) has been applied to social networks to measure the statistical similarity of individuals [25]; 3) explicitly spatial representations facilitate practical implementation of decision-making in certain social network application domains (e.g., infectious disease control) [26]; and 4) certain geo-social systems (e.g., human migration, international trade network) can be better understood or predicted through mathematical models considering physical and social space [27,28].

As discussed above, understanding geo-social systems requires consideration of both geographical relationships and social network relationships. Therefore, it is necessary to involve explicitly geographical and social network representations. Andris [29] lists five benefits to having an explicit network representation within a geo-spatial framework: 1) the group of connected geographical regions can be studied as a unit with social closeness based on a network community detection approach; 2) the social power of places can be represented by node measures (i.e., degree, betweenness); 3) the social role of interconnected places over the whole system can be represented by network system measures (i.e., degree distribution, betweenness distribution); 4) the complex social interaction between places can be understood through adding multiple social flow layers on Geographical Information System (GIS); and 5) the geo-social systems in which spatial closeness and social closeness do not match can be better modeled with an explicit network representation.

The above discussion illustrates that there is the lack of explicitly spatial and social network representations in current geovisual analytics and the importance of such representations to understand geo-social systems [30]. It is also still a challenge to statistically support the hypotheses developed through visual exploration [31], particularly the hypotheses directed to geo-social interaction. To fill the gap, this paper introduces the GeoSocialApp with the 2005 international trade network as a case study to understand the interaction between spatial and social relationships, and introduces the use of a Monte-Carlo approach to validate the hypothesis developed in our geo-social visual exploration.

Methods

In this paper, we extend and apply the GeoSocialApp, a geovisual analytics tool initially introduced in preliminary form in Luo et al. [12]. The GeoSocialApp implements traditional network analysis methods within the context of an environment that links explicitly spatial and social representations to understand the interaction of spatial and social relationships in the ITN. The GeoSocialApp is an extension of the GeoViz Toolkit (GVT) developed in the GeoVISTA Center at Penn State [32]. The research presented here makes use of the existing choropleth mapping capabilities of GVT to support geographical analysis as well as the component coordination methods that enable dynamic linking and brushing across views, and adds a dendrogram component that supports multiple graph-based views to represent a varying network hierarchy. Details about other GVT components that could be used to extend the analysis presented here can be found in http://www.geovista.psu.edu/GeoSocialApp/ (The source code for the GeoSocialApp is open source under the Library General Public License, version 2 (LGPL 2.0). We plan a public release of a binary version usable by non-programmers in the future).

GeoSocialApp Components

As noted above, we use two components in the GeoSocialApp for this study: a dendrogram view and a choropleth view. The dendrogram view implements the convergence of the iterated correlations (CONCOR) algorithm [33,34] to group nodes with equivalent positions in a single network or multiple social networks together. Equivalent positions refer to collections of actors that have similar ties to and from all other actors in the network. The implication of actors having equivalent positions is that they play similar social roles in a relational network. We can describe the relational network by an adjacency matrix A, which can generate a position similarity matrix R to measure the equivalent positions, whose element value r_{ij} is defined as:

$$r_{ij}=\frac{\sum (x_{ki}-\overline{x}_{\bullet i})(x_{kj}-\overline{x}_{\bullet j})+\sum (x_{ik}-\overline{x}_{i\bullet})(x_{jk}-\overline{x}_{j\bullet})}{\sqrt{\sum (x_{ki}+\overline{x}_{\bullet i})^2+\sum (x_{ik}+\overline{x}_{i\bullet})^2}\sqrt{\sum (x_{kj}+\overline{x}_{\bullet j})^2+\sum (x_{jk}+\overline{x}_{j\bullet})^2}} \quad (1)$$

where $\bar{x}_{i\bullet}(\bar{x}_{j\bullet})$is the mean of the values in row i (j) of the matrix A and $\bar{x}_{\bullet i}(\bar{x}_{\bullet j})$is the mean of the values in column i (j) of the matrix A. At the initial level of analysis, CONCOR performs the above equation calculations iteratively on the position similarity matrix R until all values converge to either 1 or −1, resulting in all nodes being grouped into one of two categories. Two groups can be too generalized for some studies, so hierarchical structures can be achieved by running CONCOR on each subgroup. In this way, CONCOR can continue to split nodes into successively smaller groups: two become four, four become eight, and so on. Although this algorithm was developed originally for application to social networks of individuals, it has been demonstrated to be an effective method to empirically locate structural positions in terms of the ITN [12,35].

Equivalent positions in terms of the ITN refer to collections of countries that have a similar import and export trade relationships with all other countries [36]. The implication of countries having equivalent positions is that they play similar social roles in the ITN. According to world system theory, the economic development of different countries is affected by their structural positions: core, semi-periphery, and periphery through unequal economic exchanges among them [37]. Core countries focus on capital-intensive production, periphery countries provide low-skill labor and raw materials, and semi-periphery countries are the industrializing countries positioned between the periphery and core countries. The CONCOR algorithm can classify the ITN into these three structural equivalence positions [38,39].

A tree layout and a radial layout are implemented in the dendrogram view to visualize the hierarchical structure of CONCOR results (Figure 1). The tree layout organizes the graph in a hierarchical way by placing child nodes under their common ancestors. An informationally equivalent radial view can be transformed from the tree by putting child nodes in the enclosing circle of their common ancestors [40,41]. The dendrogram view in the GeoSocialApp also provides a slider to control the hierarchical level of CONCOR results.

The dendrogram view of social space is dynamically linked to a choropleth map view used for visual exploration in geographical space. Each node in the dendrogram view corresponds to a geographical unit (i.e., states, countries) in the choropleth map. The choropleth map allows users to choose the number of classes, the classification method (i.e., equal intervals, quantiles), the variable to display, and the ColorBrewer palette [42] for color selection. Thus, the linked dendrogram and map views allow exploration of social positions and social groups and their corresponding spatial positions and spatial groups simultaneously. With the hierarchical level control in the dendrogram view, the linked views further support the explicit exploration of interaction between social space and geographical space and its impact on outcomes of interest at different network hierarchy (Figure 2). This capability will be illustrated in the case study presented below, after the data used in that case study are first described.

Data

Our analysis of the interaction between spatial and social relationships in the ITN is based upon import and export data among 192 countries in 2005. These data were extracted from the CorrelatesOfWar (COW) Database and include volume of imports and exports in current U.S. dollars [43]. We convert the 2005 ITN data into a directed network in which countries are the nodes of the network and an import/export trading relationship is represented by a link between two countries. We then organize the data into a binary matrix form to fit the CONCOR algorithm with columns as exporting countries and rows as importing countries. As an illustration, Table 1 is the original import and export data among sample countries in 2005, and Table 2 is the binary matrix for the first 10 countries in our data; "1" represents presence of import/export trade between countries, "0" represents no trade. A binary matrix is used rather than a weighted matrix for twofold reasons: one basic idea of the CONCOR algorithm is that the primary indicator of a relationship is the absence of links between individuals rather than the occurrence of the links [44]; given this idea, the past research in international trade has typically used the binary matrix with the CONCOR algorithm to identify three structural equivalence positions: core, semi-periphery, and periphery [38,39,45].

We use three additional data variables: GDP, population, and geographical distance, to validate the hypothesis developed through visual exploration using the GeoSocialApp. We downloaded 2005 GDP and population data for each country from the World Bank website (http://data.worldbank.org/). We calculated the linear distance between national capitals to measure the geographical distance between countries with ArcGIS. This measure of between-country distance is picked over others (e.g., distance between country centroids, distance between the nearest points of country borders, etc.), because gravity models used in

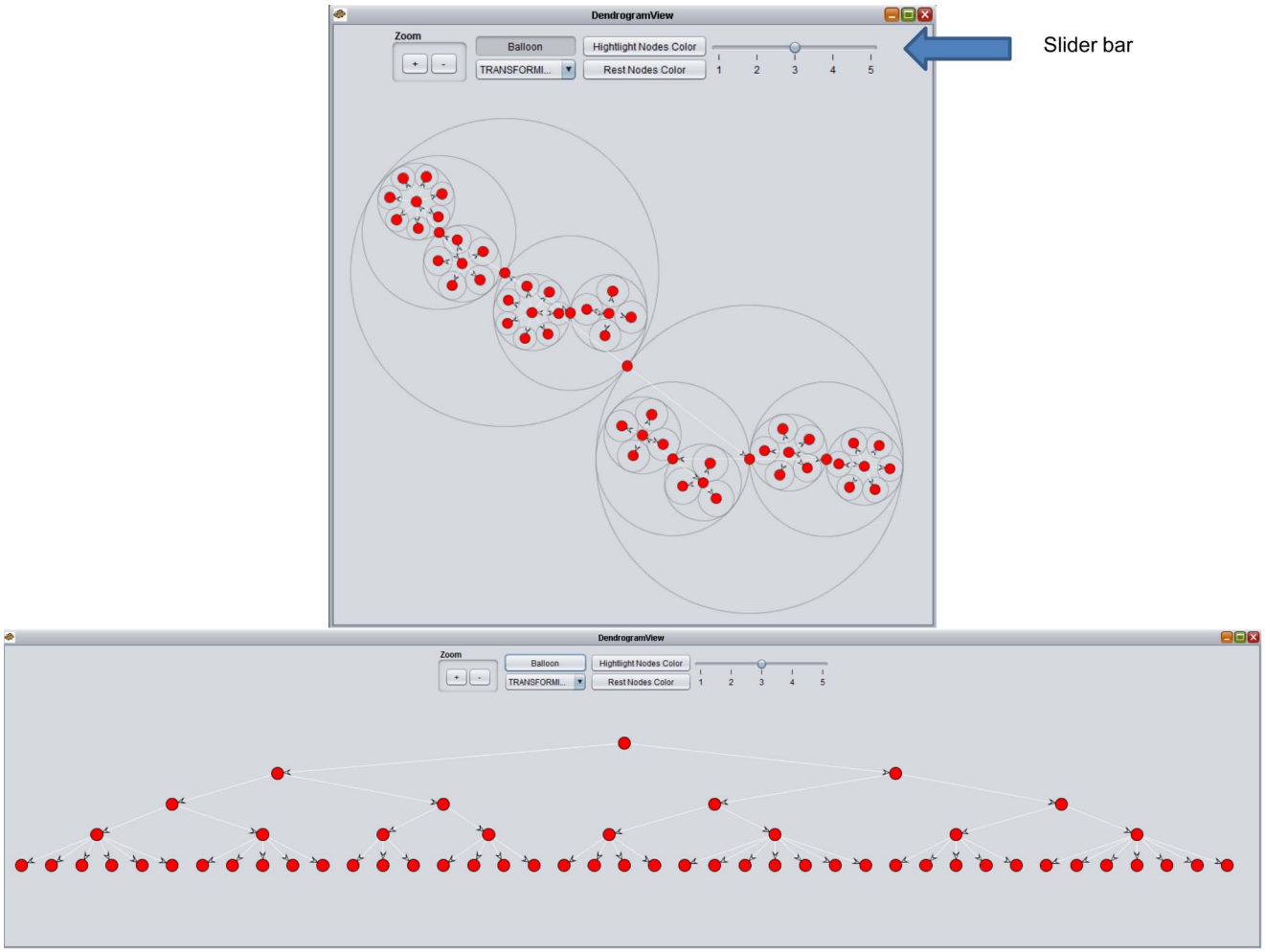


Figure 1. Dendrogram View. Two layouts to visualize the hierarchical structure of CONCOR results: the left one is a tree layout and the right one is a radial layout. Slider bar is used to control the level of CONCOR results.
doi:10.1371/journal.pone.0088666.g001

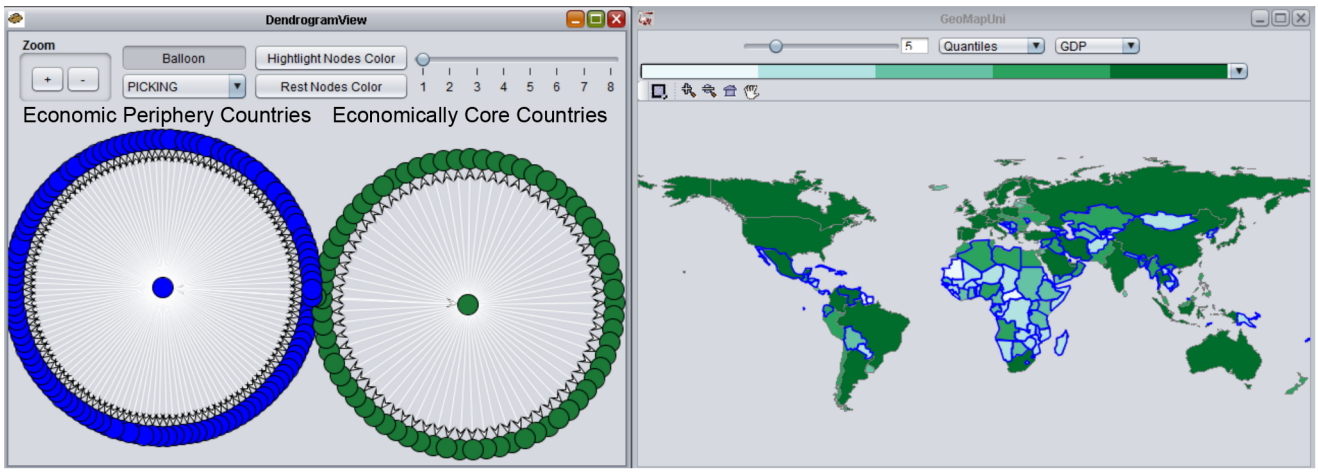


Figure 2. Dendrogram view and choropleth map view. The choropleth map depicts GDP by country. Data are divided into quintiles (5 categories with an equal number of countries in each category) depicted by 5 sequentially ordered shades of green, from low GDP (very light green) to high GDP (very dark green). Each node in the dendrogram view corresponds to one country in the choropleth map view (The highlighted nodes in blue correspond to countries with borders highlighted in blue). The first run of CONCOR process reveals two positions in the 2005 ITN.
doi:10.1371/journal.pone.0088666.g002

Table 1. Imports-exports relationship among partial countries in 2005.

year	importer1	importer2	flow1	flow2
2005	United States of America	Canada	291944	195151
2005	United States of America	Bahamas	726.3	1945.79
2005	United States of America	Cuba	0	397.87
2005	United States of America	Haiti	458.5	756.91
2005	United States of America	Dominican Republic	4721.4	5179.24
2005	United States of America	Jamaica	410.9	1962.2
2005	United States of America	Trinidad and Tobago	8342.2	1583.01
2005	United States of America	Barbados	33.4	595.28
2005	United States of America	Dominica	3.8	67.43

Flow1 means imports of importer1 from importer2 in current US millions of dollars, and flow2 means imports of importer2 from importer1 in current US millions of dollars.

doi:10.1371/journal.pone.0088666.t001

other international trade network studies use the same distance measure [46].

Results

Spatial and Social Interaction at the First Level of CONCOR

We use the dendrogram view in the GeoSocialApp to explore Table 2 to identify social relationships among all countries, and the univariate choropleth map to visualize the spatial distribution of GDP for all countries (Figure 2). Comparing the dendrogram view and the map view, and using the dynamic linking between them to explore specific details for individual and groups of countries, can provide insight about spatial and social interactions within the ITN.

Initially, we use the dendrogram view to divide the network data into two groups. After highlighting one group (blue nodes in the dendrogram view and blue outlines in the map view), we find that most countries in the highlighted group are economic periphery countries (i.e., most countries in Central America and Africa) and most countries in the other group are economically core countries (i.e., North America and European Union). The univariate choropleth map depicts GDP for each country. The sequential colors reinforce this classification: economically less-productive countries are indicated by light green, whereas other, more economically productive countries are indicated by dark green. The two classifications identified by CONCOR imply that economically core countries tend to have similar international trade partners, and economic periphery countries tend to have similar trade partners. This study focuses on the interaction between spatial and social relationships in the ITN. At the first level of CONCOR in Figure 2, all countries with close social relationships tend to exhibit spatial proximity.

Spatial and Social Interaction at the Second Level of CONCOR

The second application of CONCOR to the ITN subdivides the first two categories, resulting in a total of four groups as shown by Figure 3 (A list of countries for each group is in File S1.). The core countries and the periphery countries are partitioned into four new geographies, which further indicate a core–periphery arrangement: the mean GDP for each geography is sorted in Table 3. Figure 3A mainly includes more developed countries in the economically core group: North America, most countries in Europe, Australia, South Africa, and economically more-important countries in Asia (i.e., China, India), whereas Figure 3B mainly consists of less developed countries in the economically core groups: Russia, most countries in South America, and a small number of countries in Europe. Figure 3C mainly includes more developed countries in the economic periphery group: Central America, and a few countries from Eurasia (i.e., Vietnam, Iran), whereas Figure 3D mainly consists of the less developed countries in the economic periphery group: countries from Africa and some countries from Asia (e.g., Mongolia). In terms of spatial and social interaction identified by the second level of CONCOR, economically core countries in Figure 3A and Figure 3B (i.e., North America, Europe), as well as more developed periphery countries in Figure 3C exhibit regional patterns (i.e., Central America, Central Asia) that also fall into the same social groups across the globe. It suggests that international trade partners for those countries are related to both spatial proximity and similar economic development level (Figure 3A, 3B, and 3C). Economic periphery countries in Figure 3D have one major cluster (i.e.,

Table 2. International trade relationships among partial countries in a binary matrix for 0% threshold in 2005.

	GUATEMALA	BOLIVIA	PARAGUAY	URUGUAY	SURINAME	GAMBIA	MOROCCO	MALI	LIBERIA
GUATEMALA	0	1	1	1	0	0	1	0	0
BOLIVIA	1	0	1	1	0	0	1	0	0
PARAGUAY	1	1	0	1	0	0	1	0	0
URUGUAY	1	1	1	0	1	0	1	0	0
SURINAME	1	0	0	1	0	0	1	1	1
GAMBIA	0	0	0	0	0	0	1	1	1
MOROCCO	1	0	1	1	1	1	0	1	1
MALI	0	0	0	1	0	1	1	0	0
LIBERIA	0	0	1	1	1	0	1	0	0

doi:10.1371/journal.pone.0088666.t002

Africa). Compared to 3A, 3B, and 3C, Figure 3D suggests that spatial proximity has a stronger impact on the least developed countries in terms of international trade partners they have.

Spatial and Social Interaction at the Third Level of CONCOR

The third run of CONCOR applied to the ITN again subdivides the previously identified groups into seven different subgroups (Figure 4) (A list of countries for each group is in File S1.). At this level the geographies are considerably more complex but this research highlights three features. First, only seven new subgroups are identified in this level: CONCOR does not divide countries depicted in Figure 3A any further, resulting in the same group of countries in Figure 4A, because economically core countries in this group have highly similar import and export trade partners. Second, some groups of countries at this level further confirm a core-periphery hierarchical structure in terms of the ITN: the top economically core countries in Figure 4A; a clear distinction between east African countries (the second least developing places) in Figure 4F and west African countries (the least developing regions) in Figure 4G. Third, the role that spatial and social relationships play in terms of the ITN identified by the third level of CONCOR becomes more noticeable. Core countries in Figure 4A, Figure 4B, and Figure 4C have their own distinct geographical regions (i.e., North America, Europe), but social relationships to connect different regions are also strong. Figure 4D and Figure 4E identify two distinct geographical regions (Central America and Central Asia) compared to Figure 3C that put both into the same social group. The distinct geographical regions suggest that spatial constraints are stronger than social connections between the two regions at this network level. Comparing the two distinct geographical regions identified in Figure 4D and Figure 4E to distinct geographical regions (i.e., North America, Europe, and Austria) in Figure 4A suggests that spatial constraints have less impact on economically core countries and more impact on economic periphery countries to determine the international trade partners they have.

Validation

As outlined above, using an interactive visual approach, we found that developing countries with structural equivalence tend to exhibit a pattern of geographical proximity, and developed countries with structural equivalence tend to exhibit a pattern in which geographical proximity remains a factor, but one that is overcome by some connections to distant places. Based on the patterns, we develop the two-part hypothesis that: international trade network clusters with structural equivalence are strongly 'balkanized' (spatially fragmented) according to geography of trading partners, and the geographical distance within each network cluster has a positive relationship with the development level of countries. However, we wish to verify this visual finding with a more robust statistical verification. We have two steps to verify the hypotheses. The first step introduces two indicators (degree of balkanization and Pearson of correlation) to quantify the observed patterns, and the second step uses a Monte-Carlo method to measure the statistical level of the two indicators. It is also important to note that these two linked parts of the analytic process (visual hypothesis generation and confirmatory analysis) provide an iterative means of arriving at stronger conclusions.

Degree of balkanization

The first part of our hypothesis is that the network cluster with structural equivalence is strongly 'balkanized'. First, we calculate the average distances between countries that (i) belong to the same cluster and (ii) belong to two distinct clusters. The difference between both distances is a quantification of the degree of balkanization, denoted as B. That is to say:

$$B = \overline{D_{i,j}} - \overline{D_{m,n}}\ i,j \in \text{thesamecluster};\ m,n \in \text{differentclusters}$$

$$D_{x,y}\ \text{isthedistancebetweencountry}x\text{andcountry}y$$

$$\overline{D}\text{istheaveragedistance}$$

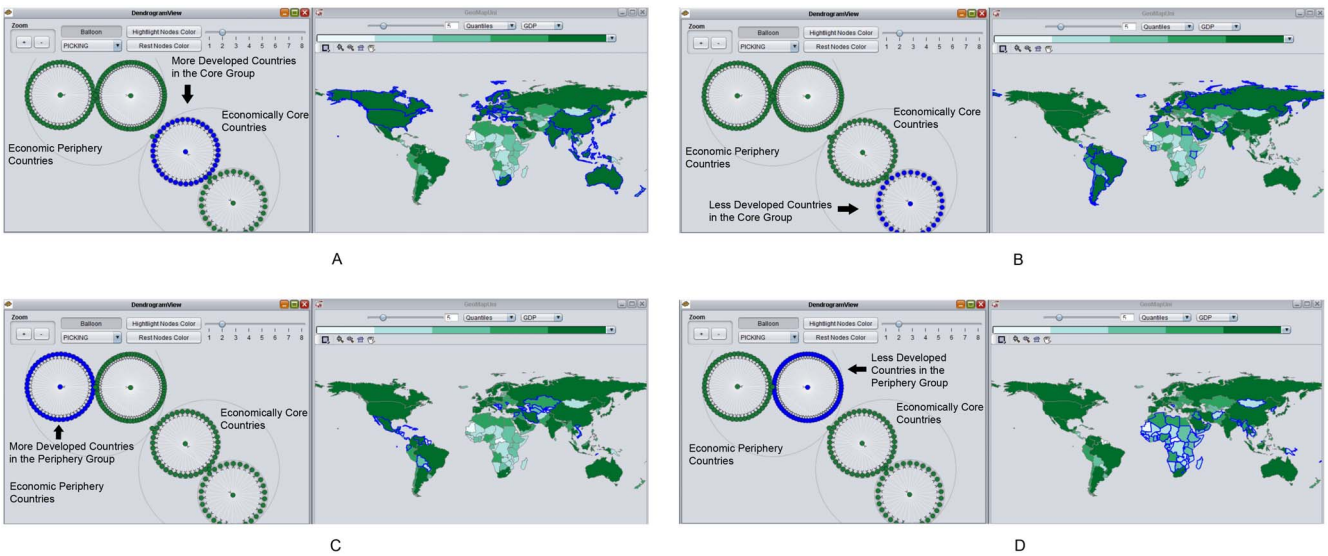


Figure 3. The second run of the CONCOR process subdivides each of the first two groups. Figure 3A: One subgroup of economically core countries; Figure 3B: The other subgroup of economically core countries; Figure 3C: One subgroup of economic periphery countries; Figure 3D: The other subgroup of economic periphery countries.
doi:10.1371/journal.pone.0088666.g003

Table 3. CONCOR group level attribute data.

FigureID	Mean GDP(billions of dollars)	FigureID	Mean GDP(billions of dollars)	Mean Distance(km)	Weighted Distance	Mean GDP(billions of dollars)
3A	912.00	4A	912.00	6664	5.55E+19	912.00
3B		4B	384.00	7146	9.76E+18	384.00
3B	250.00	4C	116.00	8086	5.15E+18	116.00
3C		4D	48.10	3403	1.84E+17	48.10
3C	34.90	4E	21.80	5125	4.67E+17	21.80
3D		4F	13.70	8838	1.19E+18	13.70
3D	12.30	4G	10.90	5833	6.37E+17	10.90

*Mean GDP in 2005 for 4 groups identified at the second level of the CONCOR, mean GDP in 2005, mean distance, weighted distance by population for 7 groups at the third level of the CONCOR.

doi:10.1371/journal.pone.0088666.t003

A positive value of B means that countries that belong to the same trade cluster are geographically grouped: the higher the positive value, the higher the degree of balkanization. If B is equal to zero, the countries from the same cluster have no geographic proximity at all and display a random geographic distribution. A negative value of B indicates that countries from the same trade cluster are geographically dispersed. The degree of balkanization of 2005 international trade data set is denoted as $\tilde{\mathbf{B}}$, with value of

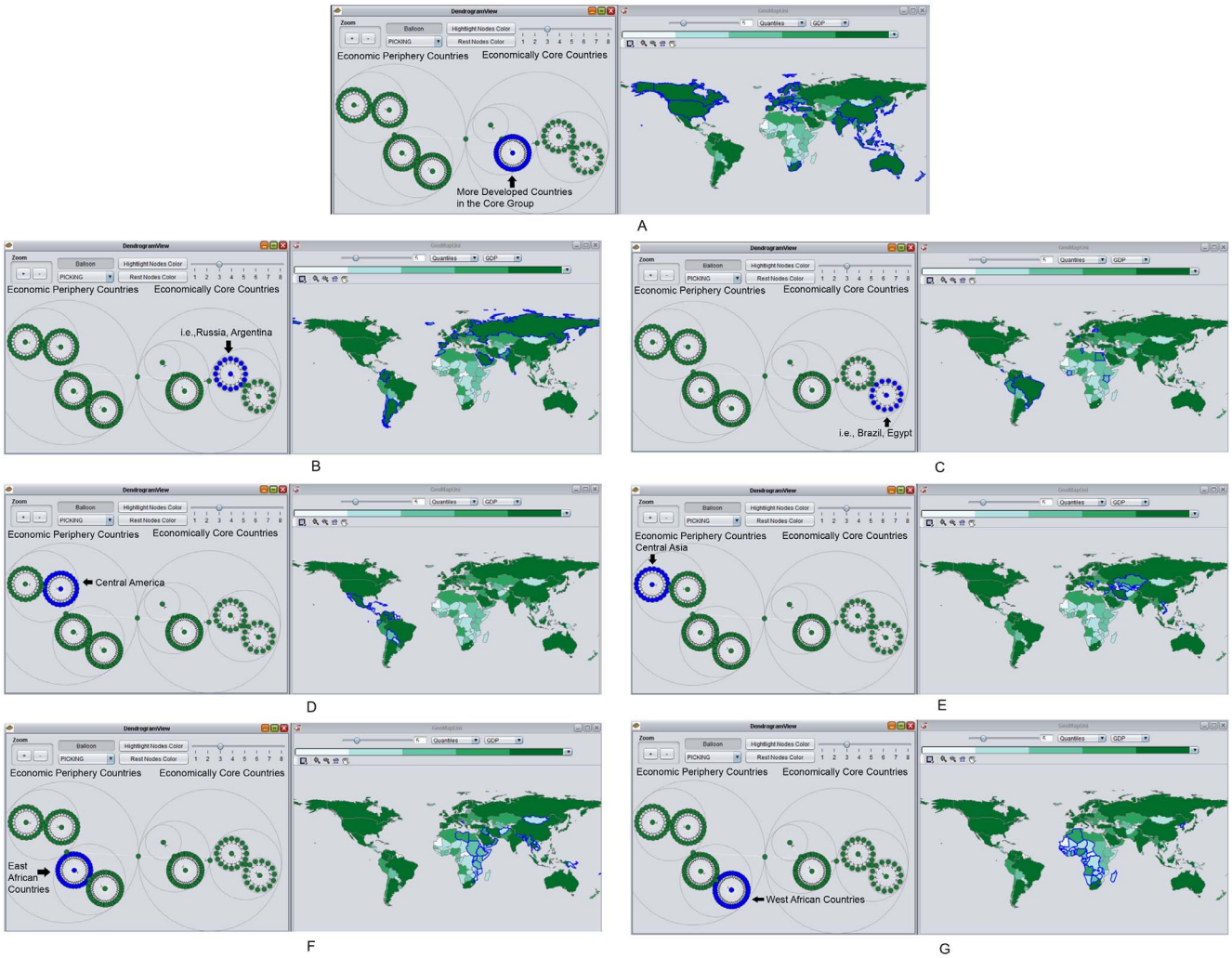


Figure 4. The third run of the CONCOR process continues to subdivide groups. Figure 4A 4B, and 4C belong to the economically core countries, whereas Figure 4D, 4E, 4F, and 4G belong to the economic periphery countries.

doi:10.1371/journal.pone.0088666.g004

2774.008 km. The absolute value indicates little about the degree of balkanization unless it is compared to some benchmark. The Monte-Carlo method can provide such a benchmark and produce a statistical significance measure of the absolute result, which we will discuss after describing our approach to measuring the relationship between GDP and distance by network cluster.

Pearson correlation

We use Pearson correlation [47] to measure the positive relationship between geographical distance within each network cluster and the development level of countries, which is determined by GDP in this paper.

$$P_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(G_i - \overline{G})(D_i - \overline{D})}{\sigma_G \sigma_D}$$

G_i is the average GDP of each cluster. D_i is the average within-cluster distance of each cluster. σ_G is the standard deviation in terms of average GDP of each cluster. σ_D is the standard deviation in terms of average within-cluster distance of each cluster. P ranges from -1 to 1. A positive P value implies that there is a positive relationship between geographical distance within each network cluster and GDP. A negative P value implies that geographical distance increases as GDP decreases. If P is around zero, it means that the geographic factor of each network cluster is independent from GDP.

When we calculate the average within-cluster distance, we give more weight to the countries that are more populous by weighting the distance by the population. The reason for this is explained below. The Pearson correlation between the average within-cluster distance without weight and GDP is only 0.13; this does not reflect the strong relationship that is apparent between the two variables as observed visually from the GeoSocialApp. We checked the GeoSocialApp again in order to figure out the reason behind this initial result. We found that simply calculating the average distance between any pair of countries may introduce some noise. For example, island countries in the middle Pacific (Figure 4F) that are far away from any other countries may raise the average within-cluster distance. The cluster in Figure 4F includes mainly developing countries in North Africa and the Mideast, as well as some island countries (e.g. Solomon Islands, Vanuatu). These islands only represent 1.5% of the population and 3.8% of the GDP for the cluster, but increase the within-group distance by 47.71%. Such a dramatic rise of within-group distance makes the distance-GDP nexus indistinct and brings down the Pearson correlation. We test the impact of those islands on the Pearson correlation through removing those islands in Figure 4F, which raises the correlation to 0.36. Given the similar issue existing in some of the other clusters (i.e., Figure 4D, 4E), we weight the distance between all countries proportionally to their population without removing any island countries (Table 3). Following from these preliminary results, we refine our hypothesis into: the geographical distance weighted by population within each network cluster has a positive relationship with the development level of countries. The 2005 international trade data set's Pearson correlation ($\tilde{\mathbf{P}}$) between average GDP per cluster and population weighted within-cluster distance is determined to be 0.97.

Validation Method

Here, we use a Monte-Carlo method to assess the hypothesis generated from visual-computational exploration. Monte-Carlo methods are a set of mathematical tools that use randomly generated data to evaluate mathematical expressions or to achieve the distribution of some desired variables [48]. Results that are generated from the random inputs serve as benchmarks to determine whether the phenomenon we have observed exhibits a statistically significant difference from that generated by a random process, thus whether the phenomenon is unlikely to have occurred by chance.

To start, we generate 10,000 random international trade networks. The basic idea of this data simulation process is to create trade networks with equal numbers of nodes and links, but to connect the nodes randomly. We keep the number of nodes and links constant to make clustering results from random trade networks comparable to results from the actual ITN data. For each random network, the degree of balkanization B and Pearson correlation P are calculated after performing the CONCOR algorithm. The 10,000 results offer a numerical approach to calculate the statistical significance of the original degree of balkanization and Pearson correlation by counting the percentage of random networks that have an equal or larger degree of balkanization or Pearson correlation. For the 2005 international trade data set, the degree of balkanization ($\tilde{\mathbf{B}}$) and the statistical significance (p value) of the Pearson correlation ($\tilde{\mathbf{P}}$) is calculated as follows:

$$p_{\tilde{B}} = \frac{\text{Number of random networks with } B \geq \tilde{B}}{\text{Totalnumberofrandomnetworks}}$$

$$p_{\tilde{P}} = \frac{\text{Number of random networks with } P \geq \tilde{P}}{\text{Total number of random networks}}$$

For this analysis, we set the confidence level for p at 0.05. Figure 5 shows the histogram of the degree of balkanization (B) based on all of the random trade networks. This figure shows an imperfect bell-shaped curve, culminating around 0. Its average mean is -0.54, which is very close to 0. An intuitive explanation is that the countries that belong to the same cluster have a random geographic distribution for most random trade networks. The p value of $\tilde{\mathbf{B}}$ is <0.0001, which means that less than one trade network within every 10,000 random trade networks has a clustering structure that equals or exceeds that of the 2005 international trade network. In other words, the observed high degree of balkanization within the 2005 trade data is unlikely to be a randomly produced result. Thus, the network cluster with structural equivalence exhibits statistically significant geographical clustering.

The Pearson correlation values calculated between the average GDP and the weighted within-cluster distance for all random trade networks are displayed in Figure 6. Unlike the previous result in Figure 5, the distribution of Pearson correlation values is irregular with one peak around 0.1 and another mini-peak around 0.9. That the majority of results are associated with the peak around 0.1 can be interpreted to mean that if trade networks were random, the relationship between GDP and the weighted within-cluster distance would be irrelevant or have very weak positive or negative relationship. The bi-modal distribution could be caused by a combination of clusters of countries with similar GDPs and the weighting procedure used. A nearly perfect correspondence between trade clusters and GDP is possible, but if trade links are broken, the patterns rapidly decohere into the default slight positive correlation. Only a small portion of random trade networks exhibit a strong positive relationship between these two variables. The p value is 0.0171, which is significant at 0.05

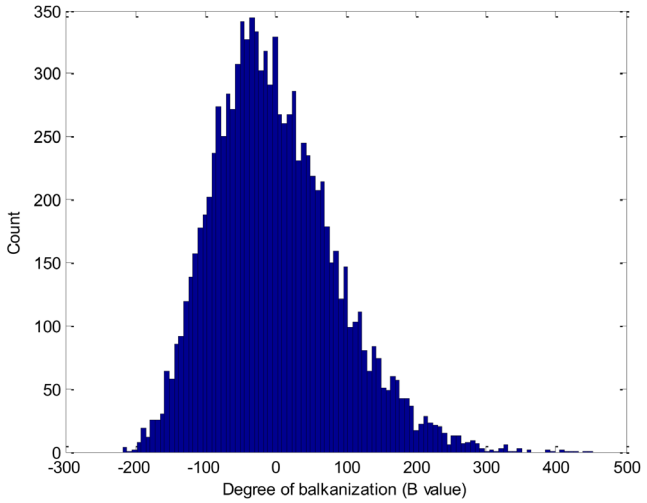


Figure 5. The degree of balkanization of all random trade networks.
doi:10.1371/journal.pone.0088666.g005

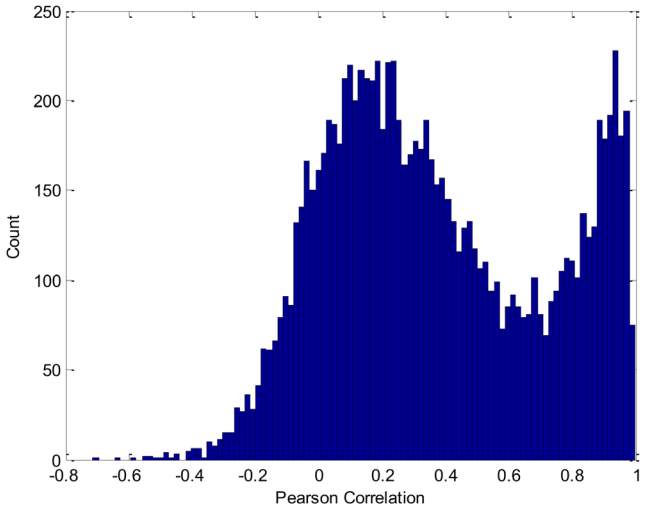


Figure 6. The Pearson correlation values between GDP and weighted within-cluster distance of all random trade networks.
doi:10.1371/journal.pone.0088666.g006

confidence level. It indicates that less than 2 of every 100 random trade networks display a stronger correlation between GDP and weighted within-cluster distance than found in the actual 2005 ITN data. In other words, the observed strong positive relationship from the visual exploration is unlikely to occur randomly, and the positive relationship between weighted geographical distance within each network cluster and the development level of countries is statistically significant.

Robustness of the validation method

We use two approaches to test the robustness of the validation results. The first approach is to change the number of runs for each Monte-Carlo validation. The second approach is to create random trade networks with different total connection numbers. For both approaches, we keep the number of nodes constant to make clustering results from random trade networks comparable to original results. If two tests exhibit consistent results with minor fluctuations, such results support that our validation method is robust against these kinds of changes. Similar test approaches have been used in other fields, such as meteorology [49].

The first approach examines whether the number of runs in each Monte-Carlo validation influences the final results. If results are robust, validation results will converge as the number of runs increases. Figure 7 displays the results in which the number of runs (N) is 1,000, 2,000, 5,000 and 10,000. When N is small, such as 1,000, the results display some reasonable fluctuations. As the number of runs rises, those results are smoothed and finally converge (as shown by the turquoise line on each plot representing 10,000 runs).

The second approach uses different numbers of connections among nodes to test the robustness of the validation. We examine the robustness with 50%, 75%, 100%, 150%, and 200% of the original connection number and rerun the validation methods. Figure 8 shows that the distributions of degree of balkanization and Pearson correlation are largely consistent based on the five different scenarios.

This section applies Monte-Carlo methods to validate the hypotheses developed from the GeoSocialApp-based visual-computational exploration of the 2005 ITN. Monte-Carlo simulation produces many randomized pseudo-networks, calculates statistical indicators, and compares the results with those from the original ITN. The results from the 2005 ITN analysis are shown to be statistically significant. In other words, the Monte-Carlo method verifies that the patterns we observe from the GeoSocialApp are unlikely to have resulted from random processes. Moreover, we test the robustness of the validation methods by changing the number of runs and the number of connections. In both scenarios, the Monte-Carlo method produces consistent results, which provides evidence that our validation method is robust.

Conclusion & Contribution

In this paper, we present the GeoSocialApp, a visual analytics application that supports exploration of the complex interaction between spatial and social network relationships and demonstrate its capabilities by investigating the ITN across geographical regions at different levels of the network hierarchy. The explicit focus of the GeoSocialApp on both geographical and social representations enables a process that generates insight related to the different roles that spatial and social relationships have within the varying network hierarchy levels. To address the network relationships, the GeoSocialApp implements the CONCOR algorithm that has been used in many past studies of the ITN. Although this algorithm has known limitations [50], our focus here is on demonstrating the potential of a geovisual analytics approach that integrates spatial and network analysis methods, not on developing novel methods to measure structural equivalence in networks. In addition, the CONCOR algorithm is still frequently used to measure structural equivalence of the ITN in recent research [28,45]. Thus, relying on a method with a long history was appropriate. The first run of CONCOR applied to our ITN data suggests a complex interaction between spatial and social relationships for the ITN, but also obscures the separate roles that each relationship has. The second and third run of CONCOR, identifying successively more homogeneous clusters, makes it clear that spatial constraints exist for all groups, but suggests that they are more influential for groups that include economic periphery countries.

Developing hypotheses about phenomena through visual-computational exploration is one major goal of visual analytics; but recent research recognizes that a weakness of many visual

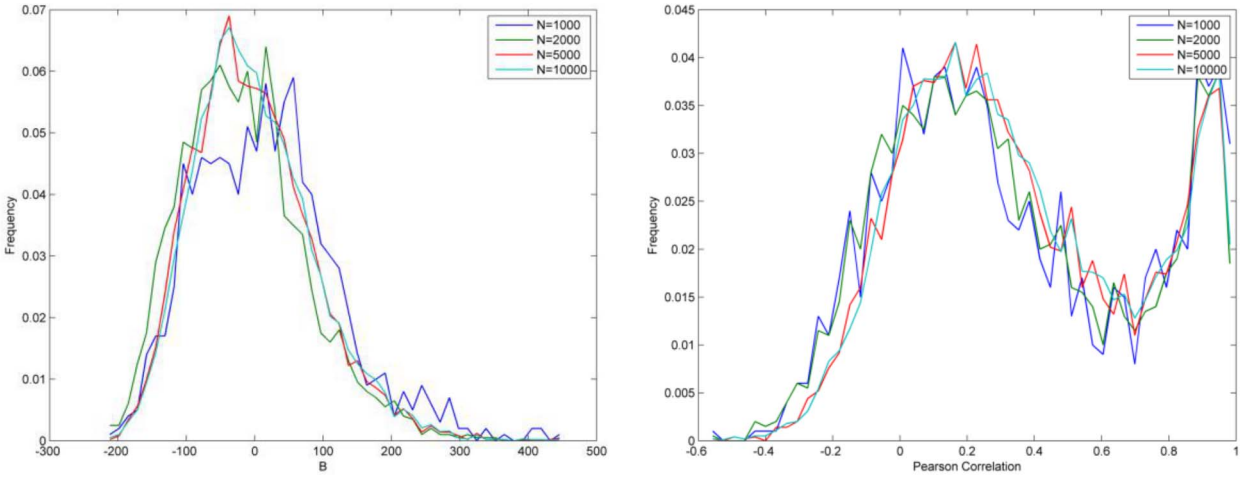


Figure 7. Validation results as a function of number of runs (N).
doi:10.1371/journal.pone.0088666.g007

analytics methods developed thus far is that they lack mechanisms to validate the hypotheses that are generated [14,31]. This research develops two indicators to quantitatively assess the patterns identified through visual-computational analysis and then uses a Monte-Carlo method with robustness tests to support our hypothesis with statistical evidence. In addition to using this method to test our hypothesis, we also use the feedback of our first statistical analysis, as discussed in the validation section, to refine our hypotheses. We propose that the approach outlined here may open a new research direction to support iterative hypothesis development, testing and refinement through combined visual-computational exploration and statistical validation.

A future goal for the GeoSocialApp specifically is to integrate this validation method directly within the tools. Monte-Carlo methods are suitable to validate the statistical significance of patterns identified through visual analytics for two reasons: a) patterns revealed through visual analytics tend to be complex and at the same time knowledge about their statistical distributions is absent in most situations; and b) one goal of Monte-Carlo methods is to achieve the distribution of some desired variables with randomly generated data [48]. To effectively integrate Monte-Carlo methods into the visual analytics tools, there are two major challenges: a) how to generate random data to provide baseline distributions based on different applications; and b) Monte-Carlo methods are time-consuming processes because they need to generate a sufficiently large number, e.g., 10,000, of new random data and then calculate the distribution of the desired variables. To address the first challenge, one solution is to understand the process of pattern revelation theoretically and mathematically, and to design Monte-Carlo methods accordingly. To address the second challenge, since each Monte-Carlo realization is completely independent, one solution is to design parallel Monte-Carlo methods, and apply them within a parallel computing environment, e.g., cluster computing frameworks [51].

In addition to integrating the validation method within the application, another future goal for the GeoSocialApp is to convey more information with novel visual designs to improve the process of hypothesis generation. For example, in the radial graphical view, more information (e.g., the distance or GDP distribution within each cluster) could have been symbolized. For the map

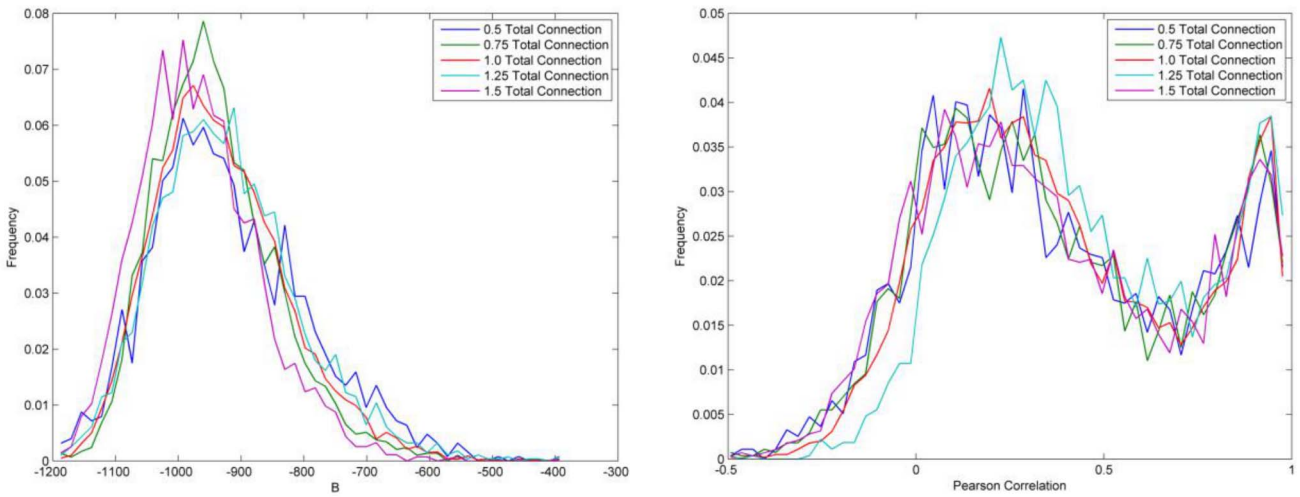


Figure 8. Validation results as a function of total connection numbers.
doi:10.1371/journal.pone.0088666.g008

view, one potentially useful addition might be a paired distance histogram (with 5–7 bins of short to long distance) that summarizes the distribution of between country distances for any selected cluster. In this way, more attribute information can be visualized on the map and network views to understand the interaction between geographical space and social network space.

Social network approaches have been widely applied to study the ITN, with a focus on the importance of network positions and relationships [52,53,54,55]. Fagiolo et al. [56] argue that the role of geographical proximity in shaping the structure of the ITN has not been explored, especially across geographical regions. To fill this gap, recent research integrates two important approaches in the study of global trade: social network analysis and the gravity model [28,57]. The researchers add network parameters into gravity models to represent the impact of the global trade network on bilateral trade, but those models are still not complex enough to consider both relationships across different geographical regions at varying levels. The hypothesis we developed through visual-computational exploration and then assessed through statistical validation can be considered as another effort toward future international trade models that consider more fully the complex geo-social interactions that occur across different geographical regions at varying levels. Our next step will extend our analysis to the temporal domain in order to understand how such geo-social patterns do change over a longer time period (e.g., from 1989 to 2009).

Given that Pearson correlation is sensitive to the sample size, the high correlation of 0.97 between geographic proximity weighted by population and the development level of countries should be interpreted with caution. However, the goal of this paper is not to produce the definitive analysis of the ITN but to demonstrate the value of applying a geovisual analytics approach as a method to account for both geographic and social network factors in complex processes. Application of the visual-computational methods was able to generate hypotheses about the interaction between level of economic development for countries and relative proximity of international trading partners and the statistical analysis (of which the Pearson correlation is a part) was used to provide support for the hypotheses. The positive relations are further validated statistically and robustly through application of a Monte-Carlo method. In future work, we will consider using a Wilcoxon rank sum test [58] and other similar non-parametric methods to complement the results from Pearson correlation for three reasons: Wilcoxon rank sum test works well even if sample size is small; Wilcoxon rank sum test conducts a formal statistical test and computes a p-value, which provides quantitative information in comparison with descriptive methods like Pearson correlation; non-parametric methods have fewer assumptions and are applicable to more general situations.

The combination of spatial and social network context supports exploration of the interaction between these components and consideration of their impact on outcomes of interest [7], but the combination has not received enough attention generally, not just with respect to the ITN. The GeoSocialApp provides generic frameworks to explore any analysis contexts that include spatial and social relationships among geographical regions (e.g., human migrants among different states in the U.S., war conflicts among different countries in the world, vector borne disease propagation, or the impact of social media on behavior in the world). To our knowledge, this is the first tool to allow users to explore the interconnections of spatial and social relationships at a geographical region level.

Supporting Information

File S1 A list of all countries with corresponding group IDs at the second level and third level of CONCOR.
(PDF)

Author Contributions

Conceived and designed the experiments: WL QD. Performed the experiments: WL PFY QD. Analyzed the data: WL QD. Contributed reagents/materials/analysis tools: PFY FH WL. Wrote the paper: WL AMM QD PFY FH.

References

1. Bailey TC, Gatrell AC (1995) Interactive spatial data analysis: Longman Scientific & Technical Essex.
2. Batty M (2003) Network geography: Relations, interactions, scaling and spatial processes in GIS. In: Unwin D, editor. Re-presenting GIS. Chichester, UK: John Wiley pp. 149–170.
3. Holland JH (1996) Hidden order: How adaptation builds complexity. Cambridge, MA: Perseus Books.
4. Batty M (2008) The size, scale, and shape of cities. Science 319: 769–771.
5. Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS ONE 6: e16939.
6. Mao L, Bian L (2010) Spatial-temporal transmission of influenza and its health risks in an urbanized area. Computers, Environment and Urban Systems 34: 204–215.
7. Adams j, Faust K, Lovasi GS (2012) Capturing context: Integrating spatial and social network analyses. Social networks 34: 1–5.
8. Andrienko N, Andrienko G (2012) Visual analytics of movement: An overview of methods, tools and procedures. Information Visualization: 3–24.
9. MacEachren AM, Jaiswal A, Robinson AC, Pezanowski S, Savelyev A, et al. (2011) Senseplace2: Geotwitter analytics support for situational awareness. IEEE. pp. 181–190.
10. Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, et al. (2010) Space, time and visual analytics. International Journal of Geographical Information Science 24: 1577–1600.
11. Guo D, Chen J, MacEachren A, Liao K (2006) A visualization system for space-time and multivariate patterns (vis-stamp). IEEE Transactions on Visualization and Computer Graphics 12: 1461–1474.
12. Luo W, MacEachren AM, Yin P, Hardisty F (2011) Spatial-Social Network Visualization for Exploratory Data Analysis; Chicago, Illinois. ACM.
13. Andrienko G, Andrienko N, Keim D, MacEachren AM, Wrobel S (2011) Challenging Problems of Geospatial Visual Analytics (editorial introduction). Journal of Visual Languages & Computing 22: 251–256.
14. Keim D, Kohlhammer J, Ellis G, Mansmann F (2011) Mastering the information age: solving problems with visual analytics. Goslar, Germany: Eurographics Association.
15. Andrienko N, Andrienko G (2010) Spatial generalisation and aggregation of massive movement data. IEEE Transactions on Visualization and Computer Graphics 17: 205–219.
16. Guo D (2009) Flow mapping and multivariate visualization of large spatial interaction data. Visualization and Computer Graphics, IEEE Transactions on 15: 1041–1048.
17. Andrienko G, Andrienko N, Rinzivillo S, Nanni M, Pedreschi D, et al. (2009) Interactive visual clustering of large collections of trajectories. IEEE. pp. 3–10.
18. Demšar U, Virrantaus K (2010) Space–time density of trajectories: exploring spatio-temporal patterns in movement data. International Journal of Geographical Information Science 24: 1527–1542.
19. Guo D, Liu S, Jin H (2010) A graph-based approach to vehicle trajectory analysis. Journal of Location Based Services 4: 183–199.
20. Wood J, Dykes J, Slingsby A (2010) Visualisation of origins, destinations and flows with OD maps. The Cartographic Journal 47: 117–129.
21. Larsen J, Axhausen KW, Urry J (2006) Geographies of social networks: meetings, travel and communications. Mobilities 1: 261–283.
22. Ahmed A, Fu X, Hong SH, Nguyen QH, Xu K (2010) Visual Analysis of History of World Cup: A Dynamic Network with Dynamic Hierarchy and Geographic Clustering. In: Huang ML, Nguyen QV, Zhang K, editors. Visual Information Communication: Springer. pp. 25–39.
23. Thiemann C (2011) SPaTo Visual Explorer. RoCS, Northwestern University.
24. Verdery AM, Entwisle B, Faust K, Rindfuss RR (2012) Social and spatial networks: Kinship distance and dwelling unit proximity in rural Thailand. Social networks 34: 112–127.
25. Mercken L, Snijders TA, Steglich C, de Vries H (2009) Dynamics of adolescent friendship networks and smoking behavior: Social network analyses in six European countries. Social Science & Medicine 69: 1506–1514.

26. Mao L, Bian L (2010) A Dynamic Network with Individual Mobility for Designing Vaccination Strategies. Transactions in GIS 14: 533–545.
27. Andris C, Halverson S, Hardisty F (2011) Predicting migration system dynamics with conditional and posterior probabilities; 2011; Fuzhou, China. IEEE. pp. 192–197.
28. Zhou M, Park C (2012) The cohesion effect of structural equivalence on global bilateral trade, 1948–2000. International Sociology 27: 502–523.
29. Andris C (2011) Metrics and methods for social distance: Massachusetts Institute of Technology. 189 p.
30. Luo W, MacEachren AM (2013) Geo-Social Visual Analytics. Journal of spatial information science: In press.
31. Cusumano-Towner M (2009) Exploring the Functional Landscapes of Gene Sets with Interactive Multidimensional Scaling; April 4–9; Boston, Massachusetts. ACM.
32. Hardisty F, Robinson A (2010) The geoviz toolkit: using component-oriented coordination methods for geographic visualization and analysis. International Journal of Geographical Information Science 25: 191–210.
33. Breiger R, Boorman S, Arabie P (1975) An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology 12: 328–383.
34. Wasserman S, Faust K (1994) Social network analysis: Methods and applications: Cambridge Univ Pr.
35. Smith DA, White DR (1991) Structure and dynamics of the global economy: Network analysis of international trade 1965–1980. Social Forces 70: 857–893.
36. Breiger R (1981) Structures of economic interdependence among nations. In: Blau P, Merto R, editors. Continuities in structural inquiry. New York: The Free Press. pp. 353–380.
37. Wallerstein IM (1974) The Modern World-System. New York: Academic Press.
38. Snyder D, Kick EL (1979) Structural position in the world system and economic growth, 1955–1970: A multiple-network analysis of transnational interactions. American Journal of Sociology: 1096–1126.
39. Nemeth RJ, Smith DA (1985) International trade and world-system structure: A multiple network analysis. Review (Fernand Braudel Center) 8: 517–560.
40. Jeong C, Pang A (1998) Reconfigurable disc trees for visualizing large hierarchical information space. pp. 19–25.
41. Carriere J, Kazman R (1995) Research report: Interacting with huge hierarchies: beyond cone trees. Proc IEEE Information Visualization' 95: 74–81.
42. Harrower M, Brewer CA (2003) ColorBrewer. org: an online tool for selecting colour schemes for maps. The Cartographic Journal 40: 27–37.
43. Barbieri K, Keshk OMG, Pollins B (2008) Correlates of war project trade data set codebook, Version 2.01. Online: http://correlatesofwar.org.
44. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks. I. Blockmodels of roles and positions. American Journal of Sociology: 730–780.
45. Cassi L, Morrison A, Ter Wal AL (2012) The Evolution of Trade and Scientific Collaboration Networks in the Global Wine Sector: A Longitudinal Study Using Network Analysis. Economic geography 88: 311–334.
46. Zhou M (2011) Intensification of geo-cultural homophily in global trade: Evidence from the gravity model. Social Science Research 40: 193–209.
47. Rodgers JL, Nicewander WA (1988) Thirteen Ways to Look at the Correlation Coefficient. The American Statistician 42: 59–66.
48. Gentle JE (2003) Random number generation and Monte Carlo methods: Springer.
49. Anderson JL (2012) Localization and sampling error correction in ensemble Kalman filter data assimilation. Monthly Weather Review 140: 2359–2371.
50. Clark R (2010) World-system mobility and economic growth, 1980–2000. Social Forces 88: 1123–1151.
51. Liu JS (2008) Monte Carlo strategies in scientific computing: Springer.
52. Shutters ST, Muneepeerakul R (2012) Agricultural trade networks and patterns of economic development. PLoS ONE 7: e39756.
53. Ercsey-Ravasz M, Toroczkai Z, Lakner Z, Baranyi J (2012) Complexity of the international agro-food trade network and its impact on food safety. PLoS ONE 7: e37810.
54. Kali R, Reyes J (2007) The architecture of globalization: a network approach to international economic integration. Journal of International Business Studies 38: 595–620.
55. De Benedictis L, Tajoli L (2011) The world trade network. The World Economy 34: 1417–1454.
56. Fagiolo G, Reyes J, Schiavo S (2009) World-trade web: Topological properties, dynamics, and evolution. Physical Review E 79: 0361151–03611519.
57. Fagiolo G (2010) The international-trade network: gravity equations and topological properties. Journal of Economic Interaction and Coordination 5: 1–25.
58. Rosner BA (2011) Fundamentals of biostatistics. Boston, MA: Brooks/Cole, Cengage Learning.

Designing a Web Service to Geo-Locate Subjects of Volunteered, Textual Geographic Information (#183)

R. Mullins [1], **F. Hardisty** [1], S. Pezanowski [1], S. Das [2], A. Savelyev [1], A. MacEachren [1], P. Mitra [3], A. Jaiswal [4]

[1]The Pennsylvania State University Department of Geography, 302 Walker Building, University Park, PA 16802, United States
[2]The Pennsylvania State University Department of Computer Science and Engineering, 313F Information Science and Technology Building, University Park, PA 16802, United States
[3]The Pennsylvania State University College of Information Science and Technology, 313F Information Science and Technology Building, University Park, PA 16802, United States
[4]Reunify, LLC. Record Linkage Research, 2043 Colorado Avenue, Suite #3, Santa Monica, CA 90404, United States

In recent years, the amount of publicly available spatial, or spatially-enable, data has grown tremendously, due in large part to the proliferation of GPS-enabled technologies in mobile devices and in-car navigation systems, and from the location information integrated into web applications, especially social networking services. Networks like Twitter, Four-Square, Facebook, and others allow users to provide insights into current events in real time via short form textual updates or statuses. Parallel to the availability of this type of citizen-produced data, there has been a growing interest in analyzing this data to examine sentiment or track how information disperses through networks. Many modern social networks provide a means to locate the contributor of status updates. The location of a contributor is typically given as geographic coordinates, latitude and longitude, that is accurate to the position provided by the web-enabled device used to submit the status update. This spatial information, along with the temporal information inherent to status updates, enables spatial and temporal analysis of contributor patterns. However, although some updates include information on the location of a contributor, little capacity is provided for geographically locating the subject, or subjects, that contributors are referencing.

In this paper we describe a web service, in development at the Pennsylvania State University, that enables the geolocation of people, places, and events described in common status updates from online social networks. We describe the use of techniques from a wide array of research areas – applied linguistics, natural language processing, search engine optimization, and geographic information science – to parse out people, places, and events explicitly or implicitly mentioned in status updates, and then analyze and contextualize these entities to locate them in geographic space. Finally, we outline how this service can be integrated into the development of dynamic, map-based, visual analytical interfaces, specifically in the context of crisis management and emergency response.

Keywords: geocoding, geospatial web services, natural language analysis, volunteered geographic information

Designing Map Symbols for Mobile Devices: Challenges, Best Practices, and the Utilization of Skeuomorphism

Joshua E. Stevens, Anthony C. Robinson, Alan M. MacEachren

GeoVISTA Center, Department of Geography, The Pennsylvania State University

Abstract. In this work we make three contributions to the design and use of map symbology on mobile devices. First, we present an overview of the current state of mobile symbology and best practices based on previous empirical findings. Second, we demonstrate the design of a new set of map symbols for mobile devices based on these guidelines and proposed design strategies. These new symbols were developed for comparison with an existing standard used by emergency management and disaster relief professionals. Lastly, we discuss the role of skeuomorphism in the context of affording interaction in map symbol design. We believe this work advances the science of mobile symbology and demonstrates a practical application of skeuomorphic design in modern mapping applications.

Keywords: Design, Symbology, Interaction, Skeuomorphism

1. Introduction

A number of recent technologies and tools have demonstrated the utility of mobile mapping devices in a wide range of settings. The potential utility of mobile maps is particularly clear in mission-critical environments such as law enforcement or emergency response and recovery. Within such settings, mobile devices must not only be reliable in both form and function, they must have utility to the mission. This requirement places particular emphasis on the visual interface of the device, as it connects the user to the application where information must be interpreted and acted upon quickly. With a growing number of location-based applications and support tools being

developed for these purposes, great care must be taken in the way events are symbolized.

In addition to mission-critical usage by professional users, mobile maps have permeated consumer environments and are ubiquitously employed to aid navigation, enhance tourism, and enable a wide array of location-based services that range from tracking a lost phone to local dating services (Lee, Zhu and Hu 2005). Whether for profession or pleasure, the success of many mobile maps relies on interactive functionality that transforms geographic data into actionable information.

At the same time, most maps are still largely static; base maps rarely need updating, key points of interest (POI) tend to reflect a constant location and information (e.g., a name and address), and many elements of cartographic design are typically not interactive (e.g., highway labels, POI). For these reason, we argue that an effective mobile map is one that successfully distinguishes interactive and non-interactive symbology.

This paper reviews research on mobile mapping with an emphasis on map symbology for and interaction with mobile maps. Based on the review, we elucidate some of the design challenges and considerations required for providing meaningful maps on a range of mobile devices reliant on interactive symbology. Emphasis is given to applications in emergency management and law enforcement where decisions must be made quickly and accurately, though we extend our findings to the broader range of mobile map uses. We conclude the paper with preliminary guidelines derived from the literature for design of mobile maps and then outline some key research and development challenges focused on leveraging technological advances to achieve effective mobile mapping applications.

1.1. Mobile Devices: Scope, Limitations, and Considerations

Mobile devices are considered here to be ones that are essentially portable and can be used remotely. These include laptop computers, smartphones, tablet computers, wearable devices, Personal Digital Assistants (PDAs), and in-vehicle units that include but are not limited to navigation systems. For the purpose of this paper, we focus on the subset of mobile devices that are held in or used by one's hands while mobile. This focus excludes laptop computers as well as wearable devices used primarily for augmented reality applications. Accordingly, the remainder of this article deals primarily with smartphones and tablet computers.

Mobile devices defined in this way have many potential benefits, including: collecting information in the field or during a police patrol (Bragdon et al. 2011), obtaining live updates remotely when in the field (Weakliam et al.

2005), communicating with natural/manmade disaster first responders (Erharuyi and Fairbairn 2003, MacEachren et al. 2006), and visual collaboration among large teams during urban emergencies (Monares et al. 2009).

Despite the advantages suggested above, mobile devices are not without their limitations. Most notable of these are reduced screen size and display resolution (Burigat and Chittaro 2011), reduced capability for input and interaction while the user is in motion (Bragdon et al. 2011), and limited processing power and memory (Follin and Bouju 2008). Each of these limitations influences how symbology can be applied and whether or not a particular approach is effective. These limitations can often be compounding. For example, limited screen space will dictate limits on the number and size of symbols that can be displayed legibly on the screen and the extent of territory that can be displayed. A desktop environment, in addition to supporting display of larger territories and more symbols can alleviate the constraints it does have through zooming, panning, and other interactions. However, the mobile utility of smaller devices is hamstrung by the increased effort similar interactive behaviors require when using the device in concurrence with another common task, like walking or driving. If this increase in effort to interact and the attention that must be directed to that interaction is too onerous, it can impede the device's advantages as a mobile aid (Willis et al. 2009).

A number of attempts have been made to understand and combat these limitations. The remainder of this paper will discuss several of these efforts that are most relevant to mobile mapping with an emphasis on issues of map symbology and interaction with the map display for mobile devices. Through this discussion, the latest directions in the literature and key challenges that remain for future research are also identified.

The existing literature on mobile map symbology is diverse and addresses many specific goals and interests. It is therefore useful to identify common themes that are prevalent. We have identified three themes that appear within the broader context of map symbology and interface design for mobile devices; specifically, the literature can be categorized as emphasizing:

1. **Symbology.** This work addresses use-specific symbol design, the salience of the symbols due to display parameters, figure-ground relationships, semantics, and the performance of symbols in task-driven evaluations.
2. **Interaction.** Research in this area emphasizes how users interact with devices, operate menus, allocate attention, and respond to displays that use symbols as part of a greater suite of functions. Examples include graphical hints that alert the user to symbols existing outside of the currently active view, symbols that aggregate when

they are too numerous for the screen, and the role of touch, gestures, and buttons to manipulate the display.

3. **Remote information access and collaboration.** Mobile devices are frequently used to report, update, collect, and communicate information remotely. A growing body of literature is centered on the use of map symbology and interface design to facilitate these collaborative tasks

There is considerable overlap between these themes and they are by no means absolute. We continue in the following sections with key lessons learned, emphasizing those most relevant to items 1 and 2: symbology and interaction.

2. Previous Research: Symbol Design and User Experience

Map symbols have been categorized in multiple ways and there is a wide range of terminology used to discuss map symbology in the literature. To create consistency across the studies that will be cited below, this paper will use the terminology described by MacEachren (1995), focusing on the relative abstractness or iconicity of symbols as it related to three categories of positional symbols that each exhibit a range on this continuum: pictorial, associative, and geometric.

Many mobile devices in current production utilize touch-screen displays, thus it is crucial that new symbol designs consider existing work in this area. Morrison and Forrest (1995) conducted one of the earliest studies evaluating pictorial symbols on touch-screen devices within the context of tourist maps. Their work highlights the need to consider design not only from the standpoint of variables affecting individual symbols (e.g., size and hue), but also semantic relationships across and between multiple symbols. For example, their results show that for many symbols, size does not influence the accuracy of visual search tasks but may greatly affect how quickly symbols are found. This relationship is moderated by the semantic context suggested by other symbols on the map. A telephone symbol for example may be interpreted as the location of a pay phone when used in isolation or in tourism maps, while the same symbol might be interpreted as a service for calling help if nearby symbols reflect first-aid and medical care. In other words, how users interpret a symbol is influenced as much by context and nearby symbols as by the symbol's own design.

In addition to suggesting semantic context, the design of nearby symbols can also influence the effectiveness of individual symbols by affecting the

overall salience of a particular symbol. Kuo-Chen refers to this as complexity contrast and it greatly influences the time required to identify symbols (2008). Related work suggests that associative symbols, such as the simple monochrome pictorial symbols that are typically part of standards-compliant recommendations, are not as strongly affected by changes in size as are realistic, multi-colored, sketch-based, or 3-dimensional pictorial symbols (Elias and Paelke 2008). Although this limitation is important to consider for map symbols in general, it is especially important for mobile devices with limited screen space. Such a limitation encourages the use of simple, abstract symbols over complex symbols with more detail and realism (Lee, Forlizzi and Hudson 2008).

While the interpretation of abstract symbols is less affected by changes in symbol size, abstract symbols (particularly geometric symbols) are subject to misidentification since the relationship between the symbol and what it refers to is often arbitrary. A trade-off therefore exists between accommodating a limited screen size through smaller, abstract symbols and maintaining semantic clarity.

Isolating symbology from the total experience of using a mobile device is at first an attractive idea since it removes complexities of human-computer interaction and a potentially limitless variety of application contexts. However, the appropriate design of symbols for mobile devices requires a complete understanding of how existing symbols are used, which symbol design traits and interactions between the devices and symbols improve or impede performance, and under which conditions or scenarios certain symbol types or designs are most useful. The factors influencing user experience are subject to varying abilities of individual users and an ever-changing range of device capabilities (Baus, Cheverst and Kray 2005, Meng 2005).

On par with the contentions of the previously mentioned work, Apple asks designers of their mobile applications to “embrace simplicity,” (2011, p. 152-153). In addition to generic advice, the human interface guidelines provided by Apple impose specific requirements for size and quality of icon designs to ensure that the designs are effective visually and tactically. The requirements derive from limitations on the user’s ability to see a symbol and the device’s ability to recognize the user’s fingertip when touched. Similar recommendations are provided by the interface guidelines from Google, which offers less specific advice but again reiterates the need to avoid complex, highly detailed, and realistic icons (2011). Additionally, both firms insist that designers consider icon design in the full context of the other elements of the interface and the purpose for which they will be used.

3. Design Recommendations and Suggested Guidelines

Together, the studies reviewed allow us to make the following suggestions:

1. Well-designed symbols should utilize black and white figure-ground relationships or be based on mixed colors that have established meanings, such as the U.S. interstate shield's use of red and blue, and have high contrast against the base map. Since the majority of vector base maps are by default light in color, this suggests that symbols for such maps should maximize figure-ground with a dark frame and light symbol.
2. Symbols with strong semantic relationships with their referent can be identified more quickly on smaller screens than those that are arbitrary and geometric, sketch-based, or based on a 3-dimensional rendition of the referent. However, increasing levels of abstraction may impede the accuracy of symbol identification, and even readily identifiable symbols can be misidentified if their purpose is ambiguous (e.g., the telephone symbol in Morrison and Forrest's research (1995)). Thus it is important to balance the *speed* with which users identify semantically strong pictorial symbols and the *accuracy* with which simpler abstract symbols can be located on the map.
3. Symbols should be smaller when displayed in large numbers - potentially removing the symbol frame if necessary as long as figure-ground is maintained and the symbols remain touchable (for symbols that require interactivity).
4. Symbols intended for concurrent use should be similar in complexity to avoid a large contrast gradient in symbol design, except in the case where greater salience is desired for particular symbols (e.g., a hospital symbol). High complexity contrast between regular symbols and those deemed important would then be preferable (e.g., a hospital symbol made with an "H" will stand out amongst other symbols made from simple geometric shapes, like squares and circles).

Three additional factors that should be taken into account by symbol designers are:

1. ***The capabilities of the target device(s).*** Touch-screen devices have additional symbol size requirements not present on devices that use other input methods. If the symbol requires tactile interaction, it must be large enough to be touched by a fingertip. This complicates the task of accommodating small screens and avoiding clutter.

2. ***The purpose of the device or application should influence the method of interaction***. Symbols used alongside other tasks that are cognitively engaging (like driving) should require as little interaction as possible. If interaction is required, it should be in the form of gestures, spoken commands, or make use of hard buttons that do not require the user to look at the screen.
3. ***Symbols that can be interacted with should be visually distinguishable from those that cannot***. How this is achieved will depend on the other variables employed in the symbol design. A bold frame may alert the user that a symbol is an aggregate and can be clicked for more information, however this approach would be less effective if symbol frames were employed for other purposes (e.g., event status or the degree of damage due to a disaster).

There is a complex relationship between each of these design decisions. It is unlikely that a symbol will be optimal for every setting, to every user, on every device. The application developer, cartographer, or designer must weigh the costs against the benefits and evaluate the performance of their symbology whenever possible.

3.1. Symbol Shape and the Map Symbol Box Model

In addition to the position of icon shape on the abstract-pictorial continuum, the shape of the frame surrounding the map icon is a vital consideration. The shape of the frame dictates how much space exists for the icon, and thus the actual size of the interpretable icon depends on the symbol frame's overall shape and border thickness (Figure 1, right). In general, symbols with square frames (or rounded rectangles as in Figure 1) provide more internal space than other frame shapes to maximize the icon size, which is beneficial in visual search tasks (Morrison and Forrest 1995). Moreover, the additional space afforded by maximizing the space around an icon can be utilized for other cues, like indicating interactivity as discussed in 3.2.

To aid in design decisions and the specification of design variables, we propose the *map symbol box model* (Figure 1, left). Similar to the box model specified for cascading style sheets (CSS) for HTML documents[1], the map symbol box model clarifies the foundational elements upon which a symbol is constructed. Defined in this way, individual design variables can be isolated and discussed explicitly, greatly enhancing the design process for new symbols and allowing precise critique of existing symbology.

---

[1] http://www.w3schools.com/css/css_boxmodel.asp

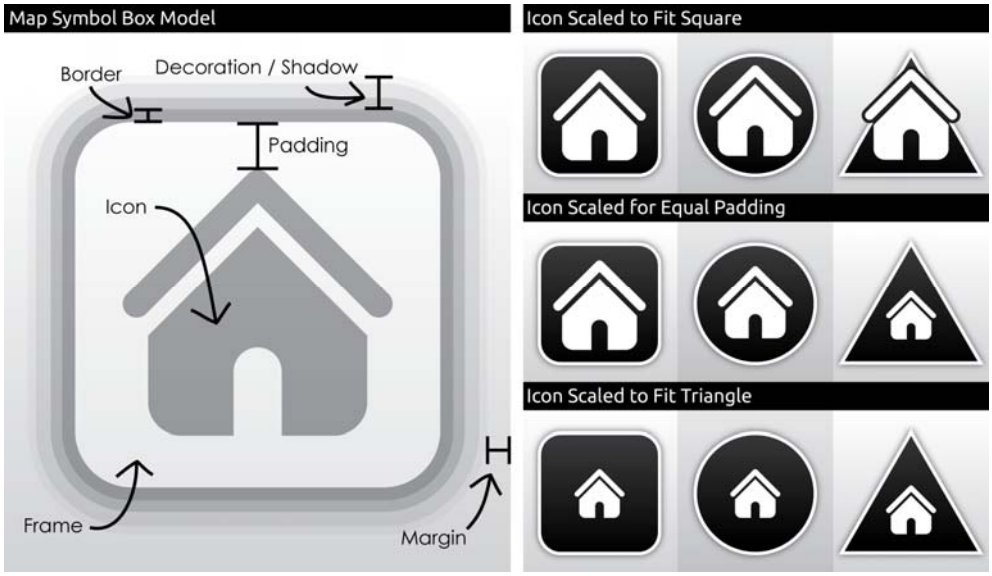


Figure 1. The proposed map symbol box model (left) and an example of specifying symbol size relationships using the padding element of this model (right).

By deconstructing map symbols into 'boxes within boxes,' each with a specific attribute, the design process can more accurately reflect the relationships between the various features of a complete map symbol.

3.2. Interactive Cues via Skeuomorphism

Skeuomorphs, typically defined as design elements that reflect ornamental references to previous (and potentially obsolete) analogs, are ubiquitous in current user interfaces. The uses of skeuomorphism in digital interfaces are typically aesthetic, such as realistic textures (leather and stitching) used in the design of mobile applications. Skeuomorphism can also be used to suggest semantic relationships. For example, the 'save file' feature in most software applications is represented by a 3.5" floppy disk, even though these disks are rarely used and new computers no longer come with the ability to read floppy disks. Despite a lack of academic literature on skeuomorphic interface designs and criticism for their form-over-function nature, we contend that when used sparingly, skeuomorphic designs have the potential to enhance the performance of map symbols.

We hypothesize that an effective use of skeuomorphism is to indicate that a map symbol is clickable (or touchable). By evoking a heuristic response akin to what we experience in the real world, map symbols – like elevator buttons and doorbells – can indicate interactivity by appearing to have a touchable surface different than their surroundings.

Figure 2 provides an example of a skeuomorphic interactive cue:

Figure 2. A subtle, 3-dimensional knurled skeuomorphic cue gives the interactive symbol (right) the appearance of being touchable.

4. GeoVISTA Symbology and Future Work

Within a project supported by the Department of Homeland Security, the Penn State GeoVISTA Center designed mobile symbology as a possible alternative to the existing Homeland Security Working Group (HSWG) symbology[2]. The GeoVISTA symbology (Figure 3) was designed in compliance with the guidelines presented in this paper. In follow up research, we plan to compare the symbol sets in a user study with several tasks completed on a mobile device.

Figure 3. The DHS HSWG symbology (left) and their redesigned GeoVISTA counterparts (right). The GeoVISTA symbols also have an interactive, skeuomorphic version using the cue in Figure 2.

Notable design decisions that distinguish the GeoVISTA symbology from the HSWG set are outlined below.

---

[2] http://www.fgdc.gov/HSWG/index.html

1. The GeoVISTA symbols are arranged in a square frame. This provides space within the symbol frame to increase the size, and therefore legibility, of the icon.
2. To have high figure-ground contrast with the greatest number of base maps, which are typically light in color, the GeoVISTA symbology uses a light icon within a dark frame.
3. The GeoVISTA symbology is intended to have a strong semantic relationship with the events or places being depicted.
4. The GeoVISTA symbology is designed as an entire set to have consistent visual complexity from one symbol to the next.
5. To further promote figure-ground relationships, the GeoVISTA symbology has a slight shadow and raised appearance, which helps separate the symbols from the base map.

Symbols intended to be interactive with feature a skeuomorphic cue that distinguishes them from non- interactive counterparts.

The user study evaluation will be based on four tasks that cover visual search, semantic relationships, interactivity, and preference. Results are forthcoming.

References

Apple Inc. 2011. iOS Human Interface Guidelines: User Experience. 1-170.

Baus, J., K. Cheverst & C. Kray. 2005. A Survey of Map-based Mobile Guides. In *Map-based Mobile Services: Theories, Methods and Implementations,* eds. L. Meng, A. Zipf & T. Reichenbacher, 193-209. Berlin: Springer.

Bragdon, A., E. Nelson, Y. Li & K. Hinckley. 2011. Experimental analysis of touch-screen gesture designs in mobile environments. In *Proceedings of the 2011 annual conference on Human factors in computing systems,* 403-412. Vancouver, BC, Canada: ACM.

Burigat, S. & L. Chittaro (2011) Visualizing references to off-screen content on mobile devices: A comparison of Arrows, Wedge, and Overview + Detail. *Interacting with Computers,* 23**,** 156-166.

Elias, B. & V. Paelke. 2008. User-Centered Design of Landmark Visualizations. In *Map-based Mobile Services,* eds. L. Meng, A. Zipf & S. Winter, 33-56. Springer Berlin Heidelberg.

Erharuyi, N. & D. Fairbairn (2003) Mobile geographic information handling technologies to support disaster management. *Geography,* 88**,** 312-318.

Follin, J.-M. & A. Bouju. 2008. An Incremental Strategy for Fast Transmission of Multi-Resolution Data in a Mobile System. In *Map-based Mobile Services,* eds. L. Meng, A. Zipf & S. Winter, 57-79. Springer Berlin Heidelberg.

Google Inc. (2011) User Interface Guidelines. http://developer.android.com/guide/practices/ui_guidelines/index.html (last accessed Nov. 4, 2011).

Kuo-Chen, H. (2008) Effects of computer icons and figure/background area ratios and color combinations on visual search performance on an LCD monitor. *Displays,* 29**,** 237-242.

Lee, D. L., M. Zhu & H. Hu (2005) When location-based services meet databases. *Mobile Information Systems,* 1**,** 81-90.

Lee, J., J. Forlizzi & S. E. Hudson (2008) Iterative design of MOVE: A situationally appropriate vehicle navigation system. *International Journal of Human-Computer Studies,* 66**,** 198-215.

MacEachren, A. M. 1995. *How maps work : representation, visualization, and design*. New York :: Guilford Press.

MacEachren, A. M., G. Cai, M. McNeese, R. Sharma & S. Fuhrmann. 2006. GeoCollaborative Crisis Management: Designing Technologies to Meet Real-World Needs. In *7th Annual National Conference on Digital Government Research: Integrating Information Technology and Social Science Research for Effective Government*, 71-72. San Diego, CA.

Meng, L. (2005) Egocentric Design of Map-Based Mobile Services. *Cartographic Journal,* 42**,** 5-13.

Monares, A., S. F. Ochoa, J. A. Pino, V. Herskovic & A. Neyem. 2009. MobileMap: A collaborative application to support emergency situations in urban areas. In *Computer Supported Cooperative Work in Design, 2009. CSCWD 2009. 13th International Conference on*, 432-437.

Morrison, C. & D. Forrest (1995) A study of point symbol design for computer based large scale tourist mapping. *The Cartographic Journal,* 32**,** 126-136.

Weakliam, J., D. Lynch, J. Doyle, M. Bertolotto & D. Wilson. 2005. Delivering Personalized Context-Aware Spatial Information to Mobile Devices. In *Proceedings, 5th International Workshop, W2GIS 20052005*, 194-205. Lausanne, Switzerland.

Willis, K. S., C. Hölscher, G. Wilbertz & C. Li (2009) A comparison of spatial knowledge acquisition with maps and mobile maps. *Computers, Environment and Urban Systems,* 33**,** 100-110.

JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 8 (2014), pp. 27–66
doi:10.5311/JOSIS.2014.8.139

REVIEW ARTICLE

Geo-social visual analytics

Wei Luo and Alan M. MacEachren

GeoVISTA Center, Department of Geography, Pennsylvania State University, PA, USA

Received: March 28, 2013; returned: May 25, 2013; revised: July 9, 2013; accepted: August 21, 2013.

Abstract: Spatial analysis and social network analysis typically consider social processes in their own specific contexts, either geographical or network space. Both approaches demonstrate strong conceptual overlaps. For example, actors close to each other tend to have greater similarity than those far apart; this phenomenon has different labels in geography (spatial autocorrelation) and in network science (homophily). In spite of those conceptual and observed overlaps, the integration of geography and social network context has not received the attention needed in order to develop a comprehensive understanding of their interaction or their impact on outcomes of interest, such as population health behaviors, information dissemination, or human behavior in a crisis. In order to address this gap, this paper discusses the integration of geographic with social network perspectives applied to understanding social processes in place from two levels: the theoretical level and the methodological level. At the theoretical level, this paper argues that the concepts of nearness and relationship in terms of a possible extension of the First Law of Geography are a matter of both geographical and social network distance, relationship, and interaction. At the methodological level, the integration of geography and social network contexts are framed within a new interdisciplinary field: visual analytics, in which three major application-oriented subfields (data exploration, decision-making, and predictive analysis) are used to organize discussion. In each subfield, this paper presents a theoretical framework first, and then reviews what has been achieved regarding geo-social visual analytics in order to identify potential future research.

Keywords: geography, social network, visual analytics, First Law of Geography, data exploration, decision-making, predictive analysis

1 Introduction

Modern society has become an increasingly interconnected world of techno-social systems embedded with dynamic multi-scale networks (e.g., the internet, transportation). The complex interactions within and among these networks always have geographical constraints,

whereas they also change or reshape the traditional notion of geographical effects (e.g., distance) [115]. For example, small social groups are usually geographically cohesive (with the geographic span of group members a function of group size); but large social groups are less cohesive and less likely to exhibit spatial clusters [147]. Furthermore, because of advances in communication and transportation technologies over the past decade, there is a shift from networks that are both geographically and socially close (e.g., physical communities) to networks that are socially near but geographically dispersed (e.g., on-line communities) [185, 195]. To effectively understand the complex interaction between space and techno-social networks, it is necessary to encourage interdisciplinary understanding (e.g., geography, network science) through integrating current theories and methods (e.g., spatial thinking, social network theory) and to develop new theories and methods.

Recent research in physics emphasizes the power of networks in which space becomes a background to visualize and understand network analysis results [175]. Research in geography encourages the integration of spatial thinking into traditional social science through the concepts of space, place, and time [79, 80], but often treats networks in a simplistic way. This paper argues that space and social networks should be considered simultaneously when framing research on human activity. We contend that this perspective has not received enough attention and provide examples throughout the paper to support this contention. We further contend and present evidence that the new multidisciplinary research field of visual analytics provides an approach and methods that are well-suited to understanding the interaction of geographical and social network contexts. Visual analytics is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [176, p. 4]. A core objective of research in visual analytics is to provide a framework for integration of computational analytical methods with visual interfaces to both the information of interest and the computational methods that enable human analysts to cope with large, complex, and heterogeneous data sources and complex questions that these data sources make it possible to address.

Understanding large and complex techno-social networks and their interaction with space at geographic scales requires advances in computational methods. However, computational methods alone have limits and biases because of the predefined structures they have, which greatly limit their analytical power. The process and results of any computational techniques have limited value without input from human analysts to select appropriate methods, to set parameters, to interpret results, to understand what to do next, and to draw conclusions [11]. Visualization of data and computational processing gives users an intuitive representation, greatly promoting application of human perceptual and cognitive information processing capabilities. A simple combination of visualization with computational analysis, however, is not sufficient. Thus, the goal for visual analytics is to integrate human and computational reasoning in more fundamental ways, bringing the experts' background knowledge, creativity, and intuition into the analysis process through an interactive visual environment, in order to combine the strengths of humans and computers to enable an insight gaining process [105].

Visual analytics provides a potential conceptual approach and set of tools to integrate the geographic and social network contexts of human processes, but the application of visual analytics to this challenge remains relatively underdeveloped in the literature. Thus, the goal of this paper is to provide a base from which to develop and apply visual analytics methods to examine the interaction of both contexts and enable understanding of human processes. To achieve this goal, we address four objectives: (1) to present a theo-

retical framework in which geography and social network contexts can be combined and through which visual analytics methods can be developed and applied (Section 2), (2) to introduce "geo-social visual analytics" as an integrated analytical approach grounded in the conceptual framework (Section 3), (3) to review what has been achieved in relation to the integration of geographic and social network contexts in relation to three core tasks for geo-social visual analytics: data exploration, decision-making, and predictive analysis (also Section 3), and (4) to use the review of the current state of the art to identify potential future research challenges for advancing their integration (Section 4). From here on, we will refer to the interaction of geographical and social relationships as geo-social relationships, and the interaction between both in terms of visual analytics as geo-social visual analytics. The key distinction between geo-social visual analytics and prior work in *geo-visual analytics* [10] is the explicit integration of social network perspectives and methods into the approach and tools.

2 Geo-social relationships at a conceptual level

The theoretical framework we propose as a base upon which to develop and apply visual analytics methods that enable understanding of human processes draws upon a wide range of perspectives from human geography and social sciences more generally. One theme that connects the perspectives is that *social processes take place within particular contexts*. The two focused on here are spatial and social network contexts, each of which has been addressed in the past with specific methods and perspectives. Spatial analysis in social science is used to identify geographical patterns that result from social processes and to understand how space affects such processes. Most spatial analysis is based on an explicit or implicit assumption of the First Law of Geography [180, p. 236]: "Everything is related to everything else, but near things are more related than distant things." Social network analysis is used to understand how relationships among actors (i.e., individuals, groups, or other social collectives) within a network affect or are affected by social processes [187]. Social network analysis has an assumption, complementary to the First Law cited above, that actors with similar relations may have similar attributes/behaviors. Spatial analysis and social network analysis consider social phenomena in their own specific contexts, either geographical space or network space, but as we outline below, considering both contexts together when they contribute simultaneously has the potential to achieve new insights about human processes.

Both contexts demonstrate strong conceptual overlaps. Hess [92] proposes a geographically informed theoretical framework to understand the behaviors of social actors in geography through integrating territorial embeddedness, network embeddedness, and societal embeddedness (Figure 1). This framework is used to study economic actions from a critical human geography perspective, but this paper aims to extend it into a generic framework to study geo-social relationships. The concept of embeddedness has been prominently used by geographers to understand the behaviors of social actors in specific contexts [151]. Societal embeddedness refers to societal (i.e., cultural, political, etc.) background from which actors come, in which actions of actors are influenced, and to which actors contribute. Network embeddedness refers to the importance of relational aspects (i.e., social relations, cultural relations) among social actors to shape the actors' behaviors and of actors' behaviors to change relations. Territorial embeddedness refers to the specific places in which

the actors behave: how the places influence actors' behaviors and attributes; how the actors' behaviors change the territory. The overlap area in Figure 1 between the territorial embeddedness and the societal embeddedness fits the First Law of Geography [180]. The overlap area between the network embeddedness and the societal embeddedness fits the homophily principle in social network analysis theory: similarity breeds connection [142]. The overlap area between the territorial embeddedness and the network embeddedness fits a common phenomenon: how space constrains the development of networks and how networks reshape the space. The three overlap areas in Figure 1 indicate that they are not mutually exclusive, but interact with each other; the emphasis is on the interaction between geographic and social network context and the impact of the interaction on outcomes of interest. The overlap area suggests a possible extension of the First Law of Geography: *Everything is related to everything else, but near things are more related than distant things. Nearness can be considered a matter of geographical and social network distance* [73].

In addition to distance, the other two important concepts implied in the First Law of Geography are relationship and interaction: "everything is related to everything else." Flint [72, p. 33] argues that "the nature of a place is the combination of both locations and their connections to the rest of the world." Prager [149] also argues that geographical locations would be unrelated without relationships and interactions, whereas such relationships and interactions would be meaningless without the context geographical locations provide. Some of those social relationships may be constrained within the place, whereas others may stretch out to link geographical locations to wider relations and processes [139]. There is an increasing understanding of the importance of combining geographical space and networks from different sub-disciplines within geography, such as political geography [122], economic geography [166], and geographical information science (GIS) [29]. Staeheli [169, p. 160] argues that spaces become "social locations" embedded in "webs of cultural, social, economic, and political relationships." Ashdown [16] even argues that the political power is now shifting from a dominance of western culture to a collective governance at the global scale, because we have come into a new interlocked age (i.e., of complex, interconnected networks) that causes our destinies to be shared with our enemies.

Taking a perspective that complements those cited above, Massey [140] explores relationships between identity of place (and group and individual) and responsibility of place (and group and individual) and the geographical components of each. From the perspective of our paper, an important component of Massey's overall conceptual argument (the details of which are well beyond our focus here) is that connections to other locations are an essential component to understanding both the identity and responsibilities of places (and the groups and individuals connected to those places). By linking individual location to group location to place, the science of social networks can explain key aspects of how observed spatial-social patterns evolve. Therefore, to know the spaces, we must understand how spaces are connected geographically and socially. This contention may be increasingly true in today's digitally hyper-connected world. We are now living life in even more complex and interrelated networks: we check our e-mail, make a phone call, take transportation, or update our status in Facebook [119]. The lifestyle causes the emergence of the new human geo-social relationships: people with strong emotional ties may live geographically far away [118]. Christakis and Fowler [51] make a list of certain rules regarding networks: people shape their network, networks shape people, their friends affect them, their friends' friends' friends affect them, and a network has its own life. The inherent network structures in our lives affect our ideas and behaviors (i.e., emotional, sex-

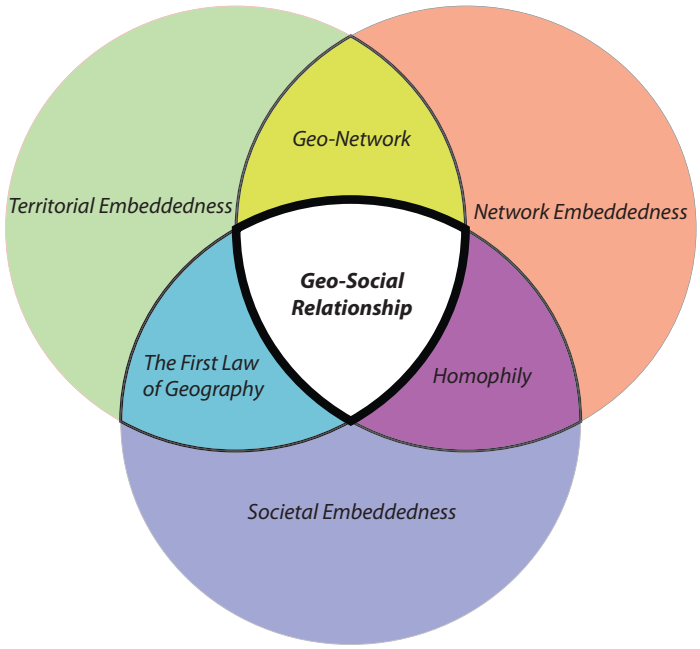


Figure 1: Proposed conceptual framework for geo-social relationships based on fundamental categories of embeddedness [92]. The framework consists of three kinds of embeddedness: territorial, network and societal. The paired overlaps each match with specific perspectives, as indicated and the joint overlap of all three areas suggests the extension of the First Law of Geography.

ual, and health-related), and the interaction of such individual-level behaviors develops macrosocial phenomena observed in a spatiotemporal framework.

Massey, in a report edited by Urry, et al. [186] goes on to argue that considering things from a "social" perspective is to use "how we are going to live together" as the motivating question for all social, political, and ethical questions framed in spatiotemporal frameworks to understand the world. This social perspective aims to develop general social theory to address today's highly interconnected society. Network science [41] is regarded by many authors as the approach to study such phenomena [192,193]. For example, current research projects from the Santa Fe Institute focus on studying the underlying principles and mathematical relationships concerning the evolution of human society and modern human social organizations (i.e., cities) with a network approach [30,31]. A growing literature in network science explores how everything is related to everything else [23,100,171]. For example, the famous small world experiment [144] formalizes the notion that each person only has six degrees of separation from anyone else on earth [108]. Sui [172] further argues that Tobler's First Law of Geography is a big idea for a small world, because it only takes a few steps to turn a large world into a small world. Based on all of the discussions, we argue that geo-social relationships in terms of the First Law of Geography should be extended into the notion: *everything/everyone is related to everything/everyone else, but near things/ones*

are more related than distant things/ones; Nearness and relationship can be considered a matter of geographical and social network distance, relationship, and interaction.

In addition to the exploration of geographical and social relationships and interactions, the conceptual framework (Figure 1) allows such relationships to be put into particular societal contexts (i.e., political, economic, cultural). On the one hand, the premise "near things are more related than distant things" in the First Law of Geography implies that certain local factors and circumstances can make geo-socially close areas different from geo-socially distant areas. For example, spatial proximity together with connections that link spatially heterogeneous groups in the population (i.e., city-wide travel of select individuals) are two major factors that determine the spatial layout and the temporal sequence of disease transmission [138]. And, recent research by Onnela, et al. [147] suggests that the size at which spatial cohesion of groups breaks down (about 30 members) coincides with the optimal group size for cooperation in social dilemma situations. On the other hand, the premise "everything is related to everything else" indicates that there are multiple factors that make contributions to patterns and connections among areas. For example, geographical homophily has a major impact on international trade among developing countries, whereas political and cultural homophily matters the most for bilateral trade between developed and developing countries [101, 206]. The societal context provides the framework to explore different factors (i.e., political, economic, cultural) behind observed geo-social patterns, and how such geo-social patterns interact with those factors to generate new geo-social patterns.

In work that complements the discussion above of geo-social integrations at the conceptual level, Adams, Faust, and Lovasi [2] identify five conceptual strategies for the integration based on current geo-social relationship research: (1) spatial impacts on the development of social networks over varying spatial scales, such as offices [159, 197], communities [55, 70], and so on; (2) the impact of social network on the places people select to inhabit [188]; (3) the use of peer network structures to determine neighborhood boundaries [93]; (4) the interactive impacts between spatial and social relationships [125, 150, 164]; and (5) multiple context impacts on outcomes related to social, health, and other processes [58, 143, 173]. The five conceptually geo-social integrations also fit the conceptual framework (Figure 1). The first integration focuses more on spatial constraints; the second emphasizes the network effects on residence selection; the third and fourth stress the interaction between geographical and network relationships; and the last highlights the multiple context impacts on outcomes.

The proposed conceptual framework comes from a critical human geography perspective. Here, we adapt it to inform the application of visual analytics to the study of geosocial processes. Previous research already provides examples of integrating critical human geography perspectives into visualization, e.g., in feminist visualization [113] and grounded visualization [107]. Knigge and Cope [107] present six connections between grounded theory and visualization, which are adopted here to integrate our proposed geo-social relationship conceptual framework and visual analytics at the theoretical level. First, the proposed conceptual framework and visual analytics are exploratory approaches which involve iterative explorations with matching mental models to construct knowledge. Second, both are iterative approaches which involve recursive processes of data collection, visualization, and analysis with critical thinking at each step. Third, both methods are enriched through connecting real-world phenomena and human experiences to broader processes. Fourth, both methods facilitate multiple interpretations and representations because of the notion

that there is no single correct way to interpret and visualize data [168]. Fifth, both methods acknowledge the importance of situating knowledge construction into the historical, geographical, and cultural context. Lastly, both methods recognize the existence of uncertainty in the data.

3 Geo-social visual analytics at a methodological level

Building upon the introduction to geo-social relationships at a conceptual level, this section discusses how to put the geo-social relationships into practice. From an application perspective, visual analytics methods can be classified into three groups focused on support for: data exploration, decision-making, and predictive analysis. This classification is used here to organize the methodological level in terms of geo-social visual analytics. Visual analytics aims to enable the human reasoning process, so it must build on an understanding of that reasoning process [177]. For the above three categories in visual analytics, this paper presents reasoning frameworks related to geo-social perspectives for each group first, and then discusses corresponding geo-social visual analytics technologies in order to identify potential future research directions.

3.1 Data exploration

3.1.1 Conceptual framework

Data exploration is a primary task in visual analytics to make sense of overwhelming amounts of disparate, conflicting, and dynamic data in a novel manner. Here, we use the Feature-ID model for geovisualization (Figure 2) [128], an extension of an earlier pattern-matching model of cartographic visualization [130], as the reasoning framework for data exploration in terms of geo-social visual analytics. The pattern-matching model, in turn, draws upon a general scientific visualization perspective based on human cognition [76] to support understanding of human-display interaction in the context of map-based geovisualization.

The iterative process between human and computer interaction through "seeing," "interpreting," and "constructing-knowledge" shown in Figure 2 is an insight-gaining process, which is a primary goal for visual analytics. Yi et al. [205] characterize insight gaining as a multi-step process: provide overview, adjust, detect pattern, and match mental model. A reasonable insight gaining process starts with seeing an *overview* of a domain area. Having a big picture may or may not lead to direct insight, but it is a good starting point for people to make additional inquires about areas of interest to gain more knowledge. *Adjust* refers to a process through which people adjust the level of abstraction to explore a data subset of interest in order to make more sense of the data. *Detect pattern* refers to identifying interesting results that can include specific distributions, anomalies, clusters, and trends in the datasets. *Match mental model* (equivalent to "instantiate schema" in Figure 2) refers to reducing human cognitive load and amplifying recognition through providing a visual representation of data to decrease the gap between the data and user's mental model of it, as well as the gap between the visual representation and real-world knowledge. The whole process of visual interactive exploration is a mental model building process from knowledge development to critical breakthrough [45] that involves the iterative interaction

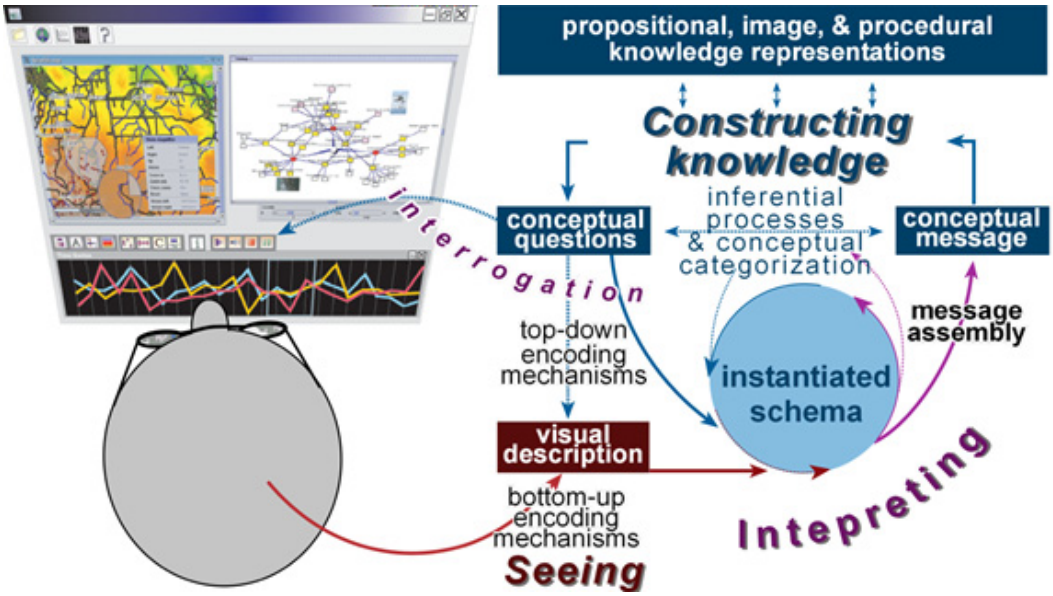


Figure 2: Feature-ID model of geovisualization; elaboration of ideas first presented in Figure 8.1 in [128]. The model focused on identifying key components in visually-enabled reasoning about geographic phenomena and relationships.

among "seeing," "interpreting," and "constructing-knowledge" at each step. Therefore, geovisualization works as a larger cognitive system to support a human reasoning process.

The framework provides a general model for understanding the visual and cognitive processes involved in interpreting and reasoning with geographic representations to address place-based questions. Extending the framework to geo-social analysis will require integrating: what has been learned about human perception and cognition of spatializations [65, 66], an understanding of how geographic scale social processes are conceptualized, and an understanding of the more complex reasoning process required by analysts attempting to understand the potentially complex relationships in geo-social processes. More generally, geo-social visual analytics aims to understand the interaction between two contexts and the impact of multiple contexts on reasoning outcomes through integrating social network space into geographic space to provide users a broader perspective. The following section illustrates the perspective through examples from existing research that each offers a step toward the objective of geo-social visual analytics to enable data exploration.

3.1.2 Geo-social visual analytics in data exploration

Network theories and representations have not been fully considered in geographical information science [149], but they have great potential to offer insight into complex geographical phenomena in terms of geo-social interactions. This section aims to address this gap. Given complex relationships between geographical space and social space at different spatial scales, this paper characterizes the relationships into two groups: (1) geo-social

relationships among geographical areas (e.g., nation, state, county), and (2) geo-social relationships among individuals at discrete locations (e.g., locations of mobile phone use, individual household locations, etc.).

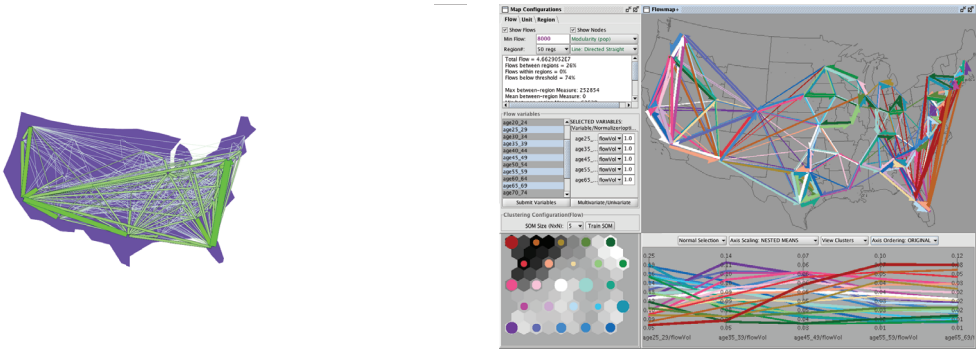
Geo-social relationships among geographical areas Geo-social relationships among geographical areas cover a wide range of topics: such as migration flows at the city scale [152], state scale [148, 182], or country scale [90]; transportation flows [13, 57, 89]; international trade among countries [68, 69]; sports competition among countries [4]; and so on. In an early example shown in Figure 3(a), Tobler [181] uses a network representation to describe the migration among different states in the U.S. In work grounded in geovisualization and geovisual analytics, Guo [86] proposes an integrated interactive visualization framework that is used to effectively discover and visualize major flow patterns and multivariate relations from the county-to-county migration data in the U.S (Figure 3(b)). In complementary recent research, Wood et al. [199] propose an origins and destinations (OD) map to preserve all origin and destination locations of the spatial layout through constructing a gridded two-level spatial treemap.

The above work assumes that geographic location defines the spatial-social process with explicitly spatial representation and implicitly network representation, but the assumption only holds partially true for the modern interconnected world. Such representations reflect a situation in which the current integration of network analysis in GIS only focuses on a mathematical perspective that emphasizes graph theory and topology components of networks [54]. Such representations do not allow users to explore the relationship between geographical space and social space, because they ignore network theory behind the network representation. Miller [145] also suggests that geographic phenomena with strong social components (e.g., infectious disease propagation [52]) do not appear to follow a Euclidean metric. Thus, geographical proximity does not necessarily mean social closeness and conversely, geographical long distance does not necessarily result in social isolation. From a cognitive perspective, explicitly spatial representation and implicitly network representation can mislead human intuition about social relationships among actors and their relationships to place. Thus, it is necessary to involve explicit network representations to consider the importance of social position, social distance, and social space.

Andris [14] proposes five benefits to involving an explicit network representation within a geographical environment:

(1) network community structure methods can identify clusters to understand the group of interconnected places as a unit rather than as dense collocations;
(2) node measures (i.e., degree, betweenness) can show the power of places;
(3) network system measures like degree distribution, closeness distribution, and clustering coefficients can indicate the role of any connected geographic region over the whole system;
(4) multiple social flow layers can be added simultaneously like spatial overlay functions in a geographical information system (GIS) to better evaluate interaction between places; and
(5) explicit network representation performs better to model the case in which spatial closeness does not correspond to stronger social flows between places.

Given the benefits of explicit network representations in a geographical environment, in [127] the authors developed a spatial-social network visualization tool, the *GeoSocialApp*,



a. b.

Figure 3: (a) State to state migration from 1995 to 2000. Adapted from Figure 7 in [181] with Flow Mapper program from CSISS.org. (b): Flow mapping and multivariate visualization of large spatial interaction data. Figure courtesy of Diansheng Guo. For additional information on the research it is derived from, see [86].

that supports network, geographical, and attribute spaces in this way to allow the exploration of spatial-social networks among them (Figure 4). The GeoSocialApp is a preliminary tool based on an early, less comprehensive version of the conceptual framework for geo-social relationships discussed in this paper. In the GeoSocialApp, different interactive and linked views for network, geographical, and attribute spaces, respectively, correspond to network embeddedness, territorial embeddedness, and societal embeddedness (see Figure 1). However, the GeoSocialApp is only a first step toward instantiating our overall framework and its present capabilities connect the three components without fully integrating them and without a full range of spatial and network analysis methods that will be required to support many of the ideas presented in this paper. That said, a brief description of the GeoSocialApp will serve to illustrate initial strategies for implementing the overall framework.

With explicit network spaces (in a dendrogram view and node-link view), GeoSocialApp users can have an intuitive understanding of social position, social distance, and social groups directly. For example, with the international trade network among 192 countries in 2005 as a case study, two groups identified through the dendrogram view show a core-periphery structure in the node-link view in which the red nodes are in the core and the yellow nodes are in the periphery. Since each node represents one country in the map view, the results in the map also show that the countries in the world have a hierarchical structure in which red nodes in the node-link view are economic core countries without highlight and yellow nodes are economic periphery countries with highlight. The parallel coordinate plot allows users to explore the power of places based on the network measures for each country. Core countries that have a low clustering coefficient (a measure of degree to which nodes in a network tend to cluster together) have high values with the other four variables (in-degree, out-degree, closeness, and Eigenvector). The negative relationship implies that rich countries may benefit more from more diversified trade partners and small economies may benefit more from concentrated trade partners.

In complementary work, Thiemann [174] develops SPaTo Visual Explorer to allow the exploration of spatial-social networks with multiple spatial and network representations.

Unlike geographical distance measures, a new shortest-path distance based on node centrality measures is implemented into SPaTo Visual Explorer [200]. This tool can easily identify the shortest social distance among different cities based on the worldwide air-transportation network. While the SPaTo Visual Explorer implements analytical methods that are relevant to the overall geo-social visual analytics perspective we present here, the authors do not ground the work explicitly in any conceptual framework for integration of spatial with network contexts. Their related work (cited above) on the nature of borders in "spatially embedded multi-scale interaction networks" [175], however, illustrates how multiscale human mobility can be understood through what we would categorize as geo-social visual analytics methods (although the authors do not use those terms). Although Theimann, et al. [175] make some use of network statistics in their analysis, we would position the analysis and its interpretation in the overlap of territorial and network embeddedness in relation to the conceptual framework outlined in Section 2.

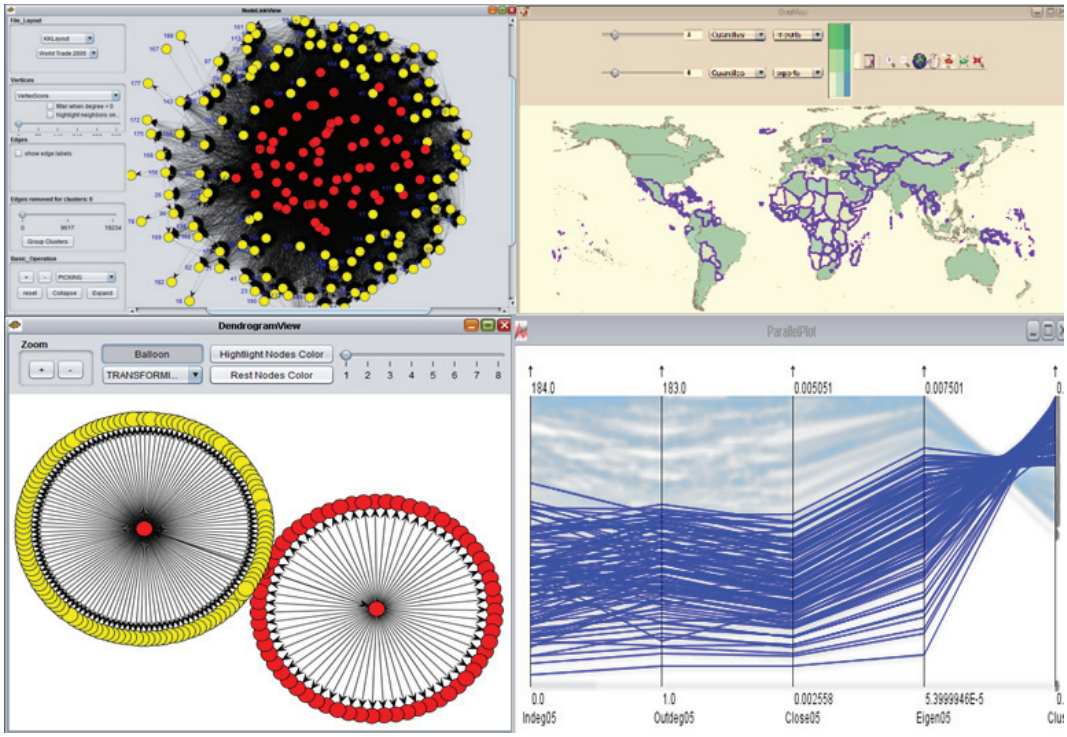


Figure 4: The analysis of international trade data among 192 countries in 2005 with GeoSocialApp: node-link (upper left), dendrogram (lower left), bivariate choropleth map (upper right) and parallel coordinate plot (lower right). For a preliminary view of this tool, see Luo, et al. [127]. Each node in the node-link view and the dendrogram view corresponds to one country in the bivariate choropleth map; the parallel coordinate plot view is used to explore network attributes for countries (from left to right, the attributes depicted are: indegree, outdegree, closeness, eigenvector, and clustering coefficient). Red dots in the node-link and dendrogram view are periphery countries and these are highlighted on the map and parallel coordinate plot with blue outlines.

Geo-social relationships among individual locations Geo-social relationships among individual locations include ones reflected in road networks [19, 56], commuting behaviors [112], location-based social networks [60], and social media networks [204]. Kwan and Lee [116] apply geovisualization methods to explore activity patterns of women in Lexington, Kentucky as they relate to the transportation network. Specifically, they use GPS data collected through a travel data collection test to explore trips by women without children under 16 years of age. Their analysis used 3D GIS-generated space-time path depictions to help find that trips by this subset of women mainly use highways and major arterials, thus illustrating that individuals within a particular demographic group can share activity patterns that use geographic networks in similar ways. In subsequent work, Kwan [114] applies the concept of human extensibility (the ability of individuals to utilize space-adjusting technologies, including transportation and communication, to overcome the friction of distance) to more deeply explore the complex daily interactions of individuals grounded in interlinked geographic and social networks across scales from local to global. In complementary work in the geovisual analytics domain, Shen and Ma [165] create MobiVis which allows visual analytics of social and spatial information in a human interaction network over time, and they illustrate how easily this tool supports comparison of individual and group behavior patterns (using the MIT Reality Mining Dataset [61]). Those studies, and most similar research studies, illustrate that geovisualization and geovisual analytics can reveal distinctive patterns of spatial and social behaviors of different human interaction groups in a straightforward way [49, 99, 120, 154, 167].

Complementary to the visualization advances outlined above, computational methods to explore spatial-social human interactions focus on developing quantitative representations of human movements. For example, with mobile phone data, Gonzalez et al. [78] and Rhee et al. [155] find that human trajectories are characterized by a regular, time independent characteristic length scale and are more attracted to more popular places, like home or work. With the circulation of bank notes in the United States, Brockmann et al. [39] find that the traveling distances of human mobility decay as a power law, and that the distribution of the time people stay in one small, spatially confined region follows algebraically long tails. Chaintreau et al. [44] and Karagiannis et al. [102] observe that inter-contact time between mobile devices shows an approximate power law in the range of 10 minutes to 1 day. Overall, all of the above studies suggest the existence of scale-free characteristics observed in most networks in which a small number of nodes have a high degree distribution and a large number of nodes have a small degree distribution [6, 22] in both spatial and temporal dimension. In other words, the number of people in terms of spatial distance or inter-contact time has a scale-free distribution: a small number of people travel a long distance or have a long inter-contact time with others, whereas a large number of people travel a short distance or have a short inter-contact time with others. The scale-free characteristics identified in the human movement data over spatial and temporal dimensions are a representation of the whole system distribution of all individuals. Representing other attributes (i.e., location, demography) of each individual in other visual analytics views and using standard linked brushing methods to connect points in the scale free graph to their matching entities in other views allows users to explore questions in terms of "who," "where," "when," and "what" on each individual.

Cho, Myers, and Leskovec [50] further identify the impact of spatial and social factors on human movement: short-range travel is spatially and temporally periodic with little impact by the social ties, which have a strong impact on long-distance travel. Balcan et

al. [20] develop a unified model to study the multiscale nature of human mobility and its relationship with epidemic spread, including airline traffic networks and short-range commuting interactions. Crandall et al. [53] take work a step further to develop models that quantify how likely it is that two people know each other, if they have a very close geographic distance at approximately the same time. These results open new directions for new perspectives on not only link prediction but also network dynamics with spatial, social, and even multiscale considerations. The topic of predictive analysis will be addressed directly in Section 3.3 below.

3.2 Decision-making

In addition to enabling an efficient insight gain from a complex dataset, another major application in visual analytics is to use the insight to support a decision-making process. To design a visual analytic tool to effectively support human decision-making related to the interactions among geographic and social contexts, it is important to understand how people process information and how people make decisions in real situations.

3.2.1 Conceptual framework

Decision-making is a process to reduce uncertainty and doubt, enabling individuals to take a reasonable course of action facing complex decision problems, often in time pressured situations [91]. The process of decision-making consists of three steps: analyze the situation, find out relevant alternatives, and select an alternative by certain criteria [109]. Here, we draw upon two theoretical perspectives to frame the discussion of geo-social visual analytics for decision-making: situation awareness and spatial multicriteria decision analysis.

The conceptual framework from situation awareness (SA) can represent the decision-making process from a cognitive perspective and also integrate data exploration, decision-making and predictive analysis in the context of visual analytics. SA can be defined as "the human user's internal conceptualization of a situation" [110, p. 3609]. Endsley [64] defines three levels of SA: the first level is the perception of elements in the current situation, the second level is the comprehension of the current situation, and the third level is the projection of future status (Figure 5). A reasonable decision-making process should be based on an understanding of the current situation from the first two steps of SA, and also a prediction for future situations. The first two steps also match Figure 2 in terms of a mental model building process to comprehend the current situation.

Spatial multicriteria decision analysis aims to integrate GIS and multicriteria decision making (MCDM), and both of them can provide different techniques and methodologies to transform geographical data and the decision-maker's preferences to obtain information and knowledge to support decision-making [133]. More details regarding GIS-based MCDM (GIS-MCDM) can be found in Malczewski [134]. Spatial decision analysis is an inherently multicriteria decision process, involving economic, social, environmental, and political dimensions [106]. The territorial embeddedness and societal embeddedness in Figure 1 can represent the essence of multicriteria decisions in spatial decision analysis. In the section below, we propose adding another dimension into spatial multicriteria decision analysis, the social network.

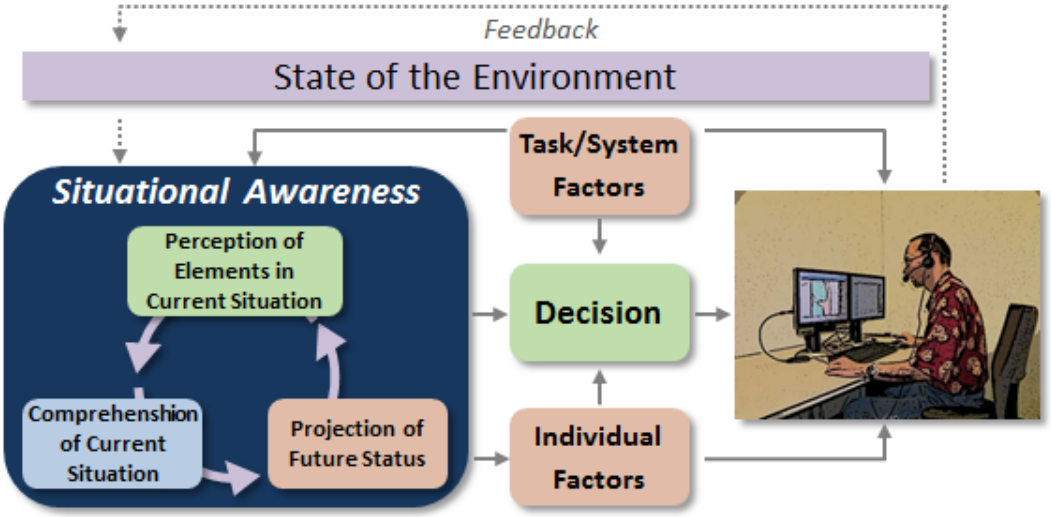


Figure 5: The process of decision-making for situation awareness. Situation awareness is defined as the human user's internal conceptualization of a situation, which plays a key role in effective decision making. Figure adapted from [109]. Spatial data analysis and social network analysis can support decision-making from different perspectives, so the effective integration of both can enhance human user's internal conceptualization of a situation.

3.2.2 Geo-social visual analytics in decision-making

Spatial data analysis and social network analysis have their independent advantages to support decision-making. As discussed in the proposed conceptual framework for geo-social relationships (Figure 1), both analysis approaches should be considered together. This section illustrates how integrating spatial data analysis and social network analysis frames the decision-making problem in a unique and insightful way. We argue here that their integration can be more powerful in support of decision-making than the sum of the parts. Limited research has been carried out thus far that achieves such integration. This section proposes two categories of geo-social integrations in terms of decision-making: the first one discusses an integrative approach toward spatial and social network factors to support a decision-making process; the second one discusses how social network structures impact GIS-based MCDM.

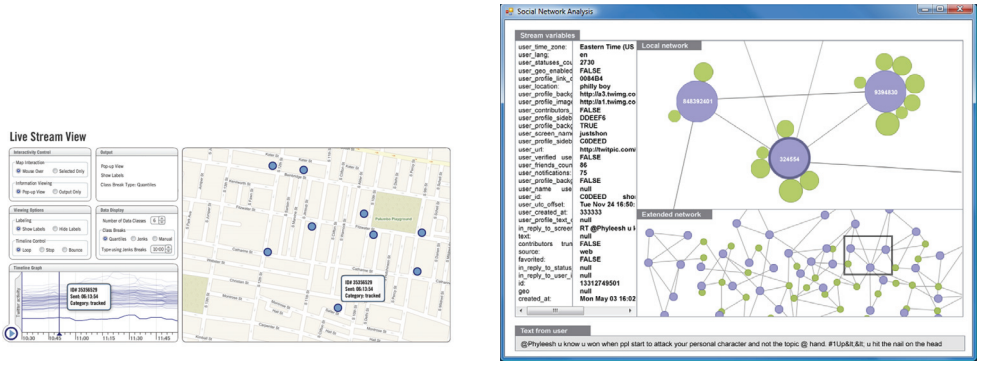
An integrative approach of spatial and social network factors Spatial data analysis is used to detect and visualize spatial patterns (e.g., disease, crime), and relate these patterns to salient explanatory covariates (e.g., economic and demographic factors), and then these insights are used to give decision-making support for polices [18]. However, many social phenomena are complex systems that mainly grow from the bottom-up, while traditional spatial data analysis focuses on top-down methods that cannot deal with the question of how the phenomena being analyzed evolves over space and time [24, 94]. Network science, a bottom up approach, provides the potential to link individual behaviors and inter-

actions among individuals to the size, scale, and shape of social phenomena observed in a spatiotemporal framework [26]. For example, urban sprawl that used to be understood through deterministic methods within economic location theory [7] has more recently been considered as a problem of organized complexity [97]. Network science has been proposed to study both urban physical networks (e.g., transport systems, water delivery) and urban social networks (e.g., industrial ecosystem) to build theories of how cities function as complex systems [17, 25–27]. We argue here that, while the application of network science has achieved important insights, a problem such as urban sprawl (or any other phenomenon that is both place-based and the product of complex societal factors) can be most fully understood through conceptual approaches that integrate geographical with social context and methods that integrate spatial with network analysis. As noted elsewhere in this paper, the complexity of such problems is what visual analytics methods are designed to address.

The two most typical network phenomena: small-world networks (characterized by high local clustering and short average node-to-node distance) [194] and scale-free networks (in which, as noted above, a small number of nodes have a high degree distribution and a large number of nodes have a small degree distribution) [6], have shown a strong relationship with space and time. For example, the famous small world experiments to study the average path length for social networks of people observed these relationships at two geographical levels: U.S. [144] and world [193]. Additionally, small-world and scale-free properties have been demonstrated to exist in many spatial-social networks (i.e., World Wide Web graph, power grid graph, and road networks) [5,22,82,124,194,202]. Even many traditional spatial phenomena exhibit scale-free characteristics such as city and company growth [25,170]. Finally, as discussed in the data exploration section, the existence of scale-free characteristics has been extended from social relationships into spatial and temporal dimensions through analyzing mobile phone and GPS data.

Geo-social visual analytic tools have the potential to directly enable decision-making that incorporates understanding of both geographic and social factors in an integrated way. One prototype of how such tools might work, TwitterHitter, was introduced by White and Roth [196]. The objective of TwitterHitter is to harvest information from Twitter.com to support the functions of crime analysis; these functions include decisions related to ongoing investigations as well as those related to deployment of personnel. TwitterHitter provides functions to plot a linked map-timeline view of the recent spatiotemporal activities of suspects on Twitter, and also can generate a directed network graph of the suspect's known associates (i.e., Twitter friends) (Figure 6). Some other spatial data analysis methods can also be used, such as geographically weighted regression [74], to understand the etiology (scientific analysis of the causes) of the criminal activity, with the collected tweets or their attributes as potential explanatory variables in the analysis. In complementary work focused on decisions related to disease outbreaks, Guo [85] proposes a geo-social visual analytic approach to analyze large spatial human interaction data to support effective pandemic control measures. The approach includes two linked views: a reorderable matrix and a map view to enable pattern interpretation in a geographical context and social context simultaneously. The geo-social interaction patterns provide valuable insight toward identifying critical locations and regions to suggest hypothetical control strategies for a pandemic outbreak based on synthetic population data.

The prototype tools developed by White and Roth [196] and by Guo [85] illustrate that spatial data analysis and network analysis can support decision-making from different perspectives. For example, spatial analysis in crime analysis demonstrates that explanatory



a. b.

Figure 6: Individual linked map-timeline and social network analysis views in TwitterHitter. The left view allows analysts to retrieve a spatiotemporal record of a suspect's activity on twitter. The right view can uncover all potential connections among suspects through network group of twitterers within a region. The views are courtesy of Jeremy D. White from Figures 3 and 4 in [196].

factors relevant to spatial clusters of crime include, but may not be limited by alcohol outlet densities [81], single person households [42], and depression [158]. Social network analysis in crime analysis has many important implications for crime investigations, such as targeting criminal leaders [201] and fighting organized crime proactively [141]. Decision-making in terms of disease control should not only require the observation of corresponding spatial patterns and driving factors behind these patterns [190], but also needs the networks through which diseases are transmitted from person to person [32]. Epidemic models, based on human interaction networks, can be used to simulate disease transmission processes and test the effectiveness of proposed control strategies [104]. Recent research implements human cognitive behaviors into epidemic models to consider human preventive behaviors, which results in a very good agreement with the observed influenza data [137,162]. Similarly, decision-making related to evacuation (e.g., in response to a hurricane threat) requires both spatial analysis related to location of people, evacuation routes, etc. as well as an understanding of how social connections impact individual evacuation decisions and behaviors [8].

The impact of social network structures on group decision-making In group/participatory settings, GIS-MCDM involves a series of activities, including defining problems, selecting evaluation criteria by group members, determining individual and collective preferences in terms of evaluation criteria and/or alternatives, sensitivity analysis with evaluation criteria and alternatives, exploring alternative combinations of individual preferences into group judgments, supporting group interaction to refine individual and group preferences, and having a final ordering of alternatives to make a compromise alternative available [123, 135]. Different stakeholders can be involved in the process of GIS-MCDM to face a variety of decision-making problems, such as environmental planning, transportation, and urban planning. Social relationships among those stakeholders have significant impact on their behaviors, which further has implications

for their decision abilities [34]. However, GIS-MCDM has not taken the impact of social relationships on actors' decision making into account.

The potential importance of social relationships for GIS-MCDM is illustrated in work by Bodin and Crona [34]. They review the role of social networks in terms of different relational patterns on governance process and outcomes: 1) high network density can facilitate collective action, reduce conflicts, and enhance knowledge development; 2) low degree of cohesiveness (i.e., clearly distinguishable subgroups) has negative effects on collaborative processes among subgroups [83]; 3) bonding ties among subgroups is beneficial for conflict resolution and collective action; 4) high degree of network centralization is positively correlated with collective action [163]. Furthermore, Bodin and Crona's research shows that none of the above network characteristics has a monotonically increasing positive effect on collective actions and conflict reduction, and that increasing one characteristic may cause the reduction of another. Therefore, how to maximize the positive effects of the individual and mix level of different network characteristics presents a key research challenge in terms of group decision-making.

The integration of visualization techniques into GIS-MCDM has received increasing attention [12, 98], but most studies focus on individual decision makers rather than groups [135]. Consequently, the collaborative tasks in GIS-MCDM with visualization/visual analytics have not been explored, not to mention considering the impact of social networks on decision-making. Here, we highlight a geo-social visual analytics tool developed to analyze public decision-making processes, and discuss the possibility to extend this tool into MCDM domains in order to consider the impact of social network structures.

Aguirre and Nyerges [3] introduce a novel geo-social visual analytics method that they label "grapevine" (Figure 7) that is directed to analysis of the very complex geo-social information generated within applications of web-based public participation systems for participatory learning and decision-making. The authors applied the grapevine tool to analysis of data collected during a month-long, online and asynchronous citizen advisory activity focused on planning for transportation in Puget Sound. The analysis enabled by the tool allowed Aguirre and Nyerges to partially confirm a hypothesis about analytic–deliberative decision-making, "that decisions are better when they come from a combination of analysis and deliberation rather than from analysis alone" [3, p. 320]. But, it also allowed them to identify key challenges in supporting deliberative processes that attempt to engage a wide cross-section of the public in deliberation that includes technical information and complex problems.

The grapevine tool is intended to help researchers understand the complex geo-social activities making up technology-enabled public decision-making. There are three underlying network structures in this tool: the main stem of the grapevine connects one node to another that represents users' posts; participants vote for each other's posts; participants reply to each other's posts. Aguirre and Nyerges discuss the potential to use social network analysis to understand the frequency of interactions and roles of people from a theoretical perspective, but how to use social network analysis in the real case study with the grapevine tool has not been explored. Therefore, the grapevine tool can be extended from four perspectives to integrate social network perspectives into GIS-MCDM in group/participatory settings for future work. First, the grapevine tool focuses on individual decision makers rather than groups. Second, the impact of structural social networks on decision-making reviewed by Bodin and Crona [34] can be considered. Third, although the grapevine tool is not a GIS-MCDM, it can generate individual and collective alterna-

tives for MCDM to allow decision-makers to choose and negotiate to support collaborative tasks. Last, agent-based modeling, a collection of agents that assess their situations and make decisions based on certain rules [37], can be used to simulate the group decision-making process.

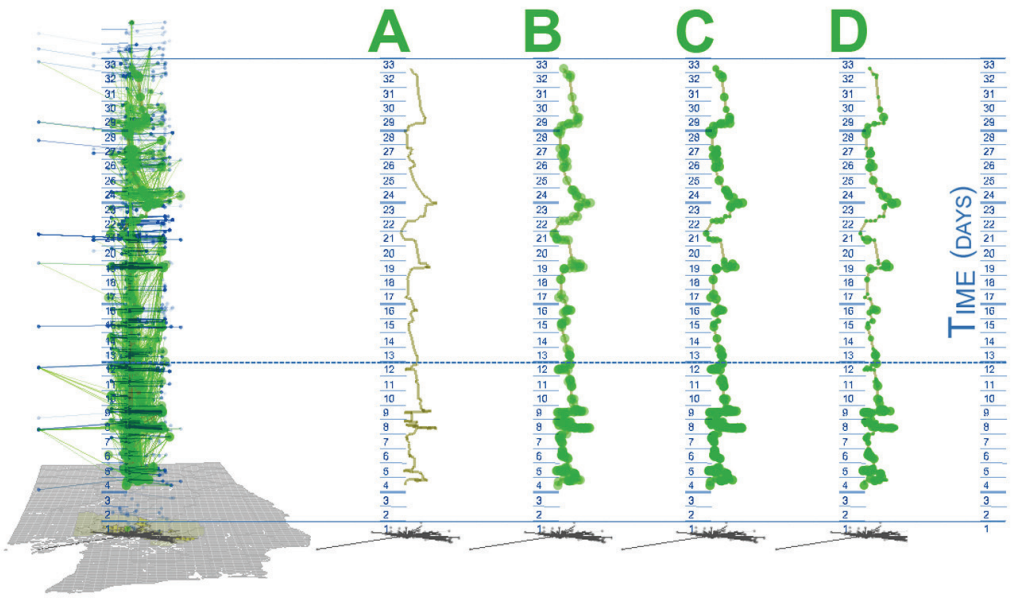


Figure 7: The static display of the grapevine. A main stem grows up with the increasing number of messages over geographical space. The main stem generates nodes when the number of votes to messages increases. The nodes generate buds when the number of replies (thus social connections) to messages increases. It shows the main stem in four states: (A) all features are turned off; (B) nodes are turned on; (C) the number of nodes is according to the number of votes; (D) the size of buds is according to the number of replies. Reproduced from Figure 4 in [3].

This section presents an argument that there are two categories of geo-social integrations in terms of decision-making: a decision-making process considering spatial and social network factors, and the impact of social network structures on GIS-based MCDM. The decision-making process is an iterative process that requires the support of a visual-interactive environment, especially in time-pressured situations [176]. One common problem with current geo-social visual analytic tools in terms of decision-making is that they are focused on helping analysts understand the current situations from the first two steps of SA, but lack the power to support decisions relevant to the current SA or predictions for future situations.

3.3 Predictive analysis

As discussed in the decision-making section, the SA model also provides the theoretical framework for predictive analysis in the context of visual analytics. Predictive analysis is

not independent from the first two steps of the SA model: it requires the understanding of the past and current situations through data exploration. An internal conceptualization of a situation, aided by predictive models and human reasoning, is the key to predictive analysis.

Developing mathematical models to support predictive analysis starts with the understanding of patterns found in real-world data. For example, based on the common property: scale-free characteristics observed in many large networks (e.g., actor collaboration graph, World Wide Web graph), Barabási and Albert [22] build a preferential attachment model to explain the development of scale-free networks in which networks tend to continue to grow with new vertices, and new vertices have a preferential attachment to vertices that are already well connected. The preferential attachment model has been used to make predictions of network growth with scale-free characteristics, but this model does not consider the impact of geographical constraints on the network growth. As discussed in the data exploration section, scale-free characteristics have been extended into spatial and temporal dimensions in terms of human mobility. Lee et al. [121] develop a new mobility model called SLAW (self-similar least action walk) that can capture all human mobility features reviewed in the data exploration section, including the Lévy flight travel patterns [39], spatial heterogeneously bounded mobility [78], power-law inter-contact times (ICTs) [44, 102], and fractal waypoints [155]. However, none of those mathematical models in terms of geo-social relationships have been implemented into geo-social visual analytics to empower prediction.

Recent studies have shown that mathematical models have a better prediction performance when they consider multiple components of context information (i.e., spatial, network, societal in Figure 1) rather than just one. For example, Andris, Halverson, and Hardisty [15] develop a new model considering physical and social space for predicting future migration, and for U.S. migration flows among major cities; the model outperforms a gravity model considering physical space alone. In complementary work, Takhteyev et al. [173] find that pre-existing ties (i.e., frequency of air travel) between places and people is the best predictor of Twitter ties compared to three other spatial and social factors including geographic distance, national boundaries, and language. A related study shows that using place-based attributes (i.e., social, economic, and ecological context) can successfully predict community membership more than 70% of the time in a large-scale social network of cell phone towers [43].

Geo-social visual analytics designed to support predictive analysis implements mathematical models in a visual-interactive environment to allow users to select appropriate methods, to set parameters, to interpret results, to understand what to do next, and to draw conclusions based on different scenarios. For example, Brigantic et al. [38] introduce a visual analytic tool (PanViz) with metapopulation-based epidemic models to rapidly assess alternative mitigation strategies in terms of pandemic influenza to give decision makers support. The PanViz system was subsequently extended and deployed in the Indiana State Department of Health Planning (Figure 8) to support analysis of potential epidemic control strategies [132]. While PanViz demonstrates the potential of geovisual analytics, it does not explicitly include capabilities to incorporate social network information into the analysis. Bisset and Marathe [33] have developed a similar tool that does include such capabilities: EPISIMS with individual-based epidemic models to simulate the dynamics of millions of individuals, traffic of entire cities, and disease spread, respectively. In related research, Broeck et al. [40] present a visual analytic tool "GLEaMviz" available to the public that

allows the user to set a variety of parameters to simulate the human-to-human infectious disease spread across the world.

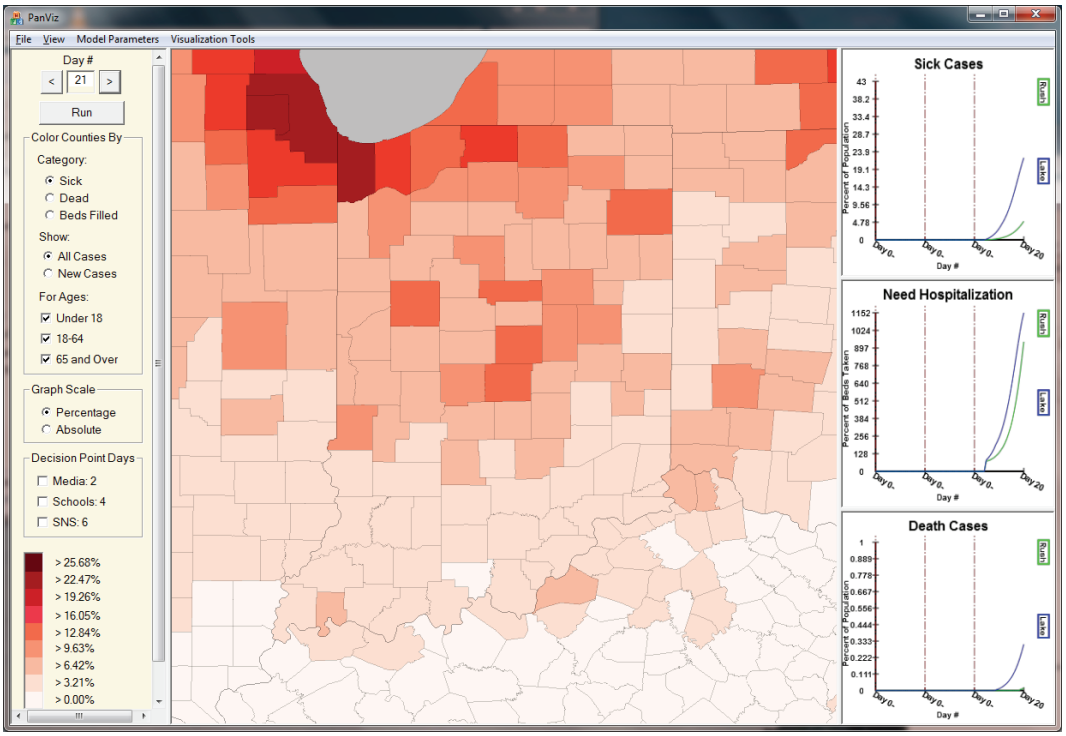


Figure 8: Here we illustrate the PanViz user interface. A user has simulated an outbreak originating in Chicago and has been exploring the spread of the infectious disease over space and time. This is day 21 of the simulation. The color of each county represents the percentage of the population that is ill. The plots on the right show the rate of spread (thus the rate of person-to-person connection), hospitalization, and death in user selected counties. Figure Courtesy of Ross Maciejewski from Figure 7 in [126]. For additional Information on the research it is derived from, see [132].

One big challenge in terms of predictive analysis is that geo-social systems are highly sensitive to social adaptive behaviors [189]. In crisis situations (e.g., pandemics, natural disasters), the geo-social systems behave abnormally which is still a challenge to predict. Vespignani [189] proposes three challenges for future work to predict social adaptive behaviors in a period of crisis: collecting data on information spread and social reactions in the period of crisis; developing models to quantify the effect of risk perception and awareness phenomena of individuals; and deploying monitoring infrastructures capable of informing computational models in real time. The three suggestions also need geo-social visual analytics to support methods to address all three challenges in an integrative system.

In research directly focused on predictive analytics for crisis events, Bengtsson et al. [28] focus on displacement after the 2010 Haiti Earthquake. Specifically, they track population with phone call data from six weeks before the disaster to five months after. Their estimates in terms of the number, the timing, and geographical distribution of population movements

correspond well with a retrospective survey. A follow-up study identifies that the destinations in which people stayed had significant social bonds and the time that the displaced population stays outside the city follows a skewed, fat-tailed distribution [126]. Bengtsson and colleagues have taken the first step to demonstrate that the prediction of population movements with the use of phone data in disaster response is possible, but substantial future work is still required before common usage. For example, cell phone data availability during natural disasters is still limited [28]; data coverage varies over space, time, and different groups of people [77]; and data privacy is always a big concern [117].

Above all, the process of predictive analysis is an iterative process that needs human interpretation, control, experiences, and imagination, especially in crisis periods. As discussed here, current geo-social visual analytic tools in terms of decision-making are good at enabling perception of elements in and comprehension of the current situations, thus the first two steps of SA, but they lack the power to make predictions for future situations. Meanwhile, there is a gap between the insight gained in the process of data exploration and predictive analysis. For example, the current predictive epidemic visual analytic tools implement epidemic models to simulate disease transmission and design corresponding control strategies without flexible approaches analyzing human spatial-social interactive clusters (like those illustrated by Guo [85] in Section 3.2.2.1), as knowledge input to improve control strategy design.

4 Discussion and challenges

Geo-social data do not make any sense when abstracted from their appropriate contexts [1]. Geo-social contexts do not only demonstrate conceptual and observed overlaps, but also shed light on data from different perspectives. Downs and DeSouza [59] argue that spatial thinking serves three purposes: 1) a descriptive function, 2) an analytic function, and 3) an inferential function. We follow Downs and DeSouza's lead and propose extending consideration of spatial thinking into spatial-social thinking. Social thinking describes how human interactions impact their ideas, emotions, and behaviors [198]. A social network approach provides a means to describe human interactions and analyze and infer how such interactions impact the social thinking process. We propose that the three purposes that spatial thinking serves can be applied to social thinking to develop integrative spatial-social thinking for people to acquire knowledge. The three purposes also match the primary goal of visual analytics: interactive visual interfaces should support human analytical reasoning in an efficient and effective way. Supporting spatial and social thinking to address the three purposes outlined is a prototypical example of meeting this goal.

Based on the above discussion, multiple geo-social visual analytics research questions need to be addressed in relation to data exploration, decision-making, and predictive analysis. In addition, data exploration, decision-making, and predictive analysis have inner connections: knowledge development in the data exploration process can support decision-making and predictive analysis, which sometimes leads decision-making and predicative analysis to develop a new knowledge construction process. Therefore, the overarching goal of visual analytics is to involve human ability with the whole complex analytical process rather than each step separately [9].

Our analysis of the literature provides several potential research directions for further investigation of geo-social visual analytics. We conclude this review paper by highlighting

nine core challenges that will require interdisciplinary efforts to meet. The challenges can be classified into two categories: the first one focuses on further integration of spatial and social network analytics from an interdisciplinary perspective, and the second one focuses on understanding the impact of geo-social relationships on society.

4.1 An effective integration of spatial and social network analytics

Understanding the interaction of geography, network, and societal space, as well as their respective and coupled impacts on outcomes of interest. This challenge underlies the entire paper. Based on our review it seems clear that understanding how spatial proximity and network relationships interact for outcomes of interest is at an early stage [173, 188]. Thus, further research is required to investigate how geographical and social relationships operate explicitly in different geo-social systems. In addition to spatial analysis and social network analysis, multivariate analyses methods, that conceptually fit the societal context in Figure 1, can be implemented into geovisual analytics tools to understand the relationships between variables and their relevance to the spatial-social patterns being studied. This paper primarily focuses on the spatial-social relationships and methods to understand them; complementary research on multivariate analysis and geovisualization can be found in the following papers [46–48, 84, 87, 88].

Developing theory, methods, and tools to consider spatial and network factors simultaneously. Most geo-social analytical approaches use independent traditional spatial analysis and social network analysis methods simultaneously to explore the same datasets [63]. It is necessary to develop new theory and methods that integrate spatial and social factors together. For example, Radil et al. [151] propose a new method that borrows the concept of social position to explore an actor's position in a spatial contiguity matrix simultaneously with his or her position in social networks. The proposed method can identify statistically significant violence patterns which cannot be captured by the classical spatial autocorrelation method, Global Moran's I [146, 178]. Radil and his colleagues' work is only one attempt to explore new geo-social theory and methods. As discussed in this paper, spatial analysis and social network analysis exhibit strong conceptual and observed overlaps, so much more future work is needed.

Understanding the dynamics of geo-social relationships and processes. Geo-social relationships are not static but dynamic, so it is important to understand the change of the relationships over time and compare the dynamic change to the static understanding [2] as well as to investigate the linked geographic and social process that drives the change. Hess [92] also discusses the need to involve a temporal concept into the proposed categories of embeddedness (Figure 1) through taking into account developments over time and changes in the spatial configuration of networks at different scales. We argue that three models can be developed to represent temporal geo-social relationships among geographical or individual units. The first model has fixed nodes with unchanging geographical relationships and varying social relationships. The model is based on the geo-social phenomena in which geographical relationships are fixed because of their relative geographical locations (e.g., cities, states, and nations), and social relationships change over time. Those geo-social phenomena include the dynamic trade network at a country-to-country scale [67, 207], human migration in the U.S. at a county-to-county scale [86], and technology adoption (i.e.,

Twitter) at a city-to-city scale [183]. The second model has fixed nodes with changing geographical relationships and relatively fixed social relationships (i.e., human movement, mobile-based social media). The last model can support more complex geo-social dynamic behaviors in which networks can expand and recede [24]. Recent developments that apply the concept of "rendezvous" (bringing sensors close to one another in space or time [96] to shed light on human mobility characteristics [95]) provides data and methods to support the last two models.

Integrating distinct applications of cognitive science to support geo-social visual analytics. Cognitive science provides theoretical frameworks for the design of geovisualization tools [129], it provides a conceptual approach (e.g., distributed cognition) to understand human reasoning as enabled by visual tools, and it also offers fundamental theories and approaches to understand and model human behaviors in network science, as discussed in the decision-making section. For example, the Organizational Risk Analyzer (ORA) uses both network theory and social psychology to model human behaviors, and ORA has been used to analyze 1500 videos made by insurgents in Iraq and effectively reduce sniper activity by 70% [35]. From a social cognitive perspective, human behaviors result from an interaction between human internal cognition and externally environmental effects [21]. The external effects also fit the three spaces in the proposed conceptual framework for geo-social relationships (Figure 1), because the external effects include human socioeconomic status (societal embeddedness), their relationships with others (network embeddedness), and materials provided by a specific location in which the personal is located (territorial embeddedness). Involving social cognitive theory in modeling human behaviors may make contributions to predict geo-social systems in crisis situations discussed in the predictive analysis section. Therefore, cognitive science should not only be used to design visualization tools [71], but to support geo-social analytic models as well.

Developing new geo-social visual analytics methods to incorporate data exploration, decision-making, and predictive analysis as a whole. As discussed above, most geo-social visual analytics methods only support one step, so visual analytics cannot effectively transform knowledge through visual exploration into complex analytical strategies directly. One possible solution is to improve inter-disciplinary cooperation through understanding the human analytical reasoning of real decision-makers to design visual analytic tools accordingly [10], such as has been attempted for maritime anomaly detection [156], bridge management system analysis [191], and other application domains. In addition, visual analytics have not synergistically integrated computational methods to maximize human conceptual, perceptual and reasoning capabilities in the whole scientific and problem-solving process. One possible theoretical framework to link the whole complex analytical process can be found in Gahegan [75], situating human reasoning, concretized representation, conceptual structures, visual representation, and mathematical models into the whole science process. To support a full range of applications of geo-social visual analytics, however, the approach must be generalized beyond the context of scientific research, which was the target of Gahegan's model.

4.2 The impact of geo-social relationships on society

Understanding the transition of daily life habits that further impact local communities and networks because of spatial decision-making. As reviewed in the decision-making section, GIS-MCDM

includes broad application domains (e.g., transportation, urban planning). Decision-making in terms of those domains involves interests among different stakeholders, but it also has a corresponding influence on the practices of everyday life. The influence will further impact local community interaction and network structures. For example, people who used to live in *hutongs* (traditional alleyway neighborhoods) in Beijing report a substantial disruption of the high quality and frequency of local interaction they had in hutong compared to that after they were relocated to mega-block high rise apartment complexes on the city periphery when their neighborhood underwent urban renewal [157]. Tita et al. [179] point to a similar impact of urban redevelopment on social networks in their argument that the clear north-south geographical division in the gang rivalry networks in a section of Los Angeles is due to a landscape feature: the San Bernadino Freeway. Although spatial decision making has a significant role in shifting local community network structures, research in GIS-MCDM has not taken such shifts into account.

Understanding the shift of the traditional decision-making approaches with the emergence of social media. The emergence of social media (i.e., Facebook, Twitter, LinkedIn) changes the world via collective power through on-line social networks. One person can communicate with hundreds or even more people about products, news, cultures, and any information. The communication occurs in a smaller world than in the pre-internet era; this is illustrated by Kwak et al. [111] who find that the average path length of a Twitter network is 4.12 compared to "six degrees of separation" [144] in the real world. The impact of people-to-people communication has greatly changed the traditional sense of decision-making, because social media based conversations help people to be accountable and occur outside of the direct control of decision-makers [136]. For example, social media is playing an increasing role in the most recent anti-government protests, including the Arab Spring, Occupy Wall Street, and the London Riots [184]. MacEachren et al. [131] leverage Twitter into a web-enabled geovisual analytics application and discuss how social media can offer strategies for disaster and emergency management. While social media have a transformative impact on traditional decision-making approaches, strategies through which responding organizations can successfully leverage these technologies are just beginning to be considered [203]. Before effective use of social media in decision-making can be achieved, there are many unexplored research questions. For example, how does information diffuse geographically and socially via social media? How do social media change human behaviors in normal and crisis situations? How do social media transform individual voices into collective power to be accountable?

Developing a framework to collect geo-social relationship data and assess their fitness for different applications while also considering the potential negative consequences for human privacy of collecting these data. The potential of geo-social visual analytics, especially during natural disasters, disease outbreaks, and similar events that put people and property at risk, provides additional motivation for future collection of network and spatial data. The most popular geo-social network data collection methods include surveys [153], crawling social media sites [162], collecting data from mobile devices such as cell phones [62], and leveraging wireless sensor technology [103]. Details about each method and their pros and cons can be found in Salathé et al. [161]. A key problem here is that there is no theory/framework to assess whether the collected data are suitable to study different applications. For example, Salathé and Jones [160] study disease transmission at individual-level through building

social interaction networks. Nodes represent individuals and edges consider both friendships in Facebook and physical proximity in real world (i.e., the same dorm, the same class). However, the demographics of social media users is a biased sample of the whole population (in relation to age, gender, race, etc.) and such networks are still at a rather coarse resolution for the study of disease transmission. As more complete data sets become available, individual-level network data with spatial and temporal information may make it possible to predict human behaviors better, but collection of individual level data raises a range of privacy concerns [36].

To sum up, geo-social visual analytics is based on the conceptual extension of the First Law of Geography: everything/everyone is related to everything/everyone else, but near things/persons are more related than distant things/persons [180]; nearness and relationships can be considered a matter of geographical and social network distance, relationship and interaction. The observed social phenomena in a spatiotemporal framework motivate the development of social, political, and ethical research questions finally to develop th general geo-social theory to understand the world. Thus, at the methodological level, geo-social visual analytics should facilitate the integration of computational methods with human reasoning abilities to answer research and application questions in the context of data exploration, decision-making, and predictive analysis. The resulting methods will, from an integrative perspective of spatial thinking and social science, enable research to understand the geo-social mechanisms and processes that underlie human behavior.

Acknowledgments

This material is based, in part, upon work supported by the U.S. Department of Homeland Security under Award #: 2009-ST-061-CI0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security; a grant from the Gates Foundation also provided partial support. We thank Guo Diansheng and Ross Maciejewski for permission to include their previously unpublished figures illustrating their research. We thank Krista Kahler for redrawing the Figure 5.

References

[1] ABBOTT, A. Of time and space: The contemporary relevance of the Chicago School. *Social Forces 75*, 4 (1997), 1149–1182. doi:10.1093/sf/75.4.1149.

[2] ADAMS, J., FAUST, K., AND LOVASI, G. Capturing context: Integrating spatial and social network analyses. *Social networks 34*, 1 (2012), 1–5. doi:10.1016/j.socnet.2011.10.007.

[3] AGUIRRE, R., AND NYERGES, T. Geovisual evaluation of public participation in decision making: The grapevine. *Journal of Visual Languages & Computing 22*, 4 (2011), 305–321. doi:10.1016/j.jvlc.2010.12.004.

[4] AHMED, A., FU, X., HONG, S., NGUYEN, Q., AND XU, K. Visual analysis of history of world cup: A dynamic network with dynamic hierarchy and geographic cluster-

ing. In *Visual Information Communication*, M. L. Huang, Q. V. Nguyen, and K. Zhang, Eds. Springer, 2010, pp. 25–39. doi:10.1007/978-1-4419-0312-9_2.

[5] ALBERT, R., ALBERT, I., AND NAKARADO, G. Structural vulnerability of the North American power grid. *Physical Review E 69*, 2 (2004), 025103. doi:10.1103/PhysRevE.69.025103.

[6] ALBERT, R., JEONG, H., AND BARABÁSI, A. Diameter of the world wide web. *Nature 401*, 6749 (1999), 130–131. doi:10.1038/43601.

[7] ALONSO, W. *Location and land use. Toward a general theory of land rent*. Harvard University Press, Cambridge, MA, 1964.

[8] ALSNIH, R., AND STOPHER, P. Review of procedures associated with devising emergency evacuation plans. *Transportation Research Record 1865*, 1 (2004), 89–97. doi:10.3141/1865-13.

[9] ANDRIENKO, G., ANDRIENKO, N., JANKOWSKI, P., KEIM, D., KRAAK, M., MACEACHREN, A., AND WROBEL, S. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science 21*, 8 (2007), 839–858. doi:10.1080/13658810701349011.

[10] ANDRIENKO, G., ANDRIENKO, N., KEIM, D., MACEACHREN, A., AND WROBEL, S. Challenging problems of geospatial visual analytics (editorial introduction). *Journal of Visual Languages & Computing 22*, 4 (2011), 251–256. doi:10.1016/j.jvlc.2011.04.001.

[11] ANDRIENKO, G., ANDRIENKO, N., KOPANAKIS, I., LIGTENBERG, A., AND WROBEL, S. Visual analytics methods for movement data. In *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*, F. Giannoni and D. Pedreski, Eds. Springer, Berlin, 2008, ch. 13, pp. 375–410. doi:10.1007/978-3-540-75177-9_14.

[12] ANDRIENKO, N., AND ANDRIENKO, G. Informed spatial decisions through coordinated views. *Information Visualization 2*, 4 (2003), 270–285. doi:10.1057/palgrave.ivs.9500058.

[13] ANDRIENKO, N., AND ANDRIENKO, G. Spatial generalisation and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics 17*, 2 (2010), 205–219. doi:10.1109/TVCG.2010.44.

[14] ANDRIS, C. *Metrics and methods for social distance*. PhD thesis, Massachusetts Institute of Technology, 2011.

[15] ANDRIS, C., HALVERSON, S., AND HARDISTY, F. Predicting migration system dynamics with conditional and posterior probabilities. In *Proc. IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)* (2011), IEEE, pp. 192–197. doi:10.1109/ICSDM.2011.5969030.

[16] ASHDOWN, P. The global power shift. TED, 2005. http://www.ted.com/playlists/73/the_global_power_shift.html.

[17] ASHTON, W. Understanding the organization of industrial ecosystems. *Journal of Industrial Ecology 12*, 1 (2008), 34–51. doi:10.1111/j.1530-9290.2008.00002.x.

[18] BAILEY, T., AND GATRELL, A. *Interactive spatial data analysis*. Longman Scientific & Technical, Essex, UK, 1995.

[19] BAK, P., OMER, I., AND SCHRECK, T. Visual analytics of urban environments using high-resolution geographic data. In *Geospatial Thinking*, M. Painho, M. Y. Santos, and H. Pundt, Eds., Lecture Notes in Geoinformation and Cartography. Springer, Berlin, 2010, pp. 25–42. doi:10.1007/978-3-642-12326-9_2.

[20] BALCAN, D., COLIZZA, V., GONÇALVES, B., HU, H., RAMASCO, J., AND VESPIGNANI, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21484–21491. doi:10.1073/pnas.0906910106.

[21] BANDURA, A. Social cognitive theory: An agentic perspective. *Annual Review of Psychology 52*, 1 (2001), 1–26. doi:10.1111/1467-839X.00024.

[22] BARABÁSI, A. *Linked: How everything is connected to everything else and what it means*. Perseus, Cambridge, MA, 2002.

[23] BARABÁSI, A. Scale-free networks: A decade and beyond. *Science 325*, 5939 (2009), 412–413. doi:10.1126/science.1173299.

[24] BATTY, M. Network geography: Relations, interactions, scaling and spatial processes in GIS. In *Re-presenting GIS* (2005), pp. 149–170.

[25] BATTY, M. Rank clocks. *Nature 444*, 7119 (2006), 592–596. doi:10.1038/nature05302.

[26] BATTY, M. The size, scale, and shape of cities. *Science 319*, 5864 (2008), 769–771. doi:10.1126/science.1151419.

[27] BAYNES, T. Complexity in urban development and management. *Journal of Industrial Ecology 13*, 2 (2009), 214–227.

[28] BENGTSSON, L., LU, X., THORSON, A., GARFIELD, R., AND VON SCHREEB, J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine 8*, 8 (2011), e1001083. doi:10.1371/journal.pmed.1001083.

[29] BERA, R., AND CLARAMUNT, C. Topology-based proximities in spatial systems. *Journal of Geographical Systems 5*, 4 (2003), 353–379. doi:10.1007/s10109-003-0115-y.

[30] BETTENCOURT, L., LOBO, J., HELBING, D., KÜHNERT, C., AND WEST, G. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences 104*, 17 (2007), 7301–7306. doi:10.1073/pnas.0610172104.

[31] BETTENCOURT, L., AND WEST, G. A unified theory of urban living. *Nature 467*, 7318 (2010), 912–913. doi:10.1038/467912a.

[32] BIAN, L., AND LIEBNER, D. A network model for dispersion of communicable diseases. *Transactions in GIS 11*, 2 (2007), 155–173. doi:10.1111/j.1467-9671.2007.01039.x.

[33] BISSET, K., AND MARATHE, M. A cyber environment to support pandemic planning and response. *DOE SciDAC Review Magazine* (2009).

[34] BODIN, Ö., AND CRONA, B. The role of social networks in natural resource governance: What relational patterns make a difference? *Global Environmental Change 19*, 3 (2009), 366–374. doi:10.1016/j.gloenvcha.2009.05.002.

[35] BOHANNON, J. Counterterrorism's new tool: "Metanetwork" analysis. *Science 325*, 5939 (2009), 409–411. doi:10.1126/science.325_409.

[36] BOHANNON, J. Investigating networks: The dark side. *Science 325*, 5939 (2009), 410–411. doi:10.1126/science.325_410.

[37] BONABEAU, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences 99*, Suppl 3 (2002), 7280–7287. doi:10.1073/pnas.082080899.

[38] BRIGANTIC, R., EBERT, D., CORLEY, C., MACIEJEWSKI, R., MULLER, G., AND TAYLOR, A. Development of a quick look pandemic influenza modeling and visualization tool. In *Proc. 7th International ISCRAM Conference* (Seattle, USA, 2010).

[39] BROCKMANN, D., HUFNAGEL, L., AND GEISEL, T. The scaling laws of human travel. *Nature 439*, 7075 (2006), 462–465. doi:10.1038/nature04292.

[40] BROECK, W., GIOANNINI, C., GONÇALVES, B., QUAGGIOTTO, M., COLIZZA, V., AND VESPIGNANI, A. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infectious Diseases 11*, 1 (2011), 37. doi:10.1186/1471-2334-11-37.

[41] BUCHANAN, M. *Nexus: Small worlds and the groundbreaking theory of networks*. W.W. Norton, New York, 2003.

[42] CAHILL, M., AND MULLIGAN, G. Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review 25*, 2 (2007), 174–193. doi:10.1177/0894439307298925.

[43] CAUGHLIN, T., RUKTANONCHAI, N., ACEVEDO, M., LOPIANO, K., PROSPER, O., EAGLE, N., AND TATEM, A. Place-based attributes predict community membership in a mobile phone communication network. *PLoS ONE 8*, 2 (2013), e56057. doi:10.1371/journal.pone.0056057.

[44] CHAINTREAU, A., HUI, P., CROWCROFT, J., DIOT, C., GASS, R., AND SCOTT, J. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing* (2007), 606–620.

[45] CHANG, R., ZIEMKIEWICZ, C., GREEN, T., AND RIBARSKY, W. Defining insight for visual analytics. *IEEE Computer Graphics and Applications 29*, 2 (2009), 14–17. doi:10.1109/MCG.2009.22.

[46] CHEN, J., CARR, D., WECHSLER, H., AND PAN, Z. Interactive visualization of multivariate statistical data. *The International Journal of Virtual Reality 5*, 3 (2006), 67–73.

[47] CHEN, J., AND MACEACHREN, A. Resolution control for balancing overview and detail in multivariate spatial analysis. *Cartographic Journal 45*, 4 (2008), 261–273. doi:10.1179/174327708X347764.

[48] CHEN, J., MACEACHREN, A., AND GUO, D. Supporting the process of exploring and interpreting space–time multivariate patterns: The visual inquiry toolkit. *Cartography and Geographic Information Science 35*, 1 (2008), 33–50. doi:10.1559/152304008783475689.

[49] CHEN, J., SHAW, S., YU, H., LU, F., CHAI, Y., AND JIA, Q. Exploratory data analysis of activity diary data: A space-time GIS approach. *Journal of Transport Geography 19*, 3 (2011), 394–404. doi:10.1016/j.jtrangeo.2010.11.002.

[50] CHO, E., MYERS, S., AND LESKOVEC, J. Friendship and mobility: User movement in location-based social networks. In *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2011), ACM, pp. 1082–1090. doi:10.1145/2020408.2020579.

[51] CHRISTAKIS, N., AND FOWLER, J. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown and Company, 2009.

[52] CLIFF, A., AND HAGGETT, P. On complex geographic space: Computing frameworks for spatial diffusion processes. In *Geocomputation: A primer*, P. Longley, S. M. Brooks, R. McDonnell, and B. Macmillan, Eds. John Wiley and Sons, New York, 1998, pp. 231–256.

[53] CRANDALL, D., BACKSTROM, L., COSLEY, D., SURI, S., HUTTENLOCHER, D., AND KLEINBERG, J. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences 107*, 52 (2010), 22436–22441. doi:10.1073/pnas.1006155107.

[54] CURTIN, K. Network analysis in geographic information science: Review, assessment, and projections. *Cartography and Geographic Information Science 34*, 2 (2007), 103–111. doi:10.1559/152304007781002163.

[55] DARAGANOVA, G., PATTISON, P., KOSKINEN, J., MITCHELL, B., BILL, A., WATTS, M., AND BAUM, S. Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social networks 34*, 1 (2012), 6–17. doi:10.1016/j.socnet.2010.12.001.

[56] DEMŠAR, U., ŠPATENKOVÁ, O., AND VIRRANTAUS, K. Identifying critical locations in a spatial network with graph theory. *Transactions in GIS 12*, 1 (2008), 61–82. doi:10.1111/j.1467-9671.2008.01086.x.

[57] DEMŠAR, U., AND VIRRANTAUS, K. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science 24*, 10 (2010), 1527–1542. doi:10.1080/13658816.2010.511223.

[58] DOREIAN, P., AND CONTI, N. Social context, spatial structure and social network structure. *Social networks 34*, 1 (2012), 32–46. doi:10.1016/j.socnet.2010.09.002.

[59] DOWNS, R., AND DESOUZA, A. *Learning to think spatially: GIS as a support system in the K-12 curriculum*. National Research Council and National Academies Press, Washington, DC, 2006.

[60] DOYTSHER, Y., GALON, B., AND KANZA, Y. Storing routes in socio-spatial networks and supporting social-based route recommendation. In *Proc. 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (2011), ACM, pp. 49–56. doi:10.1145/2063212.2063219.

[61] EAGLE, N., AND PENTLAND, A. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing 10*, 4 (2006), 255–268. doi:10.1007/s00779-005-0046-3.

[62] EAGLE, N., PENTLAND, A., AND LAZER, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*, 36 (2009), 15274–15278. doi:10.1073/pnas.0900282106.

[63] EMCH, M., ROOT, E., GIEBULTOWICZ, S., ALI, M., PEREZ-HEYDRICH, C., AND YUNUS, M. Integration of spatial and social network analysis in disease transmission studies. *Annals of the Association of American Geographers 102*, 5 (2012), 1004–1015. doi:10.1080/00045608.2012.671129.

[64] ENDSLEY, M. Theoretical underpinnings of situation awareness: A critical review. In *Situation awareness analysis and measurement*, M. R. Endsley and D. J. Garland, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 3–32.

[65] FABRIKANT, S., MONTEILO, D., AND MARK, D. The distance-similarity metaphor in region-display spatializations. *IEEE Computer Graphics and Applications 26*, 4 (2006), 34–44. doi:10.1109/MCG.2006.90.

[66] FABRIKANT, S., AND SKUPIN, A. Cognitively plausible information visualization. In *Exploring Geovisualization*, J. Dykes, A. M. MacEachren, and M.-J. Kraak, Eds. Elsevier Science, Amsterdam, 2005, pp. 667–690. doi:10.1016/B978-008044531-1/50453-X.

[67] FAGIOLO, G. The international-trade network: Gravity equations and topological properties. *Journal of Economic Interaction and Coordination 5*, 1 (2010), 1–25. doi:10.1007/s11403-010-0061-y.

[68] FAGIOLO, G., REYES, J., AND SCHIAVO, S. On the topological properties of the world trade web: A weighted network analysis. *Physica A: Statistical Mechanics and its Applications 387*, 15 (2008), 3868–3873. doi:10.1016/j.physa.2008.01.050.

[69] FAGIOLO, G., REYES, J., AND SCHIAVO, S. World-trade web: Topological properties, dynamics, and evolution. *Physical Review E 79*, 3 (2009), 0361151–03611519. doi:10.1103/PhysRevE.79.036115.

[70] FESTINGER, L., SCHACHTER, S., AND KURT, W. *Social Pressures in Informal Groups*. Stanford University Press, Stanford, CA, 1950.

[71] FISHER, B., GREEN, T., AND ARIAS-HERNÁNDEZ, R. Visual analytics as a translational cognitive science. *Topics in Cognitive Science 3*, 3 (2011), 609–625. doi:10.1111/j.1756-8765.2011.01148.x.

[72] FLINT, C. The theoretical and methodological utility of space and spatial statistics for historical studies: The Nazi Party in geographic context. *Historical Methods 35*, 1 (2002), 32–42. doi:10.1080/01615440209603142.

[73] FLINT, C., DIEHL, P., SCHEFFRAN, J., VASQUEZ, J., AND CHI, S. Conceptualizing conflictspace: Toward a geography of relational power and embeddedness in the analysis of interstate conflict. *Annals of the Association of American Geographers 99*, 5 (2009), 827–835. doi:10.1080/00045600903253312.

[74] FOTHERINGHAM, A., BRUNSDON, C., AND CHARLTON, M. *Geographically weighted regression: The analysis of spatially varying relationships*. John Wiley & Sons, 2002.

[75] GAHEGAN, M. Beyond tools: Visual support for the entire process of GIScience. In *Exploring Geovisualization*, J. Dykes, A. MacEachren, and M. Kraak, Eds. Elsevier, Amsterdam, 2005, ch. 4, pp. 83–99.

[76] GANTER, J., AND MACEACHREN, A. Cognition and the design of scientific visualization systems. Tech. rep., Department of Geography, Pennsylvania State University, 1989.

[77] GETHING, P., AND TATEM, A. Can mobile phone data improve emergency response to natural disasters? *PLoS Medicine 8*, 8 (2011), e1001085. doi:10.1371/journal.pmed.1001085.

[78] GONZALEZ, M., HIDALGO, C., AND BARABÁSI, A. Understanding individual human mobility patterns. *Nature 453*, 7196 (2008), 779–782. doi:10.1038/nature06958.

[79] GOODCHILD, M., ANSELIN, L., APPELBAUM, R., AND HARTHORN, B. Toward spatially integrated social science. *International Regional Science Review 23* (2000), 139–159. doi:10.1177/016001760002300201.

[80] GOODCHILD, M., AND JANELLE, D. Toward critical spatial thinking in the social sciences and humanities. *GeoJournal 75*, 1 (2010), 3–13. doi:10.1007/s10708-010-9340-3.

[81] GORMAN, D., SPEER, P., GRUENEWALD, P., AND LABOUVIE, E. Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *Journal of Studies on Alcohol and Drugs 62*, 5 (2001), 628–636.

[82] GOVINDAN, R., AND TANGMUNARUNKIT, H. Heuristics for Internet map discovery. In *Proc. INFOCOM* (Tel Aviv, 2000), pp. 1371–1380. doi:10.1109/INFCOM.2000.832534.

[83] GRANOVETTER, M. The strength of weak ties. *American Journal of Sociology 78*, 6 (1973), 1360–1380. doi:10.1086/225469.

[84] GUO, D. *Human-Machine Collaboration for Geographic Knowledge Discovery with High-Dimensional Clustering*. PhD thesis, The Pennsylvania State University, State College, 2003.

[85] GUO, D. Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science 21*, 8 (2007), 859–877. doi:10.1080/13658810701349037.

[86] GUO, D. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1041–1048. doi:10.1109/TVCG.2009.143.

[87] GUO, D., CHEN, J., MACEACHREN, A., AND LIAO, K. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1461–1474.

[88] GUO, D., GAHEGAN, M., MACEACHREN, A., AND ZHOU, B. Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science 32*, 2 (2005), 113–132. doi:10.1559/1523040053722150.

[89] GUO, D., LIU, S., AND JIN, H. A graph-based approach to vehicle trajectory analysis. *Journal of Location Based Services 4*, 3 (2010), 183–199. 10.1080/17489725.2010.537449.

[90] GUO, D., WU, K., ZHANG, Z., AND XIANG, W. Wms-based flow mapping services. In *Proc. Eighth IEEE World Congress on Services* (Honolulu, HI, 2012), pp. 234–241. doi:10.1109/SERVICES.2012.37.

[91] HARRIS, R. Introduction to decision making, 1988. http://www.virtualsalt.com/crebook5.htm.

[92] HESS, M. "spatial" relationships? towards a reconceptualization of embeddedness. *Progress in Human Geography 28*, 2 (2004), 165–186. doi:10.1191/0309132504ph479oa.

[93] HIPP, J., FARIS, R., AND BOESSEN, A. Measuring "neighborhood": Constructing network neighborhoods. *Social networks 34*, 1 (2012), 128–140. doi:10.1016/j.socnet.2011.05.002.

[94] HOLLAND, J. *Hidden order: How adaptation builds complexity*. Perseus Books, Cambridge, MA, 1996.

[95] HONICKY, R. *Towards a societal scale, mobile sensing system*. PhD thesis, UC Berkeley, 2011.

[96] HONICKY, R. Understanding and using rendezvous to enhance mobile crowdsourcing applications. *Computer 44*, 6 (2011), 22–28. doi:10.1109/MC.2011.129.

[97] JACOBS, J. *The death and life of great American cities*. Vintage/Random House, New York, 1961.

[98] JANKOWSKI, P., ANDRIENKO, N., AND ANDRIENKO, G. Map-centred exploratory approach to multiple criteria spatial decision making. *International Journal of Geographical Information Science 15*, 2 (2001), 101–127. doi:10.1080/13658810010005525.

[99] JIA, T., AND JIANG, B. Exploring human activity patterns using taxicab static points. *ISPRS International Journal of Geo-Information 1*, 1 (2012), 89–107. doi:10.3390/ijgi1010089.

[100] JOHNSON, S. *Emergence: The connected lives of ants, brains, cities, and software*. Scribner, New York, 2012.

[101] KALI, R., AND REYES, J. The architecture of globalization: A network approach to international economic integration. *Journal of International Business Studies 38*, 4 (2007), 595–620. doi:10.1057/palgrave.jibs.8400286.

[102] KARAGIANNIS, T., LE BOUDEC, J., AND VOJNOVI, M. Power law and exponential decay of intercontact times between mobile devices. *IEEE Transactions on Mobile Computing 9*, 10 (2010), 1377–1390. doi:10.1109/TMC.2010.99.

[103] KAZANDJIEVA, M., LEE, J., SALATHÉ, M., FELDMAN, M., JONES, J., AND LEVIS, P. Experiences in measuring a human contact network for epidemiology research. In *ACM Workshop on Hot Topics in Embedded Networked Sensors (HotEmNets)* (New York, NY, 2010), ACM. doi:10.1145/1978642.1978651.

[104] KEELING, M., AND EAMES, K. Networks and epidemic models. *Journal of the Royal Society Interface 2*, 4 (2005), 295–307. doi:10.1098/rsif.2005.0051.

[105] KEIM, D., KOHLHAMMER, J., ELLIS, G., AND MANSMANN, F. *Mastering the information age: Solving problems with visual analytics*. Eurographics Association, Goslar, Germany, 2011.

[106] KIKER, G., BRIDGES, T., VARGHESE, A., SEAGER, T., AND LINKOV, I. Application of multicriteria decision analysis in environmental decision making. *Integrated environmental assessment and management 1*, 2 (2009), 95–108. doi:10.1897/IEAM_2004a-015.1.

[107] KNIGGE, L., AND COPE, M. Grounded visualization: Integrating the analysis of qualitative and quantitative data through grounded theory and visualization. *Environment and Planning A 38*, 11 (2006), 2021–2037. doi:10.1068/a37327.

[108] KOCHEN, M. *The Small World*. Ablex, Norwood, NJ, 1989.

[109] KOHLHAMMER, J., MAY, T., AND HOFFMANN, M. Visual analytics for the strategic decision making process. In *GeoSpatial Visual Analytics*, R. D. Amicis, R. Stojanovic, and G. Conti, Eds. Springer, 2009, pp. 299–310. doi:10.1007/978-90-481-2899-0_23.

[110] KOHLHAMMER, J., AND ZELTZER, D. DCV: A decision-centered visualization system for time-critical applications. In *IEEE International Conference on Systems, Man and Cybernetics.* (2003), pp. 3905–3911. doi:10.1109/ICSMC.2003.1244498.

[111] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is Twitter, a social network or a news media? In *Proc. 19th International Conference on World Wide Web* (New York, NY, 2010), ACM, pp. 591–600. doi:10.1145/1772690.1772751.

[112] KWAN, M. Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic geography 75*, 4 (1999), 370–394. doi:10.2307/144477.

[113] KWAN, M. Feminist visualization: Re-envisioning GIS as a method in feminist geographic research. *Annals of the Association of American Geographers 92*, 4 (2002), 645–661. doi:10.1111/1467-8306.00309.

[114] KWAN, M. GIS methods in timegeographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography 86*, 4 (2004), 267–280. doi:10.1111/j.0435-3684.2004.00167.x.

[115] KWAN, M. Mobile communications, social networks, and urban travel: Hypertext as a new metaphor for conceptualizing spatial interaction. *The Professional Geographer 59*, 4 (2007), 434–446. doi:10.1111/j.1467-9272.2007.00633.x.

[116] KWAN, M., AND LEE, J. Geovisualization of human activity patterns using 3D GIS: A time-geographic approach. In *Spatially integrated social science*, M.F.Goodchild and D.G.Janelle, Eds. Oxford University Press, New York, 2004, pp. 48–66.

[117] LANDWEHR, C., BONEH, D., MITCHELL, J., BELLOVIN, S., LANDAU, S., AND LESK, M. Privacy and cybersecurity: The next 100 years. *Proceedings of the IEEE 100*, 13 (2012), 1659–1673. doi:10.1109/JPROC.2012.2189794.

[118] LARSEN, J., AXHAUSEN, K., AND URRY, J. Geographies of social networks: meetings, travel and communications. *Mobilities 1*, 2 (2006), 261–283. doi:10.1080/17450100600726654.

[119] LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABÁSI, A., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., AND GUTMANN, M. Life in the network: The coming age of computational social science. *Science 323*, 5915 (2009), 721–723. doi:10.1126/science.1167742.

[120] LEE, J., AND KWAN, M. Visualisation of socio-spatial isolation based on human activity patterns and social networks in spacetime. *Tijdschrift voor economische en sociale geografie 102*, 4 (2011), 468–485. doi:10.1111/j.1467-9663.2010.00649.x.

[121] LEE, K., HONG, S., KIM, S., RHEE, I., AND CHONG, S. Slaw: A new mobility model for human walks. In *Proc. INFOCOM* (Rio de Janeiro, 2009), IEEE, pp. 855–863. doi:10.1109/INFCOM.2009.5061995.

[122] LEITNER, H., SHEPPARD, E., AND SZIARTO, K. The spatialities of contentious politics. *Transactions of the Institute of British Geographers 33*, 2 (2008), 157–172. doi:10.1111/j.1475-5661.2008.00293.x.

[123] LIMAYEM, M., AND DESANCTIS, G. Providing decisional guidance for multicriteria decision making in groups. *Information Systems Research 11*, 4 (2000), 386–401. doi:10.1287/isre.11.4.386.11874.

[124] LIMTANAKOOL, N., SCHWANEN, T., AND DIJST, M. Developments in the Dutch urban system on the basis of flows. *Regional Studies 43*, 2 (2009), 179–196. doi:10.1080/00343400701808832.

[125] LOMI, A., AND PALLOTTI, F. Relational collaboration among spatial multipoint competitors. *Social networks 34*, 1 (2012), 101–111. doi:10.1016/j.socnet.2010.10.005.

[126] LU, X., BENGTSSON, L., AND HOLME, P. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences 109*, 29 (2012), 11576–11581. doi:10.1073/pnas.1203882109.

[127] LUO, W., MACEACHREN, A., YIN, P., AND HARDISTY, F. Spatial-social network visualization for exploratory data analysis. In *SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN)* (Chicago, Illinois., 2011), ACM. doi:10.1145/2063212.2063216.

[128] MACEACHREN, A. *How maps work: Representation, visualization and design*. Guildford Press, New York, 1995.

[129] MACEACHREN, A. *How maps work: Representation, visualization, and design*. Guilford Press, 2004.

[130] MACEACHREN, A., AND GANTER, J. A pattern identification approach to cartographic visualization. *Cartographica: The International Journal for Geographic Information and Geovisualization 27*, 2 (1990), 64–81. doi:10.3138/M226-1337-2387-3007.

[131] MACEACHREN, A., ROBINSON, A., JAISWAL, A., PEZANOWSKI, S., SAVELYEV, A., BLANFORD, J., AND MITRA, P. Geo-twitter analytics: Applications in crisis management. In *Proc. 25th International Cartographic Conference* (2011).

[132] MACIEJEWSKI, R., LIVENGOOD, P., RUDOLPH, S., COLLINS, T., EBERT, D., BRIGANTIC, R., CORLEY, C., MULLER, G., AND SANDERS, S. A pandemic influenza modeling and visualization tool. *Journal of Visual Languages & Computing 22*, 4 (2011), 268–278. doi:10.1016/j.jvlc.2011.04.002.

[133] MALCZEWSKI, J. *GIS and multicriteria decision analysis*. John Wiley & Sons, 1999.

[134] MALCZEWSKI, J. GIS-based multicriteria decision analysis: A survey of the literature. *International Journal of Geographical Information Science 20*, 7 (2006), 703–726. doi:10.1080/13658810600661508.

[135] MALCZEWSKI, J. Multicriteria decision analysis for collaborative GIS. In *Collaborative geographic information systems*, S. Balram and S. Dragićević, Eds. Idea Group, 2006. doi:10.4018/978-1-59140-845-1.ch010.

[136] MANGOLD, W., AND FAULDS, D. Social media: The new hybrid element of the promotion mix. *Business horizons 52*, 4 (2009), 357–365. doi:10.1016/j.bushor.2009.03.002.

[137] MAO, L., AND BIAN, L. Spatial-temporal transmission of influenza and its health risks in an urbanized area. *Computers, Environment and Urban Systems 34* (2010), 204–215. doi:10.1016/j.compenvurbsys.2010.03.004.

[138] MAO, L., AND BIAN, L. Agent-based simulation for a dual-diffusion process of influenza and human preventive behavior. *International Journal of Geographical Information Science 25*, 9 (2011), 1371–1388. doi:10.1080/13658816.2011.556121.

[139] MASSEY, D. *Space, Place, and Gender*. University of Minnesota Press, Minneapolis, 1994.

[140] MASSEY, D. Geographies of responsibility. *Geografiska Annaler: Series B, Human Geography 86*, 1 (2004), 5–18. doi:10.1111/j.0435-3684.2004.00150.x.

[141] MCANDREW, D. The structural analysis of criminal networks. In *The social psychology of crime: Groups teams and networks*, D. V. Canter and L. J. Alison, Eds. Aldershot, Dartmouth, 1999.

[142] MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001), 415–444. doi:10.1146/annurev.soc.27.1.415.

[143] MENNIS, J., AND MASON, M. Social and geographic contexts of adolescent substance use: The moderating effects of age and gender. *Social networks 34*, 1 (2012), 150–157. doi:10.1016/j.socnet.2010.10.003.

[144] MILGRAM, S. The small world problem. *Psychology today 2*, 1 (1967), 60–67.

[145] MILLER, H. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers 94*, 2 (2004), 284–289. doi:10.1111/j.1467-8306.2004.09402005.x.

[146] MORAN, P. The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B (Methodological)*, 10 (1948), 243–251.

[147] ONNELA, J., ARBESMAN, S., GONZÁLEZ, M., BARABÁSI, A., AND CHRISTAKIS, N. Geographic constraints on social network groups. *PLoS ONE 6*, 4 (2011), e16939. doi:10.1371/journal.pone.0016939.

[148] PHAN, D., XIAO, L., YEH, R., AND HANRAHAN, P. Flow map layout. In *Proc. IEEE Symposium on Information Visualization (INFOVIS)* (2005), IEEE, pp. 219–224. doi:10.1109/INFVIS.2005.1532150.

[149] PRAGER, S. Complex networks for representation and analysis of dynamic geographies. In *Understanding dynamics of geographic domains*, K. Hornsby and M. Yuan, Eds. CRC, 2008, ch. 3, pp. 31–48.

[150] PRECIADO, P., SNIJDERS, T., BURK, W., STATTIN, H., AND KERR, M. Does proximity matter? Distance dependence of adolescent friendships. *Social networks 34*, 1 (2012), 18–31. doi:10.1016/j.socnet.2011.01.002.

[151] RADIL, S., FLINT, C., AND TITA, G. Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in los angeles. *Annals of the Association of American Geographers 100*, 2 (2010), 307–326. doi:10.1080/00045600903550428.

[152] RAE, A. From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems 33*, 3 (2009), 161–178. doi:10.1016/j.compenvurbsys.2009.01.007.

[153] READ, J., EAMES, K., AND EDMUNDS, W. Dynamic social networks and the implications for the spread of infectious disease. *Journal of the Royal Society Interface 5*, 26 (2008), 1001–1007. doi:10.1098/rsif.2008.0013.

[154] REDA, K., TANTIPATHANANANDH, C., BERGER-WOLF, T., LEIGH, J., AND JOHNSON, A. SocioScape—A tool for interactive exploration of spatio-temporal group dynamics in social networks. In *Proc. IEEE Information Visualization Conference (INFOVIS)* (2009).

[155] RHEE, I., LEE, K., HONG, S., KIM, S., AND CHONG, S. Demystifying the Levy-walk nature of human walks. Tech. rep., North Carolina State University, 2008.

[156] RIVEIRO, M. *Visual analytics for maritime anomaly detection*. PhD thesis, Örebro universitet, 2011.

[157] ROCK, M. *Splintering Beijing: Socio-spatial Fragmentation, Commodification and Gentrification in the Hutong Neighborhoods of "old" Beijing*. PhD thesis, Pennsylvania State University, University Park, 2012.

[158] ROSS, C. Neighborhood disadvantage and adult depression. *Journal of Health and Social Behavior 41*, 2 (2000), 177–187. doi:10.2307/2676304.

[159] SAILER, K., AND MCCULLOH, I. Social networks and spatial configuration—How office layouts drive social interaction. *Social networks 34*, 1 (2012), 47–58. doi:10.1016/j.socnet.2011.05.005.

[160] SALATHÉ, M., AND JONES, J. Dynamics and control of diseases in networks with community structure. *PLoS Computational Biology 6*, 4 (2010), e1000736. doi:10.1371/journal.pcbi.1000736.

[161] SALATHÉ, M., KAZANDJIEVA, M., LEE, J., LEVIS, P., FELDMAN, M., AND JONES, J. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences 107*, 51 (2010), 22020–22025. doi:10.1073/pnas.1009094108.

[162] SALATHÉ, M., AND KHANDELWAL, S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology 7*, 10 (2011), e1002199. doi:10.1371/journal.pcbi.1002199.

[163] SANDSTRÖM, A., AND CARLSSON, L. The performance of policy networks: The relation between network structure and network performance. *Policy Studies Journal 36*, 4 (2008), 497–524. doi:10.1111/j.1541-0072.2008.00281.x.

[164] SCHAEFER, D. Youth co-offending networks: An investigation of social and spatial effects. *Social networks 34*, 1 (2012), 141–149. doi:10.1016/j.socnet.2011.02.001.

[165] SHEN, Z., AND MA, K. Mobivis: A visualization system for exploring mobile data. In *Proc. PacificVIS '08* (Kyoto, 2008), pp. 175–182. doi:10.1109/PACIFICVIS.2008.4475474.

[166] SHEPPARD, E. The spaces and times of globalization: Place, scale, networks, and positionality. *Economic geography 78*, 3 (2002), 307–330. doi:10.1111/j.1944-8287.2002.tb00189.x.

[167] SLINGSBY, A., BEECHAM, R., AND WOOD, J. Visual analysis of social networks in space and time. In *Nokia Data Challenge Workshop, Pervasive 2012* (Newcastle, UK, 2012). doi:10.1016/j.pmcj.2013.07.002.

[168] SLOCUM, T. *Thematic Cartography and Geovisualization*. Prentice Hall, Upper Saddle River, NJ, 1999.

[169] STAEHELI, L. Place. In *A companion to political geography*, J. Agnew, K. Mitchell, and G. Tuathail, Eds. Wiley-Blackwell, Malden, MA, 2003, pp. 158–170.

[170] STANLEY, M., AMARAL, L., BULDYREV, S., HAVLIN, S., LESCHHORN, H., MAASS, P., SALINGER, M., AND STANLEY, H. Scaling behaviour in the growth of companies. *Nature 379*, 6568 (1996), 804–806. doi:10.1038/379804a0.

[171] STROGATZ, S. *Sync: The emerging science of spontaneous order*. Theia, New York, 2003.

[172] SUI, D. Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers 94*, 2 (2004), 269–277. doi:10.1111/j.1467-8306.2004.09402003.x.

[173] TAKHTEYEV, Y., GRUZD, A., AND WELLMAN, B. Geography of twitter networks. *Social networks 34*, 1 (2012), 73–81. doi:10.1016/j.socnet.2011.05.006.

[174] THIEMANN, C. SPaTo visual explorer. RoCS, Northwestern University, 2005. http://www.spato.net/.

[175] THIEMANN, C., THEIS, F., GRADY, D., BRUNE, R., AND BROCKMANN, D. The structure of borders in a small world. *PLoS ONE 5*, 11 (2010), e15422. doi:10.1371/journal.pone.0015422.

[176] THOMAS, J., AND COOK, K. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society, 2005.

[177] THOMAS, J., AND COOK, K. A visual analytics agenda. *IEEE Computer Graphics and Applications 26*, 1 (2006), 10–13. doi:10.1109/MCG.2006.5.

[178] TITA, G., AND RADIL, S. Spatializing the social networks of gangs to explore patterns of violence. *Journal of Quantitative Criminology 27*, 4 (2011), 521–545. doi:10.1007/s10940-011-9136-8.

[179] TITA, G., RILEY, K., RIDGEWAY, G., GRAMMICH, C., AND ABRAHAMSE, A. *Reducing gun violence: Results from an intervention in east Los Angeles*. Rand Corporation, 2011.

[180] TOBLER, W. A computer movie simulating urban growth in the Detroit region. *Economic geography 46* (1970), 234–240. doi:10.2307/143141.

[181] TOBLER, W. Experiments in migration mapping by computer. *Cartography and Geographic Information Science 14*, 2 (1987), 155–163. doi:10.1559/152304087783875273.

[182] TOBLER, W. Flow mapper tutorial, 2007. http://www.csiss.org/clearinghouse/FlowMapper/FlowTutorial.pdf.

[183] TOOLE, J., CHA, M., AND GONZÁLEZ, M. Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE 7*, 1 (2012), e29528. doi:10.1371/journal.pone.0029528.

[184] TSOU, M.-H., AND YANG, J.-A. Spatial analysis of social media content (tweets) during the 2012 US Republican Presidential Primaries. In *Proc. GIScience* (Columbus, Ohio, 2012).

[185] URRY, J. Social networks, travel and talk. *The British Journal of Sociology 54*, 2 (2003), 155–175. doi:10.1080/0007131032000080186.

[186] URRY, J., DINGWALL, R., GOUGH, I., ORMEROD, P., MASSEY, D., SCOTT, J., AND THRIFT, N. What is "social" about social science? *Twenty-First Century Society: Journal of the Academy of Social Sciences 2*, 1 (2007), 95–119. doi:10.1080/17450140601108924.

[187] VALENTE, T. *Social Networks and Health: Models, Methods, and Applications*. Oxford University Press, Oxford, UK, 2010.

[188] VERDERY, A., ENTWISLE, B., FAUST, K., AND RINDFUSS, R. Social and spatial networks: Kinship distance and dwelling unit proximity in rural Thailand. *Social networks 34*, 1 (2012), 112–127. doi:10.1016/j.socnet.2011.04.003.

[189] VESPIGNANI, A. Predicting the behavior of techno-social systems. *Science 325*, 5939 (2009), 425–428. doi:10.1126/science.1171990.

[190] WANG, J., CHRISTAKOS, G., HAN, W., AND MENG, B. Data-driven exploration of spatial pattern-time process-driving forces' associations of SARS epidemic in Beijing, China. *Journal of Public Health 30*, 3 (2008), 234–244. doi:10.1093/pubmed/fdn023.

[191] WANG, X., DOU, W., CHEN, S., RIBARSKY, W., AND CHANG, R. An interactive visual analytics system for bridge management. *Computer Graphics Forum 29*, 3 (2010), 1033–1042. doi:10.1111/j.1467-8659.2009.01708.x.

[192] WATTS, D. *Small worlds: The dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, 2003.

[193] WATTS, D. *Six degrees: The science of a connected age*. WW Norton & Company, 2004.

[194] WATTS, D., AND STROGATZ, S. Collective dynamics of "small-world" networks. *Nature 393*, 6684 (1998), 440–442. doi:10.1038/30918.

[195] WELLMAN, B. Little boxes, glocalization, and networked individualism. In *Digital Cities II. Computational and Sociological Approaches*, M. Tanabe, P. van den Besselaar, and T. Ishida, Eds. Springer, Berlin, 2002, pp. 10–25. doi:10.1007/3-540-45636-8_2.

[196] WHITE, J., AND ROTH, R. TwitterHitter: Geovisual analytics for harvesting insight from volunteered geographic information. In *Proc. GIScience* (Zurich, Switzerland, 2010).

[197] WINEMAN, J., KABO, F., AND DAVIS, G. Spatial and social networks in organizational innovation. *Environment and Behavior 41*, 3 (2009), 427–442. doi:10.1177/0013916508314854.

[198] WINNER, M. *Thinking about you thinking about me*. Michelle Garcia Winner, San Jose, CA, 2002.

[199] WOOD, J., DYKES, J., AND SLINGSBY, A. Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal 47*, 2 (2010), 117–129. doi:10.1179/000870410X12658023467367.

[200] WOOLLEY-MEZA, O., THIEMANN, C., GRADY, D., LEE, J., SEEBENS, H., BLASIUS, B., AND BROCKMANN, D. Complexity in human transportation networks: A comparative analysis of worldwide air transportation and global cargo-ship movements. *The European Physical Journal B—Condensed Matter and Complex Systems 84*, 4 (2011), 1–12. doi:10.1140/epjb/e2011-20208-9.

[201] XU, J., AND CHEN, H. Criminal network analysis and visualization. *Communications of the ACM 48*, 6 (2005), 100–107. doi:10.1145/1064830.1064834.

[202] XU, Z., AND SUI, D. Small-world characteristics on transportation networks: A perspective from network autocorrelation. *Journal of Geographical Systems 9*, 2 (2007), 189–205. doi:10.1007/s10109-007-0045-1.

[203] YATES, D., AND PAQUETTE, S. Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International Journal of Information Management 31*, 1 (2011), 6–13. doi:10.1016/j.ijinfomgt.2010.10.001.

[204] YAU, N. Facebook worldwide friendships mapped, 2010. http://flowingdata.com/2010/12/13/facebook-worldwide-friendships-mapped/.

[205] YI, J., KANG, Y., STASKO, J., AND JACKO, J. Understanding and characterizing insights: How do people gain insights using information visualization? In *Proc. Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV 08)* (New York, NY, USA, 2008), ACM, pp. 1–6. doi:10.1145/1377966.1377971.

[206] ZHOU, M. Substitution and stratification: The interplay between dyadic and systemic proximity in global trade, 1993—2005. *The Sociological Quarterly 54*, 2 (2013), 302–334. doi:10.1111/tsq.12027.

[207] ZHOU, M., AND PARK, C. The cohesion effect of structural equivalence on global bilateral trade, 1948—2000. *International Sociology 27*, 4 (2012), 502–523. doi:10.1177/0268580912443577.

Leveraging Geospatially-Oriented Social Media Communications in Disaster Response

Susannah McClendon
GeoVISTA Center
Department of Geography
The Pennsylvania State University
susantbm@gmail.com

Anthony C. Robinson
GeoVISTA Center
Department of Geography
The Pennsylvania State University
arobinson@psu.edu

ABSTRACT

Geospatially-oriented social media communications have emerged as a common information resource to support crisis management. Our research compares the capabilities of two popular systems used to collect and visualize such information - Project Epic's Tweak the Tweet (TtT) and Ushahidi. Our research uses geospatially-oriented social media gathered by both projects during recent disasters to compare and contrast the frequency, content, and location components of contributed information to both systems. We compare how data was gathered and filtered, how spatial information was extracted and mapped, and the mechanisms by which the resulting synthesized information was shared with response and recovery organizations. In addition, we categorize the degree to which each platform in each disaster led to actions by first responders and emergency managers. Based on the results of our comparisons we identify key design considerations for future social media mapping tools to support crisis management.

Keywords

Geographic Information, Social Media, Crisis Management, Mashups.

INTRODUCTION

Crowd-sourced information has rapidly become an essential source of data in disaster response. Since the first well documented efforts of citizen journalists on September 11th, 2001 and the use of internet blogs to collect information after the 2004 East Indian Ocean Tsunami, recent emergency response efforts have included mapping SMS messages after the Haiti earthquake in 2010. The Haiti earthquake represented a paradigm shift in the use of social media for disaster response, as multiple web-based platforms emerged to collect, refine, and disseminate crisis-related social media. The use of social media to gain real time information on the ground in a disaster has been driven by the rapid speed at which information can be distributed, the cross-platform accessibility of information, and the ubiquity of social media worldwide (Vieweg, et al., 2010). The utility of this information has been enhanced by the creation of crisis maps based on location data extracted from social media communications (Liu and Palen, 2010, MacEachren, et al., 2011).

In 2011 the American Red Cross conducted a survey that showed that 33% of citizens have used social media sites, including Facebook, Twitter, Flickr and SMS text messages/alerts to gain information about an emergency (American Red Cross, 2011). About half the respondents said they would contribute information during an emergency using social media channels. Statistics from the International Telecommunication Union reveal that in 2009 there were 4.6 billion mobile phone subscribers world-wide and 1.5 billion subscribers used mobile devices to access the internet (International Telecommunication Union, 2011). We can reasonably expect the use of social media in disaster response to increase in the future.

Research on the increasing use of social media in disaster response has emerged as a new focus in the field of crisis informatics (Anderson and Schram, 2011). Extracting, categorizing, visualizing, and evaluating such information presents serious research challenges, including the problem of managing and extracting meaningful information from the large volume of contributions, applying the information to decision support workflows, and the development of formal information sharing protocols (Harvard Humanitarian Initiative, 2011). Mapping crowd sourced information in disaster response gained wide-scale media attention after the successful deployment of the Ushahidi Crisis Map during the 2010 Haiti earthquake (Starbird, 2011). There are a number

of specific challenges involved in mapping social media communications, including the extraction of accurate location information, and the application of useful and usable cartographic representations to visually support situational awareness in crises. Research on the integration of Geographic Information Systems (GIS) and crowdsourced information from social media has focused more on the challenges of extracting action items and location information from social media feeds (MacEachren, et al., 2011) and less on the utility of the extracted information and the effectiveness of associated crisis maps to support emergency response.

Our research examines two applications that have leveraged geospatially-oriented social media during recent disasters; the Tweak the Tweet (TtT) project from Project Epic (Starbird and Palen, 2011) and Ushahidi (Okolloh, 2009), both of which have been used to create crisis maps of content collected from social media sources during recent disasters. For each application we examine collected data, information products, and evidence of subsequent response actions for two recent disasters; the 2011 Joplin tornado and 2010 Fourmile Canyon fire for Tweak the Tweet, and the 2010 Haiti earthquake and 2010 Gulf Oil Spill for Ushahidi. Other efforts have used crowdsourced information during recent disasters, including the open source Sahana platform (Currion, et al., 2007) and the collective effort of the Crisis Mappers Network (crisismappers.net). However, Sahana utilized data feeds directly from Ushahidi and TtT, and the Crisis Mappers Network is focused on connecting and empowering crisis collaborators, and does not offer their own specialized technology platform.

We begin with background information, including an overview of geospatially-oriented social media, followed by a brief history of TtT and Ushahidi. Next, we evaluate how each organization collected, processed and geo-located these social media communications and compare and contrast the cartographic representations and reporting capabilities of the resulting crisis maps. We then examine the effectiveness of each application by identifying examples of actionable items used by military, government and non-government organizations that emerged from the use of these crisis maps. Finally, we conclude with key design considerations for future efforts to leverage geospatially-oriented social media in crisis informatics.

GEOSPATIALLY-ORIENTED SOCIAL MEDIA

Many social media sources, including Twitter, allow users to tag reports with coordinates to indicate their location on Earth. This information is easy to process and represent on a map, but does not necessarily represent an "actionable" location. A more substantive challenge is associated with making use of textual descriptions of place (placenames and less-specific geographic features), like those often included in an SMS text message. Placenames in text can be geocoded to assign location coordinates, but placenames are usually associated with irregular areas (for example, the New York City metro area) at least as often as one might ever associate them with a specific coordinate location on earth (the centroid of the legal boundary of New York City). Determining less-specific geographic features is also of critical importance when using geospatially-oriented social media.

Both platforms make use of social media collected from SMS and Twitter messages. SMS or "text messaging" is a short messaging service that allows for the storage and retrieval of short 160 character messages across global cellular telephone networks. Location information is not automatically attachable to SMS data, and must be inferred from the message itself. Twitter is a microblogging service that allows users to post messages up to 140 characters called Tweets via mobile phones or web accessible devices. Twitter users follow other users to see their tweets in a Twitter feed. The Twitter user community has developed linguistic markers to facilitate communication; including the @ symbol to address users (@username); the RT abbreviation to represent a retweet (RT @username); and # or hashtag to indicate keywords. Hashtags allow users to search Twitter feeds or to follow trends (Zappavigna, 2011). Location information can be added to a tweet using a phone's GPS capabilities, and can also be inferred from user profiles and mentions of placenames in messages themselves.

PROJECT EPIC AND TWEAK THE TWEET

Project Epic is research effort at the University of Colorado that aims to improve methods of public information gathering and dissemination during emergency situations. The project's mission is to couple computational methods with behavioral knowledge on how people develop information using social media in crisis situations (Palen, et al., 2010). One Project Epic research project, Tweak the Tweet, was first presented in 2009 as a simple set of standardized communication practices coupled with a technology platform for making sense of crisis Tweets (Starbird, 2011, Starbird and Palen, 2011). TtT asks users to tweet using a crisis specific micro-syntax designed to enable real-time processing of Tweets. TtT features a web-based tool for collecting and visualizing contributed information using the Twitter API to continually-update a database. The categorized information is displayed on a simple map mashup using the Google Maps API.

The TtT micro-syntax is based on primary or main hashtags that can be used in any crisis situation and are designed to indicate the "who, what, and where" of the Twitter message content. For example, #name or #contact can be used to indicate "who"; #need, #shelter, #road, #open, #damaged can be used to indicate "what"; and #loc can used indicate "where". These hashtags are used in conjunction with an event tag to organize the crisis. Event tags can be spontaneously generated during an event, like #joplin or #tornado, or prescribed by the TtT micro-syntax like #4MileFire. Used together, the primary and event hashtags format meaningful machine readable tweets.

TtT was first deployed in the aftermath of the Haiti earthquake in January 2010 with the goal of having responders and agencies on the ground use the syntax. The first deployment did not have any associated mapping functionality, and the micro-syntax was not widely adopted by first responders or the public. Despite that, volunteers from Crisis Commons, TtT and other organizations tweeted or retweeted almost 3000 unique tweets formatted with the TtT syntax (Starbird and Palen, 2011). Since 2009, TtT has added a mapping component to their system design and the application has been deployed for over twenty major crises.

USHAHIDI

Ushahidi began as a non-profit African technology company that was developed to map incidents of violence in Kenya following elections in 2008. Ushahidi's mission is to develop platforms for sharing crisis information and personal narratives (Okolloh, 2009) and has since grown to develop tools to facilitate the democratization of information in broader contexts. The open source software tools developed by Ushahidi automate the collection of incident reports using cellular phones, email, and the web and facilitate the mapping of report locations in an interactive map mashup along with descriptive data to contextualize events.

Ushahidi offers three core products: the Ushahidi Platform, the SwiftRiver Platform, and Crowdmap. The Ushahidi Platform combines interactive mapping with the ability to capture real-time data streams from mobile messaging services and Twitter, and also supports email and web forms. It also provides spatial and temporal views of collected data. The SwiftRiver Platform allows for the real-time filtering and verification of data from these multiple data streams, including the ability to automatically categorize information based on semantic analysis, provide analytics and insight into user relationships and data trends, facilitate information validation and qualification, and it offers an interactive dashboard for monitoring and reporting purposes. Crowdmap is a cloud-hosted solution designed to support rapid launches of both the Ushahidi and SwiftRiver platforms.

Since 2009, deployments of Ushahidi platforms have focused on election monitoring, reporting human rights violations, disease surveillance, wildlife tracking, and disaster response. Though there were several deployments of the Ushahidi platform prior to the 2010 Haiti earthquake, it was the Haiti crisis that brought Ushahidi international attention. Ushahidi adoption since the 2010 Haiti earthquake has seen significant growth.

DATA COLLECTION, MANAGEMENT, AND IDENTIFYING LOCATIONS

Both TtT and Ushahidi utilize technical and manual methods to collect, refine, and add meaning to data. The following sections describe how each platform is designed, how they manage data, and how they derive location information from collected social media reports.

Tweak the Tweet

After the initial launch of TtT during the 2010 Haiti earthquake, TtT refined its aims to promote *crowdfeeding* after analysis of results from the deployment in Haiti highlighted the difficulty of getting the crowd to adopt a micro-syntax for Twitter (Starbird and Palen, 2011). TtT promotes the monitoring of social media sites by volunteers (called voluntweeters) during a crisis and they disseminate information back into the crowd using the TtT micro-syntax. In addition, voluntweeters promote the use of the syntax through conversations with other responding organization volunteers and by posting instructions and links to TtT crisis maps on social media sites. For the events we researched, TtT prescribed a micro-syntax with event tags including #boulderfire, #boulder, #4MileFire, #joplin, and #tornado.

The TtT software platform utilizes the Twitter Streaming API to identify tweets based on the TtT micro-syntax and stores the tweets in a MySQL database, parsing the information into key-value pairs based on hashtags. The platform uses Google maps to map the tweet content after the location has been determined. At regular intervals a Ruby script parses the messages filtered by hashtags into a MySQL database and the script in turn updates a public Google spreadsheet. Because machine processing may miss meaningful data in the tweet, such as placenames and other locations, the TtT process uses a combination of automatic and manual processing by

volunteers to populate data in the event spreadsheet. For this research we downloaded spreadsheets for the 2010 Boulder Fourmile Canyon fire and the 2011 Joplin tornado disasters. These TtT–hosted spreadsheets contain all the events collected for each disaster.

Ushahidi

In the hours following the 2010 Haiti earthquake, Ushahidi staff deployed the Ushahidi Haiti Crisis Map. Working in the United States, they gathered information from media reports and social media sources. Approximately 85% of Haitians had access to cellular telephones and the cellular telephone infrastructure, though damaged, was quickly repaired. Within days a SMS short code number was set up in collaboration with phone companies and U.S. State Department resources and advertised through local radio stations. The messages received via SMS were sent to an automated system set up to facilitate message translation and mapping of the data by volunteers (Heinzelman and Waters, 2010).

Shortly before the 2010 Gulf Oil Spill, students at Tulane University began development of a crisis map to document oil refinery accidents using the Ushahidi platform. On the day the class presented the GIS map, the Deepwater Horizon oil rig exploded in the Gulf of Mexico (Dosemagen, 2010). The Louisiana Bucket Brigade, an environmental organization, worked with Tulane students to launch the Oil Spill Crisis Map to give Gulf residents a chance to contribute information about threats to their community and ecosystem from the oil spill. Data for the map was submitted via SMS, Email, Twitter and web forms. Citizens were encouraged to make a reports based on health issues, wildlife sightings, and other notable impacts they may witness in the region.

The Ushahidi API (Ushahidi, 2011) supports data exchange in XML (Extensible Markup Language) and JSON (JavaScript Object Notation). Ushahidi software supports PHP scripting and is designed to work with MySQL and is usually run on an Apache web server. Ushahidi software can be configured to work with common SMS gateway providers to process and deliver SMS messages, and it can be configured to use the Twitter Streaming API to process Tweets. Data can be exported from MySQL via a PHP script to a Google spreadsheet. The Ushahidi map template is designed with a link to download the raw data in a Google spreadsheet, but because the Ushahidi platform is open source and can be modified by the organizations that deploy the software, not all organizations include the ability to download the data. For the 2010 Haiti earthquake and the 2010 Gulf Oil Spill, we were able to download spreadsheets with data covering six months after the initial incidents.

Streaming Data and Scalability Challenges

Collecting data from social media communications like Twitter and SMS is difficult due to the large datasets that can be generated in a short amount of time. A key challenge has emerged in automating the extraction of useful and actionable data from such sources. In fact, applying structure to content using tweets with a micro-syntax to enhance computational automation was part of the original intent behind the TtT project. Challenges associated with filtering, managing, analyzing and translating large volumes of social media communications are being addressed through ongoing development of The SwiftRiver platform by Ushahidi.

During the first deployment of TtT for the 2010 Haiti earthquake, the syntax was not widely adopted by citizens and first responders, but the syntax was picked up by people who spontaneously volunteer during a crisis (Starbird and Palen, 2011). TtT efforts spurred a network of volunteers that helped give structure to the social media communications that were transpiring on Twitter during both the 2011 Joplin tornado and 2010 Boulder Fourmile Canyon fire crises. These volunteers adopted the TtT syntax and translated information from multiple sources using the syntax before tweeting it out to their followers. These followers were diverse, including media outlets, the American Red Cross, FEMA, and other relief organizations. This type of volunteerism was promoted to direct Twitter communications so that automatic filtering of Tweets would be more effective.

During the 2010 Haiti earthquake, Ushahidi enlisted volunteers to assist with handling the large volume of data. SMS messages began to flow at a rate of 1,000 to 2,000 a day and were passed directly from the cellular telephone provider to an automated system, designed by Ushahidi developers for coordinating volunteers. Volunteers manually translated the messages from Haitian Creole and then filtered and determined locations (Meier and Munro, 2010). The system supported message translation with a lead time of less than ten minutes.

EXTRACTING LOCATION INFORMATION

One of the most challenging aspects of using social media data during a disaster is extracting unambiguous and accurate location information. Locations are essential for determining if a message is actionable (Munro, 2011). Location can be determined in several ways, including processing location references like a place name or street

address in message content; explicit coordinates derived from geo-location services from cellular phones; and extraction of location information in a user's social media account profile (Bellucci, et al., 2010, Field and O'Brien, 2010). Table 1 illustrates examples of profile-derived location information shown in Tweets 5-8.

Tweet	Time	User	Tweet	Location
1	11.21 pm 23 May 2011	@sarahgracesitz	RT @shawncmatthews: Here is a great resource for donation centers #joplin #tornado #relief http://ow.ly/515zC	iPhone: 37.112511,-93.303925
2	11.18 pm 23 May 2011	@CajunTechie	If you want to donate clothing of all sizes to displaced residents in #Joplin #Missouri you can drop it by 702 Moffet #tornado	ÃœT: 36.874136,-94.873582
3	10.07 pm 23 May 2011	@pbdoetmee	Ronduit dramatisch fotowerk na de tornado hit-down in Joplin, Missouri, USA... http://bit.ly/jiw6Kg #indrukwekkend #joplin #tornado	51.953923,6.008155
4	11.35 pm 23 May 2011	@PeterKinder	Just confirmed I will be guest on @IngrahamAngle radio show Tuesday morn 5/24/11 9:30 CDT talking #Joplin #MO #tornado relief, update #pdk	iPhone: 0.000000,0.000000
5	10.48 pm 23 May 2011	@maryfranholm	RT @OzarksRedCross: #CBCO FB site says: CODE RED 4 blood donations 4 the #Joplin #tornado has been lifted. Thanks 2 so many of you donat ...	Lost in a good book
6	10.39 pm 23 May 2011	@Jeannie_Hartley	RT @OzarksRedCross: #RedCross update here: http://bit.ly/jZ3Ivp #Joplin #tornado	Universe
7	10.36 pm 23 May 2011	@wheelertweets	RT @Jeannie_Hartley: #Tornado #Joplin #mo @info4disasters @Redcross @kcredcross @1stAid4 @wheelertweets @jnicky63 @viequesbound @Lady1st ...	Tunis, Tunisia
8	11.02 pm 23 May 2011	@JoplinMoTornado	GOODNEWS: 7 people were rescued from the debris today! #joplin #tornado -G	Joplin, Mo
9	11.02 pm 23 May 2011	@DAOWENS44	RT @OzarksRedCross: #CBCO FB site says: CODE RED 4 blood donations 4 the #Joplin #tornado has been lifted. Thanks 2 so many of you donat ...	

Table 1. Tweet Examples from the 2011 Joplin Tornado.

Location information included within a Twitter or SMS message as a text reference (e.g. a user mentions a specific place by name) must be extracted and geocoded to obtain coordinate information. Latitude and Longitude coordinates can also be included with messages as geospatial metadata. The process of manually or computationally assigning such metadata is called geo-tagging. Twitter included the ability to geo-tag tweets in 2009 (Bellucci, et al., 2010). Because of privacy concerns, social media applications and cellular phones usually require users to opt-in to enable geo-tagging. The SMS protocol does not incorporate geospatial metadata and typical messages sent from cellular phones via SMS will not contain location information (Munro, 2011). However, GeoSMS (geosms.wordpress.com), a location-enabled SMS standard, can embed geospatial metadata into a URI (Uniform Resource Identifier). MacEachren, et al. (2011) notes that the proportion of users who enable geo-tagging is still small. However, geo-tagging alone is no guarantee the message content is meaningful. Of three geo-tagged examples in Table 1, two geo-tags are close to Joplin in Springfield, Missouri (1) and the nearby city of Miami, Oklahoma (2), while the location in Tweet 3 is in the Netherlands.

The Ushahidi platform does not contain a mechanism to automatically geocode implicit location information, but the SwiftRiver Platform does incorporate tools that use natural language processing and a gazetteer to return coordinate locations based on place names. The Ushahidi platform will extract geospatial metadata from social media feeds if it exists. For the Ushahidi 2010 Haiti earthquake map, the majority of information gathered came via SMS and not geo-tagged. Location information from SMS messages was translated by volunteers who used a variety of resources to obtain coordinate locations from the translated messages. Many of the volunteers were originally from Haiti and used their own geographical knowledge of the region combined with Open Street Map to pinpoint extract coordinates (Heinzelman and Waters, 2010). For the Ushahidi 2010 Gulf Oil Spill map we were unable to determine which specific methods were used to geo-locate implicit location information.

The software platform used by TtT extracts geospatial metadata using the Twitter API if such metadata exists. The software filters for location tags prescribed in the TtT micro-syntax or tags identifiable as spontaneously generated by the crowd that may include implicit location information, for example #loc or #lat and #long. These tags and the data after each tag were parsed into key-value pairs to populate the database. Location pairs, along with identified place names or event tags like Joplin or Boulder were geocoded using GeoKit (geokit.rubyforge.com), which can geocode textual information across a number of different geocoding services. Volunteers could review the resulting coordinate pairs, which were then entered into the database if approved.

COMPARING USHAHIDI AND TWEAK THE TWEET

Here we draw comparisons between Ushahidi and TtT in three key dimensions. First, we describe what types of data and variables are captured by each effort. Next, we compare the interactive mapping tools that each platform provides. We conclude our comparisons by characterizing how each platform has resulted in tangible actions by responders and emergency managers. We use four recent disasters in these comparisons. For Ushahidi, we explore its use in the 2010 Haiti earthquake and 2010 Gulf Oil Spill. For TtT we focus on the 2011 Joplin tornado and 2010 Boulder Fourmile Canyon fire. We were unable to find directly overlapping events for both platforms. Ushahidi deployments tend to focus on larger disasters rather than the localized events focused on by TtT. For one overlapping event, the 2010 Haiti earthquake, TtT had not yet implemented mapping tools, and in other overlapping events (like 2010 Pakistan floods) TtT has integrated their efforts with Ushahidi.

Field type	Ushahidi Field Name	TtT Field Name Boulder[1] Joplin[2]	Definition - TtT	Definition - Ushahidi
Record ID	#	Record ID[1] / ID[2]	Unique identifier	Unique identifier
Event		Event[1]	Event Hashtag – used for Place location	
Categorization	Category	Report Type[1,2]	Primary Hashtag – only one allowed – used for Key legend in Web Maps	Multiple categories allowed – used for Web Map Category Filter in Legend
Report	Incident Title	Report [2]	Partial parsed tweet with hashtags removed for pop-up display	Report Title for Web Report
Details		Details [1]	Partial parsed tweet with hashtags removed for pop-up display	
Original Report	Description	Text[2]	Original tweet	Original Message (in original language and translated if necessary)
Date/Time Stamp	Incident Date	Time[1,2]	Tweet time stamp	Message time stamp
Date_Time		Date_Time [2]	Time contained in tweet message	
Info		Info[1]	Volunteer added comment	
Source		Source[1,2]	Twitter user	
Contact		Contact[1,2]	Name, number, web page or other contact info contained in tweet	
Completed		Complete[1,2]	Indication if report was acted upon	
Status		Status[1]	? All N/A	
Verification	Verified	Verified[1]	? All N/A	Corroborated via incident report credibility vote
Actionable		Actionable[1]	? All N/A	
Approved	Approved			Map location approved
Author		Tweet Author[1] / Author[2]	The author of the record in the spreadsheet or author of retweet	
Tweet		Tweet[1]	Original Tweet	
Photo URL		Photo URL[1], Photo[2]	URL to photo	
Video		Video [2]	URL to Video	
Location (Text)	Location	Location[1,2]	Parsed location string	Parsed location string
Mapped		Mapped[1]	? All N/A	
Longitude	Longitude	GPS Long[1,2]	Derived Longitude	Derived Longitude
Latitude	Latitude	GPS Lat[1,2]	Derived Latitude	Derived Latitude

Table 2. Comparing data collected from TtT and Ushahidi

Raw Data

Data generated during a disaster from social media networks tend to be ephemeral and if it is not collected during the disaster, it can be difficult to conduct related research after the fact. Collecting raw data from Twitter older than two weeks has become challenging due to changes in the Twitter API that forbid certain types of archiving. Here, we conduct our analysis using the spreadsheets gathered from each application and additional analytical results from the PeopleBrowsr (www.peoplebrowsr.com) service which provides 1000 days of social media content and social analytics for marketers (not including SMS). We did not include PeopleBrowsr analytics for the 2010 Haiti earthquake because that data collected was primarily from SMS, and PeopleBrowsr analytics are not available for the 2010 Boulder Fourmile Canyon fire due to a small number of reports.

In Table 2 we list all fields we discovered in the TtT and Ushahidi spreadsheets and our interpretation of the definitions for each field type for each application. Common fields which we think share a common meaning across both platforms are highlighted. The Ushahidi platform has fewer fields (eight vs. twenty-five for TtT) and

they do not vary between the two incidents. TtT has variation in field names and the number of fields. We note that it is difficult to differentiate between the terms Status, Actionable and Verified in the TtT fields.

In the content summary shown in Table 3 the Ushahidi field "Approved" always shows a rating of 100%. According to Ushahidi documentation all messages are "approved" once valid location coordinates are determined and an administrator approves the content. Reports that are not yet approved are not displayed. The "verified" field indicates a report is submitted by or corroborated by a trusted source or an administrator. Table 3 shows that 6% and 40% of the records were corroborated in the 2010 Haiti earthquake and 2010 Gulf Oil Spill events. Raw data from TtT did not reveal the meaning of the codes "Status", "Actionable", and "Verified."

The 2010 Gulf Oil Spill Ushahidi spreadsheet lists the first incident date eleven days after the Deepwater Horizon explosion. The total number of tweets with the #oilspill keyword from April 10th to October 18th, 2010 according to PeopleBrowsr, is 22,199. The Louisiana Bucket Brigade collected 2952 reports according to their spreadsheet, representing approximately 13% of the total Twitter traffic by the PeopleBrowsr estimate. Of note is that all 2952 reports were geo-located. Additionally, there were only 9 tweets on the day after the explosion and no tweets for the next 17 days. The traffic over six months highlights the extended nature of the disaster.

The 2011 Joplin tornado data starts the day after the tornado and ends 27 days after the tornado. According to PeopleBrowsr there were 333,387 total Twitter mentions of the #Joplin keyword from May 13 to June 13th, 2011. TtT identified 504 tweets that were entered into the spreadsheet. This represents approximately 0.02% of the total Twitter traffic if the PeopleBrowsr estimates are correct. This highlights the challenge associated with harvesting social media communications during temporally-limited crises. It is also interesting to note that 65% of the 504 records in the TtT spreadsheet for the 2011 Joplin tornado and 54% of the 522 records for the 2010 Boulder Fourmile Canyon fire included locations. Examination of the raw data reveals frequent status communication between volunteers that was not mapped because it was not relevant to the event itself.

	Incident	Incident Date	Reports	First Report Date	Last Report Date	%Verified	%Approved	%Actionable	%Complete	%LAT/LONG
TtT	Joplin Tornado[2]	5/22/2011 5:34 PM	504	5/23/2011 12:11 AM	6/13/2011 11:10 PM	N/A[1]		N/A[2]	0.4%[2]	65%
	Boulder Fourmile Fire[1]	9/6/2010 10:00 AM	522	9/8/2010 5:50 PM	9/17/2010 9:33 PM			N/A[1]	N/A[1]	54%
Ushahidi	Haiti Earthquake	1/12/2010 4:53 PM	3589	1/12/2010 4:08 AM	5/18/2010 4:26 PM	6%	100%			100%
	Gulf Oil Spill	4/10/2010 10:00 PM	2952	4/21/2010 1:44 PM	10/18/2010 10:07 PM	40%	100%			100%

Table 3. Summary of Ushahidi and TtT Spreadsheet Content

Maps

Cartographic representation of crisis mapping represents another challenge in the use of social media communications for disaster response because of the need to display large volumes of data while avoiding information overload. This is complicated further by the fact that potential users of crisis maps, including citizens, responders, volunteers, journalists and managers will have different expectations influenced by their social and physical relation to the crisis event (Liu and Palen, 2010). Field and O'Brien (2010) recognize that given the growth of social media communications and the geospatial component integral to an interconnected world, good cartography is crucial for creating maps with a purpose that are more than one-dimensional.

Crisis maps created by Ushahidi and TtT are quite similar (Figure 1) in terms of their core features. Both platforms utilize simple interactive map mashups and categorized point symbols to represent reports. Ushahidi has the ability to generalize dense sets of reports into aggregated symbols, making it scalable to larger datasets. Both platforms have recently introduced temporal displays to highlight report frequency over time (frequency graph in Ushahidi and time-categorized markers in TtT). The overall map and interface aesthetic is significantly more refined in current implementations of Ushahidi, perhaps reflecting its relative maturity compared to TtT.

Neither platform supports significant geospatial analysis capabilities. Basic filtering controls are available to winnow the dataset, but there are no quantitative spatial analysis methods available to identify clusters or to compare current patterns to past patterns. A significant difference between platforms is that Ushahidi provides alerting tools for users to "listen" for reports from a given area or matching a given set of thematic criteria.

Spatial data interoperability in both platforms is supported through spreadsheet downloads of raw data, making it possible for users to ingest collected information into a full-featured GIS if necessary.

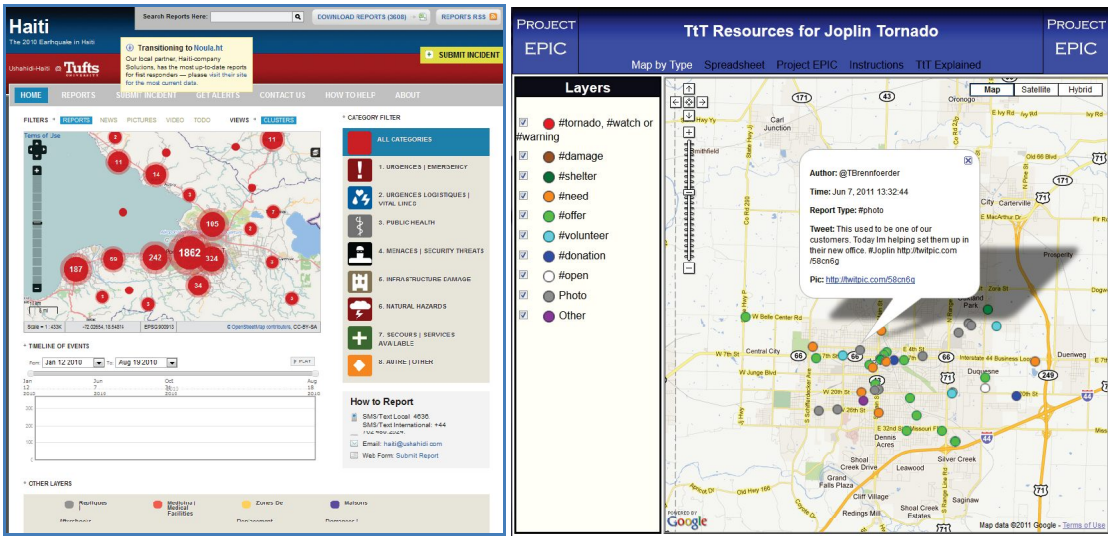


Figure 1. Examples of Ushahidi (left) and Tweak the Tweet (right) Mapping Interfaces

Evidence of Action

Understanding the effectiveness of efforts like TtT and Ushahidi is a difficult task. We concentrated our efforts on identifying impacts from media reports, after action reports, and TtT and Ushahidi's own assessments.

Project Epic and TtT gained recognition for their efforts during the 2010 Boulder Fourmile Canyon fire when CNN ran a story which credited TtT for integrating crisis information through the use of volunteers and social media (Spellman, 2010). One article cites Project Epic's use of a map with geo-located tweets to tracking the fire movements from citizen reports (Orlando, 2011). Another news report (Petty, 2010) describes how TtT was used to gather and map data from the 2010 Boulder Fourmile Canyon fire, including information not provided through official emergency response channels. We did not find many media reports of the specific use of TtT after the 2011 Joplin tornado, but the use of the syntax was promoted by organizations like Crisis Commons.

Following the 2010 Haiti earthquake, Craig Fugate, Director of FEMA, tweeted that the Ushahidi Haiti Map was the most "comprehensive and up-to-date map available to humanitarian organizations" (Heinzelman and Waters, 2010). Newsweek's profile of the Ushahidi efforts after Haiti indicate that the crisis map resulted in saving lives (Ramirez, 2010). One after action report (UN-SPIDER, 2010) notes that the US Marines used Ushahidi to coordinate locations and direct relief efforts. They also indicate that data from Ushahidi was used to direct Coast Guard responders for search and rescue. The Ushahidi Blog (blog.ushahidi.com) highlights multiple examples of action items and response efforts generated by Ushahidi, including food and water deliveries.

The 2010 Gulf Oil Spill Ushahidi map represented a different use of this technology in disaster response. The oil spill was not as much a direct threat to lives as it was a threat to local economies and the environment. Use of the map has been primarily to raise awareness of the ongoing ecological disaster and to document the damage. One article highlighted the fact that the information collected would be useful in long-term recovery efforts and could also be used in future legal actions over long-term damage to ecosystems and livelihoods (Sutter, 2010).

A common thread throughout the reports we reviewed is that the impact of geospatial social media platforms on tangible emergency response actions is not yet well-defined. While both have received media attention and have clearly captured public interest, there are few specific examples of the information leading to different decision-making patterns, widespread allocation of resources, or information leading to the rescue of disaster victims.

Future Design Considerations and Conclusions

The recent use of social media communications in disaster response is largely driven by volunteer organizations. Engagement with volunteers during the 2010 Haiti earthquake by TtT prompted TtT developers to focus research on a second layer of crowdsourcing: communication between volunteers and response organizations (Starbird, 2011). By working with crisis volunteer organizations TtT has continued to promote the use of the micro-syntax after a disaster. Spurred by lessons learned from deployments of crisis mapping efforts like Ushahidi, the Standby Task Force was recently formed to organize volunteers and provide a dedicated technical interface to the humanitarian community to assist in dealing with new sources of information like social media.

A key design consideration going forward is to ensure effective mechanisms for disseminating and sharing information between first responders and crisis managers. A UN report suggests that efforts like Ushahidi and TtT may contribute to information overload, citing that a large percentage of the information gathered was already in the hands of relief organizations on the ground (UN-SPIDER, 2010). This does not take into consideration the value of mapping the information and providing crisis managers a mechanism to identify clusters where relief organizations and first responders could concentrate their efforts.

We have outlined several reasons why extracting the location from social media is difficult, and further research is needed in this area, especially in the role of automatic geocoding and disambiguation of text descriptions of place. Ushahidi point out that from 40,000 Haiti-related SMS messages only 5% were mapped (Norheim-Hagtun and Meier, 2010), but that does not mean that the other 95% did not contain any geospatial information. To support analytical reasoning and geospatial analysis we should be able to uncover patterns that reference physical and cultural regions, types of landforms, directional information and topological relations in addition to basic point locations. A range of recent research focuses on the challenges of extracting actionable information from social media. Munro (2011) proposes models to systematically identify actionable items using trending categories or topics, subwords, and spatiotemporal clusters. MacEachren, et al. (2011) discuss how SensePlace2, a geovisual analytics application, includes a crawler application designed to systematically query the Twitter API based on crisis-relevant keywords and phrases. Vieweg, et al. (2010) have recently described an automated methodology to detect messages that will be useful for situational awareness.

Credibility and verification of information is another area that needs to be addressed in future research. One report indicates that search teams found a high proportion of SMS reports about trapped wounded victims turned out to be coming from families wanting to recover their dead relatives (Harvard Humanitarian Initiative, 2011). Some recent research has focused on identifying ways in which credibility might be automatically assessed in Tweets by evaluating message content, user profile details, and message propagation (Castillo, et al., 2011).

Finally, we must develop effective cartographic representation techniques to ensure the usability of web maps for crises. A particular challenge for crisis mapping is that there are a wide range of expectations and technical skills associated the diverse group of people that need to use crisis maps, including citizens, responders, volunteers, journalists and managers. Not all groups are equally equipped to evaluate the results of geographical analysis. Maps are likely to be seen as credible evidence, even when the underlying data is of unknown quality.

There is no doubt that the contribution of social media communications during disaster has shifted the paradigm of emergency response to include at least a one-way social media dialog from those most affected. Mapping social media content provides a way to gather and visualize information from what can arguably considered the true first responders - the affected citizens who are the first to assess the situation and request assistance through social media. Driven by volunteers and advances in web-based technology, the proliferation of this information has grown faster than the analytical capabilities of disaster management organizations and workflows. TtT has contributed a method to filter, automate and direct information from social media sources during a disaster and Ushahidi has proven to be an effective and widely adoptable platform for displaying geospatially-oriented social media communications. However, TtT and Ushahidi have only tackled simple location-related problems and provided only rudimentary situational awareness and mapping capabilities to visualize the social media communication stream. Future research must focus on applications that go beyond basic crowdsourcing to develop information collections, analytical tools, coordination of communications, and mapping visualization to support all phases of disaster management. Future platforms developed with the volunteer community in mind will need to incorporate social media as one piece of an overall strategy to support situational awareness and response and recovery featuring effective two-way communications with citizens through social media.

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number: 2009-ST-061-CI0001. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies of the U.S. Department of Homeland Security.

REFERENCES

1. S. Vieweg, A.L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in, ACM, Atlanta, Georgia, USA, 2010, pp. 1079-1088.
2. S.B. Liu, L. Palen, The new cartographers: crisis map mashups and the emergence of neogeographic practice, Cartography and Geographic Information Science, 37 (2010) 69-90.

3. A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford, SensePlace2: Geotwitter Analytics Support for Situation Awareness, in: IEEE Conference on Visual Analytics Science and Technology, Providence, RI, 2011.
4. American Red Cross, Social Media in Disasters, in: http://www.redcross.org/www-files/Documents/pdf/SocialMediainDisasters.pdf, accessed on 10/15/2011
5. International Telecommunication Union, Key Global Telecom Indicators for the World Telecommunication Service Sector, in: www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html, accessed on 10/15/2011
6. K.M. Anderson, A. Schram, Design and implementation of a data analytics infrastructure in support of crisis informatics research (NIER track), in, ACM, Waikiki, Honolulu, HI, USA, 2011, pp. 844-847.
7. Harvard Humanitarian Initiative, Disaster Relief 2.0: The future of information sharing in humanitarian emergencies, in, 2011.
8. K. Starbird, Digital Volunteerism During Disaster: Crowdsourcing Information Processing, in: CHI '11 Workshop on Crowdsourcing and Human Computation, Vancouver, BC, 2011.
9. K. Starbird, L. Palen, "Voluntweeters": self-organizing by digital volunteers in times of crisis, in, ACM, Vancouver, BC, Canada, 2011, pp. 1071-1080.
10. O. Okolloh, Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information, Participatory Learning and Action, 59 (2009) 65-70.
11. P. Currion, C. de Silva, B. Van De Walle, Open source software for disaster management, Communications of the ACM, 50 (2007) 61-65.
12. M. Zappavigna, Ambient affiliation: A linguistic perspective on Twitter, New Media & Society, 13 (2011) 788-806.
13. L. Palen, K.M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, D. Grunwald, A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters, in: ACM-BCS Visions of Computer Science Conference, Edinburgh, UK, 2010, pp. 1-12.
14. J. Heinzelman, C. Waters, Crowdsourcing Crisis Information in Disaster-Affected Haiti, United States Institute of Peace, 252 (2010) 1-16.
15. S. Dosemagen, Ushahidi Used to Create Oil Spill Crisis Map, in: The Ushahidi Blog, 2010.
16. Ushahidi, Ushahidi API, in: http://wiki.ushahidi.com/doku.php?id=ushahidi_api, accessed on 10/15/2011
17. P. Meier, R. Munro, The Unprecedented Role of SMS in Disaster Response: Learning from Haiti, SAIS Review, 30 (2010) 91-103.
18. R. Munro, Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol, in, Association for Computational Linguistics, Portland, Oregon, 2011, pp. 68-77.
19. A. Bellucci, A. Malizia, P. Diaz, I. Aedo, Framing the design space for novel crisis-related mashups: the eStoryS example, in: Information Systems for Crisis Management and Response (ISCRAM 2010), Seattle, WA, 2010, pp. 1-10.
20. K. Field, J. O'Brien, Cartoblography: experiments in using and organising the spatial context of micro-blogging, Transactions in GIS, 14 (2010) 5-23.
21. J. Spellman, Heading off disaster, one tweet at a time, in: CNN Tech, www.cnn.com/2010/TECH/social.media/09/22/natural.disasters.social.media/index.html,accessed on 10/15/2011
22. J. Orlando, Turning Disaster Response On Its Head, in: Continuity Insights, http://www.continuityinsights.com/articles/turning-disaster-response-on-its-head, accessed on 10/15/2011
23. D. Petty, Evacuees use social media to keep up on Boulder wildfire disaster developments, in: Denver Post, MediaNews, Denver, CO, 2010.
24. J. Ramirez, 'Ushahidi' Technology Saves Lives in Haiti and Chile, in: Newsweek, 2010.
25. UN-SPIDER, Lessons from Haiti, Coordinates, 6 (2010) 27-31.
26. J.D. Sutter, Citizens monitor Gulf Coast after oil spill, in: CNN Tech, articles.cnn.com/2010-05-06/tech/crowdsource.gulf.oil_1_oil-spill-gulf-coast-jeffrey-warren, accessed on 10/15/2011
27. I. Norheim-Hagtun, P. Meier, Crowdsourcing for Crisis Mapping in Haiti, innovations, 5 (2010) 81-89.
28. C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in, ACM, Hyderabad, India, 2011, pp. 675-684.

Sharing and Discovering Map Symbols with SymbolStore.org

Anthony C Robinson, Scott Pezanowski, Joshua Stevens, Ryan Mullins, Justine Blanford, Raechel Bianchetti, Alan M. MacEachren

GeoVISTA Center
Department of Geography
The Pennsylvania State University

Abstract. Maps are often used to support each phase of emergency management activities, including disaster planning, response activities, and long-term recovery efforts. While there are many symbol standards for emergency management, interoperable map designs remain elusive for this domain. Informal symbol conventions are frequently applied by emergency management mapmakers in place of more formal standards. And until now there have been few flexible mechanisms for discovering, sharing, and previewing these symbol sets among mapmakers. Here we highlight the Symbol Store, a web-based interactive tool intended to help mapmakers discover, share, and preview point symbols. The Symbol Store allows users to browse for symbols by keywords, category tags, and contributors. It also allows users to preview symbols on realistic maps prior to download. Moving forward, we are focused on further fostering the development of refined symbol sets through the addition of new features to the Symbol Store.

Keywords: Symbology, Map Design, Standardization

1. Introduction

One of the most critical cartographic challenges is to determine the means for representing geographic features. This task is often supported through the careful selection of ready-made symbols provided in mapping and graphic design software, or through the drafting of new symbols. Developing best practices for symbol designs and their standardization has been a topic of Cartographic research for several decades now (Rado and Dudar 1971; Robinson 1973; Morrison and Forrest 1995). While there has been much progress on how to design symbols, current mechanisms for discovering and sharing symbols tend to rely on symbol distribution through GIS

software and informal sharing through set-specific websites and personal exchanges (Robinson et al. 2011). While a great deal of effort has gone into developing new symbol standards in recent years to support map interoperability in defense, crisis management, and humanitarian mapping domains (Dymon 2003; DOD 2008; Kostelnick et al. 2008), much less progress has been made toward ensuring that symbols can be easily discovered and shared. Here we present our progress toward developing a web-based platform for sharing and refining map symbols that we call the Symbol Store. This work is driven by results from a multi-year investigation of symbol interoperability at the U.S. Department of Homeland Security (DHS). Our prior work focused first on the use of existing standards (Robinson et al. 2011) such as the ANSI 415 symbol set (ANSI 2006), and the development of a new more flexible platform for supporting the creation of new symbol standards (Robinson et al. 2012).

A primary goal for the Symbol Store is to help users who want to search for and quickly retrieve point symbols using keywords, category names, and other metadata. Supporting this core functionality enables the Symbol Store to enhance interoperability in emergency management contexts where authorities work together to create maps to support situational awareness and emergency response actions (Cutter 2003). Emergency management mapping also frequently involves planning activities and long-term recovery efforts which also stand to benefit from improved symbol sharing mechanisms.

While our initial work with the Symbol Store has focused on developing basic methods to support web-based symbol sharing, our current efforts are aimed at integrating an iterative process for standardizing domain specific symbol sets. The following sections outline our progress toward supporting both areas of concern.

2. Sharing Symbols

To support cartographers who wish to easily discover new symbols and share the symbols they have already collected, we developed a web-based platform for sharing symbols called the Symbol Store *Figure 1*. Based on the results of our prior work to study the use of symbol standards by DHS mapmakers and to design a new process to developing new and more flexible symbol standards, we developed four core goals to shape the Symbol Store application:

2.1. Search for and retrieve symbols

A fundamental aim for the Symbol Store is to support easy and efficient keyword searches for symbols that are used cartographers from government agencies as well as the private sector. Our intention is to support rapid retrieval for symbols using basic keyword search techniques. Current methods for finding symbols in GIS software often rely only on the formal names assigned to symbols, rather than their keyword description.

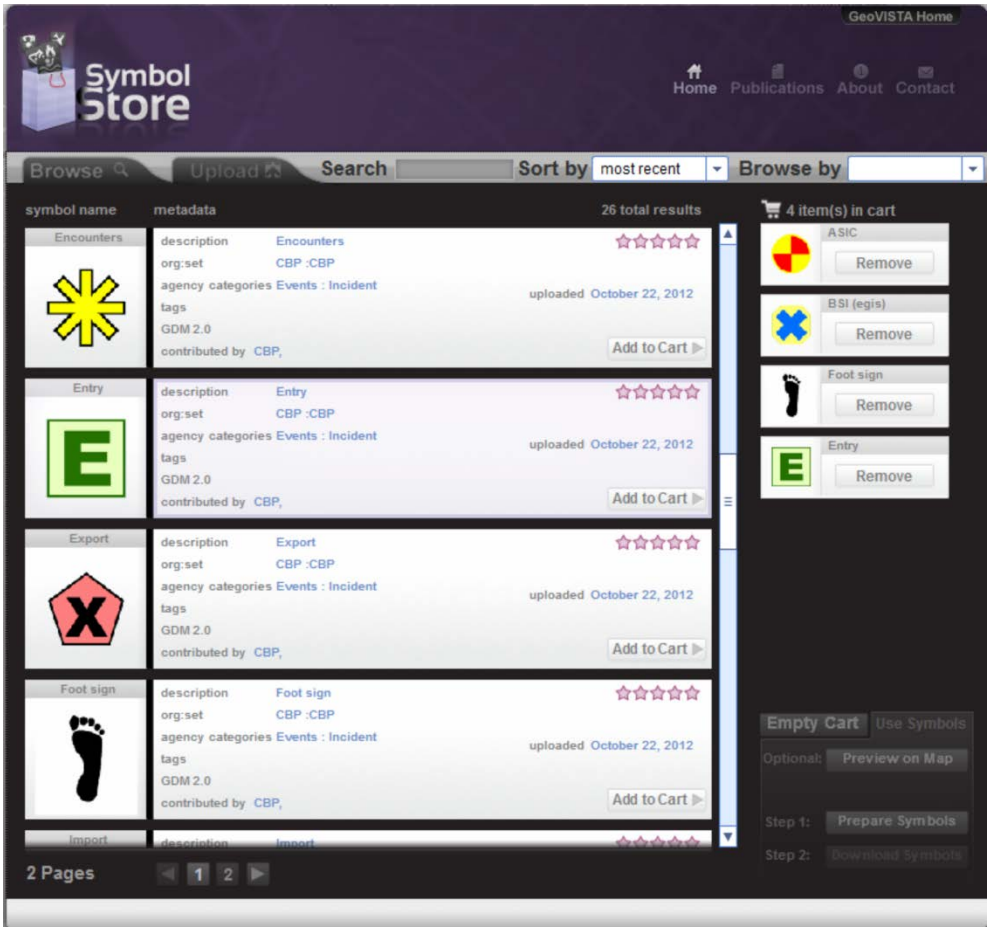


Figure 1. The primary Symbol Store interface for browsing and selecting symbols.

2.2. Preview symbols on realistic maps

Once a user has selected a set of symbols using the Symbol Store search interface, they can preview these symbols using the map preview tools shown here in *Figure 2*. These map preview tools provide a range of controls designed to allow users to change the map scale, feature density, labeling, coloring, and other common map design aspects in order to preview

the suitability of their symbols prior to downloading them. They can also switch between multiple realistic basemaps to see how symbols work (or do not work) in common cartographic design situations.

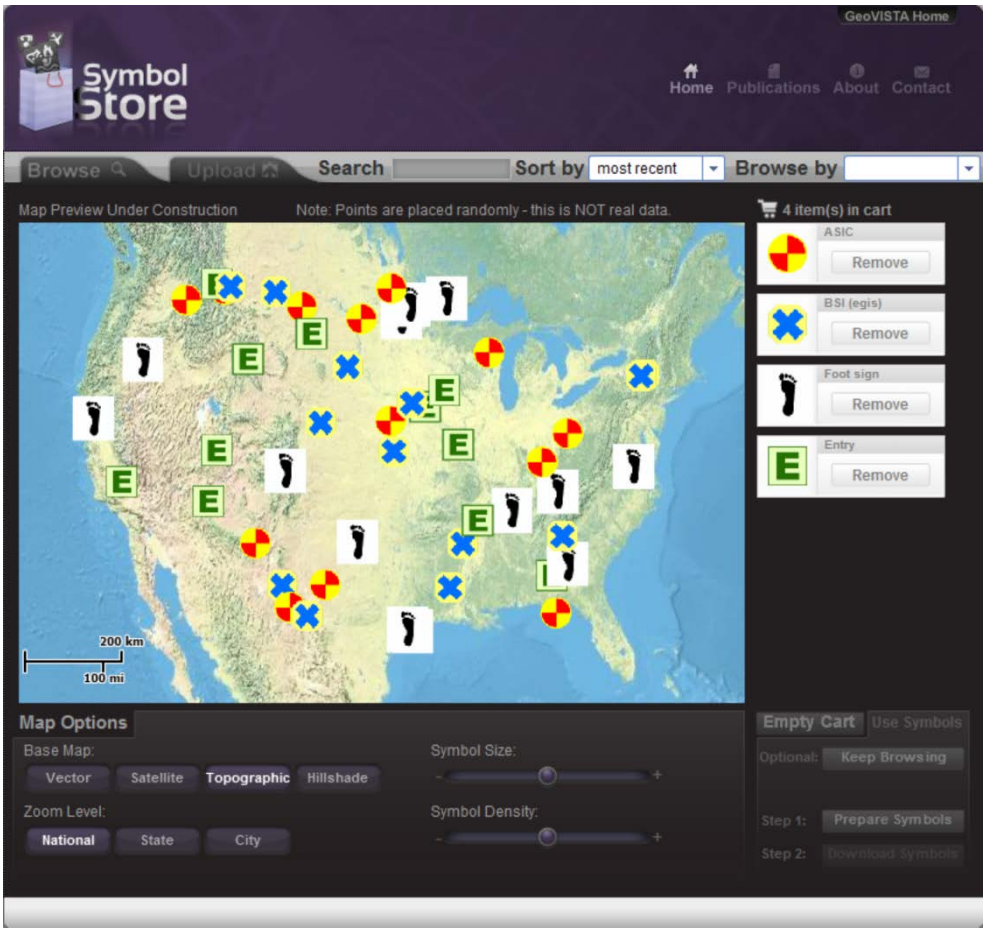


Figure 2. The Symbol Store interactive map preview interface for testing symbols.

2.3. Browse for symbols

In addition to basic search tools to find symbols using keywords, Symbol Store users can browse for symbols by time (show the most recent uploads, for example), contributor (show all symbols designed by a particular agency, for example), and symbol categories (show every symbol that is tagged as part of the category called *infrastructure*, for example). This browsing capability supports iterative and flexible symbol discovery, and it is particularly well-suited for cases in which keyword searches are not possible (e.g. you don't know the exact name for the symbol you want) or efficient (e.g. you know you need several symbols that correspond to infrastructure).

These browsing features also allow popularity measures such as subjective ratings to become one of the means by which new symbols can be discovered and shared. We anticipate this capability to be bolstered over time as increasing numbers of users provide ratings for symbols housed in the Symbol Store.

2.4. Share symbols

Users can contribute symbols to Symbol Store by uploading an Esri *.style file and associated fonts through the Symbol Store interface. Alternatively, they can choose to upload a collection of .PNG or .SVG graphic symbols if they are not commonly using Esri .style files. After uploading symbols, users can tag individual symbols or groups of symbols to assign keywords, category names, and other important metadata information (see Section 3 for details).

The Symbol Store (currently accessible at SymbolStore.org) holds over 2400 symbols collected from major U.S. government symbol sets as well as public domain symbol sets designed by independent cartographers. To help users select symbols, an interactive map preview allows users to test a set of symbols on realistic maps as shown in *Figure 2*. Once the user has found and previewed a set of useful symbols, they can be downloaded for immediate use in a range of common formats. The selected symbols are bundled together in a.zip file which includes PNG images in three common resolutions, an Esri .style file for use in ArcGIS, and SVG vector graphics that can be used in graphic design software.

3. Refining Symbols

Our most recent additions to the Symbol Store are features that allow communities of mapmakers to iteratively refine and enhance their symbol collections. These new features build on prior work to develop an asynchronous, round-based approach for refining and formalizing domain specific symbol sets (Robinson et al. 2012). In that research, we developed and tested this standardization approach. To test this method we created a web-based platform called the e-Symbology Portal, based on a customized Drupal content management system. Subsequent to our initial testing of the e-Symbology Portal and our accompanying symbol standardization process, we saw the need for closer integration between the Symbol Store and the e-Symbology Portal in order to more efficiently and effectively support communities of mapmakers who want to share and refine their symbols.

The Symbol Store now directly engages with the e-Symbology Portal so that users engaged in our iterative process for refining map symbol standards

can select sets of symbols to review, enhance and refine their associated metadata, and identify new symbol design needs. Users can review symbols through a dedicated interface that supports metadata creation (for newly uploaded symbols) and refinement (for existing symbols in Symbol Store). *Figure 3* shows an example of how our new tools can assist users in the tasks associated with iteratively refining symbol sets through modifying their metadata.

Figure 3. The Symbol Store interface for refining symbol metadata.

4. Conclusion

Looking ahead, we envision multiple new opportunities for extensions to our work to develop new means for sharing and refining map symbols. We have recently completed an initial user study with mapmakers who engage in flood mitigation mapping to rate the overall utility and usability of the Symbol Store. Our preliminary results indicate that the tool is highly rated in terms of both key aspects of performance, but that there remain a wide range of small interface improvements and new functions that we must implement in order to maximize user satisfaction and overall utility.

As a next step, we plan to evaluate the Symbol Store and e-Symbology Portal integrated process and symbol refining tools with mapmakers. Our aim for this evaluation is to enhance existing functions designed to support symbol refinement and to identify new means to support collaboration and interoperability around symbology. We envision the use of this integrated process to be limited to a relatively small group of the overall user popula-

tion for the Symbol Store, but this group is nonetheless very important as the Symbol Store will only thrive if we are able to design good mechanisms for supporting iterative refinement of symbol sets by mapmakers.

References

ANSI (2006). ANSI INCITS-415 2006 Homeland security mapping standard - Point symbology for emergency management, American National Standard for Information Technology.

Cutter, S L (2003). GI Science, Disasters, and Emergency Management Transactions in GIS 7(4): 439-446.

Dept. of Defense, USA (2008). Common warfighting symbology: MIL-STD-2525C, United States Department of Defense.

Dymon, U J (2003). An analysis of emergency map symbology. International Journal of Emergency Management 1(3): 227-237.

Kostelnick, J C, Dobson, J E, et al. (2008). Cartographic symbols for humanitarian demining. The Cartographic Journal 45(1): 18-31.

Morrison, C and Forrest, D (1995). A study of point symbol design for computer based large scale tourist mapping. The Cartographic Journal 32(2): 126-136.

Rado, S and Dudar, I (1971). Some problems of standardization of transportation map symbols in thematic mapping. International Yearbook of Cartography 11: 160-164.

Robinson, A C, Roth, R E, et al. (2012). Developing Map Symbol Standards through an Interative Collaboration Process. Environment and Planning B: Planning and Design 39(6): 1034-1048.

Robinson, A C, Roth, R E, et al. (2011). Understanding User Needs for Map Symbol Standards in Emergency Management. Journal of Homeland Security and Emergency Management 8(1): 1-16.

Robinson, A H (1973). An international standard symbolism for thematic maps: Approaches and problems. International Yearbook of Cartography 13: 19-26.

Government Information Quarterly 30 (2013) 226–240

Contents lists available at SciVerse ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Spatiotemporal crime analysis in U.S. law enforcement agencies: Current practices and unmet needs

Robert E. Roth [a,*], Kevin S. Ross [c], Benjamin G. Finch [b], Wei Luo [b], Alan M. MacEachren [b]

[a] University of Wisconsin-Madison, Geography, 550 N. Park Street, Madison, WI 53706, United States
[b] GeoVISTA Center, Department of Geography, The Pennsylvania University
[c] Nsite, LLC

ARTICLE INFO

Available online 30 May 2013

Keywords:
Law enforcement
Public safety
Crime analysis
Crime mapping
Geographic information systems (GIS)
Cartography
Spatiotemporal analysis

ABSTRACT

This article compares the current states of science and practice regarding spatiotemporal (space + time) crime analysis within intermediate- to large-size law enforcement agencies in the Northeastern United States. The contributions of the presented research are two-fold. First, a comprehensive literature review was completed spanning the domains of Criminology/Crime Analysis and GIScience/Cartography to establish the current state of *science* on spatiotemporal crime analysis. This background review then was complemented with a set of interviews with personnel from seven intermediate- to large-size law enforcement agencies in the United States in order to establish the current state of *practice* of spatiotemporal crime analysis. The comparison of science and practice revealed a variety of insights into the current practice of spatiotemporal crime analysis as well as identified four broad, currently unmet needs: (1) improve access to externally maintained government datasets and allow for flexible and dynamic combination of these datasets; (2) place an emphasis on user interface design in order to improve the usability of crime mapping and analysis tools, (3) integrate geographic and temporal representations and analyses methods to better unlock insight into spatiotemporal criminal activity, and (4) improve support for strategic crime analysis and, ultimately, public safety policymaking and administration. The results of the interview study ultimately were used to inform the design and development of a spatiotemporal crime mapping application called *GeoVISTA CrimeViz*.

Published by Elsevier Inc.

1. Introduction: the analysis of information on criminal activity

Crime analysis describes the systematic collection, preparation, interpretation, and dissemination of information about criminal activity to support the mission of law enforcement (Boba, 2005). The goal of crime analysis is the unlocking of valuable insights from the collected crime information in order to assist law enforcement with criminal apprehension and crime prevention, to the end of improving the overall quality of life for community residents (O'Shea & Nicholls, 2003). Ideally, crime analysis draws upon both quantitative and qualitative approaches in order to understand criminal activity fully, integrating descriptive and inferential statistical analyses of crime incidents with text reports, information graphics, and prior experience to determine the appropriate response tactics, strategies, and broader policies (Gottlieb, Arenberg, & Singh, 1994; Osborne & Wernicke, 2003). Influenced by the Digital Revolution and associated Information Age, research and development within crime analysis during the past two decades has emphasized the design of computer software that supports the assembly and interpretation of digitally-native crime information (Wilson, 2007). The research reported here focuses upon a critical subset of computing technologies designed to analyze the spatial and temporal (together *spatiotemporal*) components of crime information.

The field of *Geographic Information Science* (GIScience) and its technological counterpart *Geographic Information Systems* (GIS) describe the gamut of tools and techniques available to analyze geographically-referenced information (Goodchild, 1992). GIScience subsumes a variety of topics relevant to spatiotemporal crime analysis, which include geographic information collection (geocoding, GPS technology, remote sensing, and surveying), geographic information maintenance (geographic database management and multi-resolution databases), geographic information analysis (geocomputation, geographic data modeling, spatial analysis, and spatial statistics), geographic information representation (cartography and geographic visualization) and the use of geographic information and information products (geocollaboration, geovisual analytics, public participatory GIS, and spatial decision support systems) (for a general overview of these topics, see Longley, Goodchild, Maguire, & Rhind, 2005). The term *crime mapping* is used today to describe the

* Corresponding author.
E-mail addresses: reroth@wisc.edu (R.E. Roth), kevin@kross.com (K.S. Ross), bgf111@psu.edu (B.G. Finch), wul132@psu.edu (W. Luo), maceachren@psu.edu (A.M. MacEachren).

0740-624X/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.giq.2013.02.001

application of all GIScience tools and techniques for crime analysis (Getis et al., 2000), although its original use focused on applications of Cartography only (i.e., the representation of geospatial crime information in map form).

There is a substantial volume of work within GIScience examining the treatment of spatial and temporal components of information in conjunction (e.g., Andrienko, Andrienko, & Gatalsky, 2003; Hägerstrand, 1970; Langran, 1992; Peuquet, 1994; Sinton, 1978). Despite this research, there is little implementation of temporal analytical functionality in popular GIS software. Perhaps as a direct result, the analysis of the temporal component of crime has been identified as an under-supported function of crime analysis, with Ratcliffe (2009: 12) stating in an overview of current challenges to crime analysis that "At present, the most under-researched area of spatial criminology is that of spatio-temporal crime patterns." Existing reports on crime analysis indicate that spatiotemporal analysis and visualization often is limited in practice to the generation of one-off, static maps showing crime over a small period of time, usually the past 7-to-30 days (Lodha & Verma, 1999). Thus, the possible use cases for advanced spatiotemporal crime analysis remain undetermined and therefore the positive impacts of spatiotemporal crime analysis remain unrealized.

Here, we describe research to address directly this challenge of spatiotemporal crime analysis. The aim of our research was the identification of gaps between the spatiotemporal crime analysis techniques reported in the literature and the actual use of these techniques by law enforcement to combat crime. The primary contributions of the research are two-fold. We first completed a comprehensive background review to understand the current state of *science* in spatiotemporal crime analysis, disambiguating and synthesizing relevant research from the knowledge domains of Criminology/Crime Analysis and GIScience/Cartography. We then conducted a set of interviews with experts from seven intermediate- to large-size law enforcement agencies in the United States (daytime service populations of 125,000 to many millions) in order to compare the current state of *practice* in spatiotemporal crime analysis to the previously reviewed state of science. Such a critical comparison of science and practice is relevant to detectives, officers, and decision makers working in law enforcement as well as municipal, state, and federal administrators and policymakers working broadly in public safety. The interview study also served as the needs assessment stage for the design of a spatiotemporal crime mapping application called *GeoVISTA CrimeViz* (http://www.geovista.psu.edu/CrimeViz) developed in collaboration between the Penn State GeoVISTA Center and the Harrisburg (PA, USA) Bureau of Police (for details on the application, see Roth, 2011; Roth & Ross, 2009; Roth, Ross, Finch, Luo, & MacEachren, 2010). Therefore, we were interested in identifying the key crime analysis needs of law enforcement agencies that the *GeoVISTA CrimeViz* application must support, with a particular emphasis on those needs not currently supported by readily available spatiotemporal crime analysis software.

The article proceeds in four sections. In the following section, we synthesize background material from the domains of Criminology/Crime Analysis and GIScience/Cartography to establish the current state of science on spatiotemporal crime analysis. In the third section, our interview protocol and qualitative data analysis approach is described. We present the results and discuss the key findings of the interviews in the fourth section, providing an overview of the current state of practice to contrast with the background review. The fourth section is organized according to six key crime analysis needs identified from the background review: (1) geographic information, (2) cartographic representation, (3) cartographic interaction, (4) spatial analysis, (5) temporal analysis, and (6) map and analysis use. The fifth and final section contains our concluding remarks and lists several broad spatiotemporal crime analysis needs that currently are not fully support.

2. Background review: current state of science on crime analysis

A comprehensive review of existing literature was completed prior to the interview study in order to characterize the current state of science on crime analysis. The following review is organized into three sections: (1) a summary of the origins and purpose of crime analysis from the discipline of Criminology, with an emphasis on the types of crime analysis; (2) a summary of the different kinds of geographic information that may be collected to support crime analysis and the ways to represent this information cartographically (i.e., in map form); and (3) advanced statistical and computation techniques to analyze the spatial and temporal components of these information.

2.1. Origins and purpose of crime analysis

Crime analysis has its roots in 19th century London, where the first modern police department was established (Boba, 2005). August Vollmer, Police Chief of Berkeley (CA, USA) and founding professor of the UC-Berkeley School of Criminology, often is credited with the first application of crime analysis in the United States in the early 20th century, with other important early U.S. work conducted by the 'Chicago School' of sociologists (e.g., Shaw & McKay, 1942; Sutherland, 1934). Vollmer's student, O.W. Wilson, first defined the term 'crime analysis' in his recommendation of information analysis techniques to police departments in the 1950s and 1960s (Wilson & McLaren, 1977). The crime analysis capabilities of law enforcement agencies expanded through the 1970s and 1980s (Emig, Heck, & Kravitz, 1980), due in part to federal grants provided through the National Institute of Justice, a program of the United States Department of Justice. There also was increased interest at this time in crime analysis in academia; a review of this research is provided in Harries (1999).

Crime analysis therefore is informed by the discipline of *Criminology*, or the scientific study of the causes and control of crime and delinquent behavior, with the goal of understanding criminal activity, rehabilitating convicted criminals, and improving the quality of life within a community (Sutherland, Cressey, & Luckenbill, 1992). There are two popular criminological theories that emphasize the importance of spatiotemporal pattern and process (Cahill & Mulligan, 2007). Under *routine activity theory*, an individual criminal incident requires three conditions to occur concurrently in place: (1) presence of a motivated offender, (2) presence of a suitable target, and (3) absence of a proper guardian, law enforcement or otherwise (Cohen & Felson, 1979). The spatiotemporal dynamics of these three components can be analyzed both to identify locations of elevated crime risk and to prescribe the appropriate policing tactics to attenuate this crime risk (Bruce, 2008). In contrast, *social-disorganization theory* evaluates the ability of a community, or homogenous geographic unit, to combat negative community-level changes and enforce positive ones (Shaw & McKay, 1942). By analyzing the spatial and temporal differences in demographic and environment characteristics between stable and disrupted neighborhoods, long-term policing strategies can be developed and absent public policies can be established to prevent criminal activity in blighted communities (Sampson & Groves, 1989). Together, these two theories reveal the importance of spatial and temporal context during crime analysis (Wilcox, Land, & Hunt, 2003).

Boba (2005) describes five types of crime analyses, or the general applications of criminological theory and crime analysis techniques in support of the functions of law enforcement:

(1) *Criminal investigative analysis* describes the process of collecting and analyzing information about a criminal offender. Criminal investigative analysis often involves the construction of offender profiles from known information, which then allows for the inference of offender characteristics (e.g., personality type, social habits, and work habits) based on those profiles (Jackson & Bekerian, 1997); journey-to-crime analysis, described below, is

a spatial analysis technique that can be applied to build the geographic component of an offender profile.

(2) *Intelligence analysis* expands investigation of a single individual and single crime series to a larger crime syndicate, focusing upon identification of relationships among offenders, called *link analysis*. Intelligence analysis often is applied in the context of organized crime. By establishing the offender network, law enforcement can identify and target key players in the jurisdiction and diffuse crime from the top down (Innes, Fielding, & Cope, 2005). White and Roth (2010) describe the potential of harvesting the geographic information from microblogging and social networking services to build a spatially-anchored offender network for informing and structuring intelligence analysis, although noting potential ethical concerns of harvesting volunteered geographic information.

(3) *Tactical crime analysis* is the reactive investigation of recent crime spikes within a single jurisdiction or across multiple jurisdictions (Bruce, 2008). Tactical crime analysis examines key aspects of recent criminal activity (e.g., crime type, location, time, MO, suspect description) to identify overarching patterns that may explain the recent spike. Such analysis directly informs apprehension, suppression, and target hardening blue force tactics (Bruce & Ouellette, 2008). The application of tactical crime analysis is central to the *CompStat process*, where police captains are required to present statistical analyses and cartographic representations of recent crime in their jurisdiction during regularly scheduled meetings as a way to improve leadership accountability for recent crime spikes (Walsh, 2001); CompStat has the potential for application as a strong strategic tool as well (Weisburd, Mastrofski, McNally, & Greenspan, 2002).

(4) *Strategic crime analysis* is the analysis of crime and other police-related issues to identify long-term plans for reducing crime rates and improving the quality of life for a community. Strategic crime analysis embodies the concept of *problem-oriented policing* (Goldstein, 1979), which proactively seeks to understand the underlying causes of persistent criminal activity and to develop intervention strategies to attenuate this activity. There also is an important evaluation component of strategic crime analysis that determines how well previously applied intervention strategies worked to combat crime (Boba, 2001), the results of which may inform broader public policies. Such evaluation is the final step of the strategic crime analysis model recommended by Eck and Spelman (1987) called *SARA*: *S*canning, *A*nalysis, *R*esponses, and *A*ssessment.

(5) *Administrative crime analysis* presents interesting findings of crime research and analysis to audiences within police administration, city government officials, and citizens. Administrative crime analysis directly links the detectives, officers, and decision makers responding to criminal activity and the municipal, state, and federal administrators and policymakers responsible for broader issues in public safety. Such administrative activity includes the allocation of resources within the department, such as assigning cases to detectives, and any other internal collaboration (Zhao et al., 2006). This also includes the preparation of crime reports and graphics for use in court proceedings (Harries, 1999). Finally, administrative crime analysis includes the presentation of criminal activity for public consumption, through town hall meetings or websites, to the end of promoting dialogue about public policy (Rose, 2008).

2.2. Geographic information and cartographic representation in crime analysis

As introduced above, *crime mapping* describes the analysis of the geographic component of criminal activity, both at an individual level of analysis (e.g., investigative analysis of a single crime series) and ecological level of analysis (e.g., comparative analysis across neighborhoods to identify communities with unusually high concentrations of crime) (Eck, Chainey, Cameron, Leitner, & Wilson, 2005). Law enforcement agencies are required to collect and maintain several different information sets to document criminal activity and to support crime analysis. Harries (1999) identifies three geographically-referenced information sets commonly maintained internally by municipal law enforcement agencies: (1) crime reports, (2) calls for service, and (3) vehicle recoveries. The *crime report* is the primary information set used by law enforcement agencies and includes both numerical and categorical information for indexing and searching of the record as well as a lengthy, textual narrative of the event compiled by the reporting officer. Many records management systems distinguish between *crime incidents*—which focus on attributes of the crime event such as location, time of day, and characteristics of the victim—and *arrests*—which focus on characteristics of the apprehended offender (Mamalian & La Vigne, 1999). Crime reports are organized according to *uniform crime reporting* (UCR) codes for comparison across municipalities and states. Although there is some variation in the exact coding scheme used across municipalities and states, the UCR code commonly includes a two digit *UCR primary* code indicating crime type and a two digit *UCR secondary* code indicating a discriminating condition within the primary crime type. Many municipalities also use the UCR system for indexing the *modus operandi* (MO), or method of committing the crime.

The *calls for service* information set indexes all requests for law enforcement services, typically submitted by phone, and is an order of magnitude larger than the crime incident information set, as most police dispatch does not lead directly to a reported incident or an arrest. Maps of calls for service are interpreted by crime analysts as the general 'demand' for police services within the municipality (Spelman, 1995). The *vehicle recoveries* information set, maintained primarily in larger municipalities, indexes the locations from which vehicles were reported as stolen and subsequently recovered (Chainey, Tompson, & Uhlig, 2008). A fourth information set maintained internally by some law enforcement agencies is the *field interview*, or information collected by officers from potential witnesses and offenders while on patrol (Osborne & Wernicke, 2003). Finally, Harries (1999) notes that agencies often utilize external information sources, which may include federal information like Census Bureau information, national crime information like the probation and missing persons lists as well as the sex offender registry, and volunteered information from microblogging and social networking services.

As the name implies, a principle task within crime mapping is the production of *cartographic representations* (i.e., maps) of the aforementioned kinds of geographic information collected on criminal activity. Literature on crime mapping uses alternative terminology from that common in GIScience (specifically within Cartography, e.g., Dent, 1999; Slocum et al., 2005) to describe the reference and thematic maps produced in support of crime analysis; translations between lexicons are provided below. Boba (2005) describes six types of crime maps generated to support the mission of law enforcement:

(1) *Single-symbol maps* use point symbols to represent the locations of features. In crime mapping, these are commonly referred to as *push pin maps*, drawing on the analog wall map solution used prior to the move to electronic information and GIS (for more on the etiology and evolution of push pin maps, see Wallace, 2011). In GIScience, this kind of map is referred to as a *one-to-one dot map*; when the symbology varies by color, shape, or central icon to represent a nominal difference in kind, the map sometimes is described as containing qualitative point symbols (Roth, 2010). One-to-many dot maps (i.e., *dot density maps*), where one dot represents multiple crimes, are not common in crime mapping, perhaps because of the potential misinterpretation of the meaning of a dot and the associated underestimation of total crime.

(2) *Graduated maps* are described by Boba (2005) as the use of either color or size to represent aggregated information. In GIScience, the use of a color gradient to represent aggregated information is called a *choropleth map*, while the use of size to represent aggregated information is called a *proportional symbol map* (MacEachren & DiBiase, 1991). For all choropleth maps and some proportional symbol maps, the information typically is aggregated to a set of relevant boundaries (i.e., enumeration units), such as police districts or beats. In the case of proportional symbol maps, the information also might be aggregated according to a set of point locations (e.g., apartment complexes, arenas, bars, stores) or linear features (e.g., street blocks).

(3) *Density maps* aggregate crime incidents to an arbitrary grid either directly or using a moving window smoothing function, with the frequency of each grid cell represented by color; these maps are referred to as *hot spot maps* in practice, although there is conflicting use of this term in the crime mapping literature (Chainey et al., 2008). In GIScience, this technique typically is called *isoline mapping* or *surface mapping* (Slocum, McMaster, Kessler, & Howard, 2005), although the actual isolines (i.e., lines of equal crime frequency) are rarely depicted on crime maps, with the underlying interpolation grid instead color tinted. Hot spot maps have the advantage over choropleth or proportional symbol maps in that they are not restricted by political units that have little impact on criminal activity, but they suffer more heavily from the *denominator dilemma*, as the underlying population typically is not known for arbitrary grid cells (i.e., the hot spots only may be indicating where the people are and not where criminal activity is elevated above average) (Ratcliffe, 2009).

(4) *Chart maps* show relative values within a single variable at the same time, such as the percentage of crime types by district. Examples include pie charts and stacked histograms that are placed directly on the map (Andrienko & Andrienko, 1999). The concept of a chart map can be extended to any form of *multivariate symbolization* (i.e., the representation of two or more variables in one map), rather than relative values within a single attribute only. Examples from Cartography include ray glyphs (Buja, Cook, & Swayne, 1996), star plot glyphs (Klippel, Hardisty, & Weaver, 2009), and Chernoff faces (Krygier & Wood, 2005).

(5) *Buffer maps* represent a distance zone around a feature or features of interest, such as a school or bar (e.g., Grubesic, Mack, & Murray, 2007). It is possible then to aggregate crime incidents within the buffer zone, representing the frequency using a color gradient (i.e., a buffer map/graduated map combination).

(6) *Interactive maps* leverage a digital environment to allow the map user to manipulate the mapped display according to his or her needs in real time. An interactive map is not a form of cartographic representation, as with the above map types listed by Boba (2005), but rather an additional aspect of a digital map that can be added at varying degrees to any static map (MacEachren, 1994). Thus, cartographic representation (i.e., maps) and *cartographic interaction* (i.e., user interfaces to these maps) are best considered as a fundamental duality within Cartography and GIScience, both of which requiring consideration during map design and development (Roth, 2011, 2012). MacEachren, Wachowicz, Edsall, Haug, and Masters (1999) further parse cartographic interaction into six interaction operators: (1) *focusing/filtering* (increasing or decreasing the detail of a selected subset of map objects; subsequent scholars have interpreted this operator as *filtering* or reducing the number of map objects in the display according to user imposed constraints), (2) *viewpoint manipulation* (panning, zooming, or changing the user's viewing angle of the map), (3) *brushing* (selecting a portion of the map display through direct manipulation of the map in order to perform some operation to the highlighted features), (4) *sequencing* (dividing the crime information into a set of bins according to time intervals or an attribute of the information), (5) *colormap manipulation* (adjusting the map symbolization, including the map type, color scheme, classification scheme, etc.), and (6) *assignment* (associating a variable in the information set with a component of the map display).

An additional form of cartographic representation discussed by other scholars in crime analysis is the representation of time on maps. The cartographic representation of time focuses on visual depiction of entities and patterns, geographically; it can be used to monitor changing situations and to support more complex spatiotemporal analyses of crime information (which are treated in the subsequent subsection). Spatiotemporal phenomena can be represented by either static maps, which represent temporal change using one or several graphic(s), or animated maps, which represent temporal change in the phenomenon with temporal change in the map (Monmonier, 1990). Starting with the former, there are three general approaches to the representation of multiple points or intervals of time (or any other conceptually bivariate or multivariate representation) on static maps: (1) adjacent displays, (2) separable coincident displays, and (3) integral coincident displays (MacEachren, Brewer, & Pickle, 1998). *Adjacent displays*, or *small multiples*, represent each moment in time or interval of time on a separate map, producing a series of maps with the same spatial extent (Bertin, 1967, 1983; Tufte, 1983). A set of small multiples for crime incidents would divide the information set into a series of time intervals, with each interval receiving its own map and no crime incident occurring on two maps. In contrast, coincident displays juxtapose two or more time states or intervals in a single graphic; the map is termed *separable coincident* when each time period can be individually analyzed visually (e.g., crime incidents from the past 7 days in one color and incidents from the past 8–30 days in a second color) and *integral coincident* when only the difference between time periods can be analyzed visually (e.g., using color to represent the change in crime rates by district following a newly implemented policing tactic). The second general method for representing temporal change—*cartographic animation*—describes the display of individual maps (called frames) in rapid succession (DiBiase, MacEachren, Krygier, & Reeves, 1992). While several research applications of cartographic animation to crime analysis have been reported in the literature (Brunsdon, Carcoran, & Higgs, 2007; Lodha & Verma, 1999; Wolff & Asche, 2009), Ratcliffe (2009) notes that this has translated into little practical application due to a lack of easy-to-use cartographic animation tools and training.

2.3. Spatial and temporal analysis in crime mapping

In practice, the term crime mapping applies to the complete suite of GIScience tools and techniques when used to support crime analysis, including information assembly, spatial statistics, and geocomputation in addition to the aforementioned cartographic themes of representation and interaction (Harries, 1999). Although there is no established taxonomy of spatial analysis techniques for crime analysis, several methods are discussed regularly in the literature on crime analysis and crime mapping. A primary application of spatial statistics and geocomputation to crime analysis is for identification and interpretation of spatial clusters of crime incidents. The most straightforward calculation is *spatial autocorrelation*, which measures the departure from complete spatial randomness (CSR) observed in a distribution of incidents (Griffith, 1987); positive autocorrelation suggests a distribution in which spatially near objects are likely to be similar (i.e., clustered) and negative autocorrelation suggests a distribution in which near objects are likely to be dissimilar (i.e., a checkerboard pattern). Spatial autocorrelation indices such as Geary's C, Moran's I, and Getis's G provide a single value for the entire distribution; however, these

calculations have been extended to provide *local indicators of spatial autocorrelation* (LISA) that identify the location of clusters in the distribution, rather than simply reporting that a distribution is clustered (Anselin, 1995).

A second spatial analysis technique for the identification of clusters is the *spatial scan statistic* (Conley, Gahegan, & Macgill, 2005; Kulldorff, 1997; Openshaw, Charlton, Wymer, & Craft, 1987). A spatial scan statistic is a geocomputational routine that calculates a clustering metric (called the likelihood ratio) for a large number of distinct circular or elliptical sampling windows placed over a crime incident distribution; the output of these algorithms is a small subset of the sampling windows that have a significant number of incidents contained within them as compared to the area not within the window (Chen, Roth, Naito, Lengerich, & MacEachren, 2008). Numerous scholars in criminology and crime analysis have identified the potential of scan statistics for identifying clusters of elevated criminal activity (Chainey et al., 2008; Jefferis, 1998; LeBeau, 2000; Levine, 2006; Nakaya & Yano, 2010; Zeng, Chang, & Chen, 2004). Chen (2009) provides a useful discussion of the conceptual differences between spatial autocorrelation and spatial cluster measures, such as the spatial scan statistics.

Aside from geographic clustering methods, a method of specific interest to crime analysts is *journey-to-crime analysis*, which uses the locations of related crime incidents to determine the most likely areas of offender residence and to forecast the locations of future crimes (Brantingham & Brantingham, 1981). This technique also is referred to as geographic profiling (Rossmo & Velarde, 2008), although this term is being phased out of the literature due to the implication of police surveillance. Two additional, commonly applied spatial analyses are kernel density estimation and buffering, which primarily are applied to generate density maps and buffer maps respectively (described above).

Space and time are paramount to both tactical and strategic crime analysis, as indicated by the dominant theories on criminology described above. As with spatial analyses, temporal analyses and related information graphics primarily are employed for detection of temporal clusters in criminal activity. Modifications to the scan statistic are available to identify crime incident clusters in time alone or in space and time together (Block, 1995; Levine, 2006; Zeng et al., 2004); for these modifications, the scan is completed with a moving time window, rather than or in addition to a moving spatial catchment area. An alternative technique is the *cumulative summation* (*CUSUM*) *algorithm*, which also applies a sliding temporal window to detect aberrations in event activity, such as a spike in crime that is considerably higher than past incident rates (Hutwagner, Thompson, & Seeman, 2003; Maciejewski et al., 2010).

Aside from cluster analysis, there is a small amount of research within crime analysis on the use of temporal information graphics and statistical summaries to complete visually-based *trend analysis* (Chung, Chen, Chaboya, O'Toole, & Atabakhsh, 2005; Ratcliffe, 2004; Townsley, 2008). There also is work on predictive algorithms that attempt to forecast when future crime incidents will occur (Bowers, Johnson, & Pease, 2004); such research may be considered the temporal equivalent of journey-to-crime analysis. A final potentially useful temporal analysis specific to crime information is *aoristic analysis*, a technique for estimating an exact time stamp for a crime that occurred when the victim is not present (e.g., a burglary) based on the time windows of past crimes of the same crime type (Ratcliffe & McCullagh, 1998).

There are many software applications marketed for crime analysis that provide spatiotemporal analysis; most of these applications also support basic cartographic representations and cartographic interactions. Available software packages include: ATAC (Automated Tactical Analysis of Crime; Bair, 2000), Azavea HunchLab/Crime Spike Detector (Cheetham, 2010), CrimeStat (Levine, 2006), ESRI ArcGIS (http://www.Esri.com/software/arcgis/), GeoDa (Anselin, Syabri, & Kho, 2006), MapInfo (http://www.mapinfo.com), ReCAP (Brown, 1998), SaTScan (Kulldorff, 2010), STAC (Block, 1995), and STV (Buetow et al., 2003).

3. Method: needs assessment interviews

3.1. Participants

Seven law enforcement agencies in the United States participated in an interview study designed to assess the current practices and key unmet needs of spatiotemporal crime analysis. Law enforcement agencies were purposefully sampled based on two criteria: (1) the municipal law enforcement agency (six in total) had a daytime service population of 100,000 or greater (all participating law enforcement agencies ultimately had a daytime population of 125,000 or greater) and (2) the police headquarters was within a one day drive (~250 miles) of University Park, PA (the site of the research). One federal law enforcement agency was included in the study to provide a non-municipal perspective. Recruitment was completed via email, with contact information obtained through existing GeoVISTA Center contacts in law enforcement or through agency websites. The sample therefore is representative of intermediate- to large-size law enforcement agencies in the Northeastern United States. The generalizability of results may be limited beyond this context and caution must be applied in interpretation of results due to the relatively small sample of agencies at which interviews were conducted. Each responding law enforcement agency self-identified an individual most appropriate to discuss the spatiotemporal crime analysis practices across their agency. For two of the law enforcement agencies, it was necessary to interview a pair of individuals, as their responsibilities were split according to different internal units; thus, nine interview sessions were completed in total.

A background survey was administered at the start of each interview session to establish several characteristics of the interview participants. Two participants had no post-secondary education, three participants held a Bachelors degree, two participants held a Masters degree, one participant held a PhD, and one participant held a law degree (in addition to a BS in Criminal Justice); outside of the law degree, the degrees were in either Criminal Justice (5) or Geography (2). The participant sample was composed of a near even mixture of primarily producers of spatiotemporal information and associated information products (i.e., crime analysts and crime mappers) and primarily users of this spatiotemporal information and information products (i.e., administrators, detectives, officers, and decision-makers) (Table 1). The majority (7 of 9) of participants reported producing spatiotemporal information and associated information products at least monthly, with a large minority (4 of 9) completing this activity daily. The majority (7 of 9) reported using spatiotemporal information and associated information products at least weekly. Two high ranking officers stated that while they use spatiotemporal information and information products weekly, they never produce them, while two crime analysts stated that while they produce spatiotemporal information and information products weekly, they never use them for policing or decision making purposes. Four participants were sworn officers while the other five held civilian status.

3.2. Materials and procedure

Interviews vary on the degree of structure in their questioning (Robinson, 2009). Structured interviews include a series of focused

Table 1
Interview participant regularity of producing and using spatiotemporal information and associated information products.

Regularity of activity	*Produce spatiotemporal info*	*Use spatiotemporal info.*
Daily	4	2
Weekly	1	5
Monthly	2	0
Yearly	0	0
Rarely	2	2
Total	**9**	**9**

questions that typically prompt short and equally focused responses; all participants are asked the exact same set of questions in the same order. On the other end of the continuum, unstructured interviews include a set of broad discussion topics or general themes, with no preset order; these types of questions are exploratory in nature and typically prompt longer, open-ended responses that vary greatly from person to person. Many interview protocols follow a semi-structured approach, which starts with a set of focused questions but allows the interviewer to ask follow-up or probe questions as he or she sees fit and change the order of questioning if appropriate (for an example in crime mapping, see Ratcliffe, 2000).

At the end of the interview, participants were asked if there was anything else they would like to discuss before concluding or if they had any questions about the study. All interview sessions lasted between 60 and 75 minutes and were completed at the participant's work location in a private room. For consistency, the same project member acted as the interviewer for all nine interviews. The interviews were audio recorded for subsequent qualitative data analysis, as described in the following subsection.

The interview protocol for the needs assessment proceeded in six sections; a summary of the interview questions is included in Table 2. Each interview session began with an introduction to the project and

Table 2
A summary of the interview questions.

Introduction	
Background	
1	What is your agency, department, or organization, job title, and responsibilities at this position?
2	Are you a sworn officer or a civilian?
3	Please describe your prior education and formal training?
4	Please describe any previous employment relevant to crime mapping and analysis?
5	How frequently do you produce spatiotemporal information and associated information products (maps, analyses, etc.) in your daily work?
6	How frequently do you use spatiotemporal information and associated information products (maps, analyses, etc.) in your daily work?
Information	
7	Please list the types of spatial or temporal phenomena for which your agency collects information.
8	For each collected information set, describe its: format, number of entities/records, geographic and temporal resolution, scale of analysis and mapping.
9	Does your agency use any external information sources?
10	Is the information your agency collects text/report-based or entered into a table or database?
11	Are there any information sets not collected by your agency that would be useful in crime mapping and analysis?
Mapping and Analysis	
12	Please describe the kinds of maps produced by your agency.
13	Please list the reference or basemap information your agency uses on these maps.
14	What spatial analyses or data transformations does your agency apply to the collected raw information?
15	What temporal analyses or models does your agency apply to the collected raw information?
16	Does your agency aggregate your point incident information in space or time?
17	Does your agency filter your point incident information prior to mapping?
18	Does your agency represent the temporal component of your information directly on maps?
Use	
19	How are maps and analyses used in a tactical way at your agency?
20	How are maps and analyses used in a strategic way at your agency?
21	What is the workflow from generation of maps and analyses to usage of these information products at your agency?
22	Please describe a successful use of mapping and analysis at your agency?
23	Please describe an unsuccessful use of mapping and analysis at your agency?
Do you have any last questions or comments before we conclude the session?	

an overview of the goals of the needs assessment; here, participants were informed that they did not have to respond to all questions, particularly if the question was irrelevant or sensitive. Participants then were asked two sets of brief, structured questions. Participants first responded to structured questions about their general background in law enforcement and their overall experience producing and >using spatiotemporal information and associated information products in support of crime analysis (summarized in the previous subsection). After the background questioning, participants were asked a set of structured questions about characteristics of the geographic information that their agency collects and maintains, as well as any external geographic information sources that their agency leverages.

Following the structured portion of the interview, the participants were asked two rounds of semi-structured questions. The first round of semi-structured questioning focused upon the current crime mapping practices from an information producer perspective, asking about the types of maps that are generated and the types of spatial and temporal analyses that are applied to the information. The second round of semi-structured questioning focused upon the current crime mapping practices from an information user perspective, asking about tactical and strategic uses of crime mapping, the general crime analysis workflow, and examples of successes and failures when using maps and analyses to support the mission of law enforcement.

3.3. Qualitative data analysis

Qualitative data analysis (QDA) describes the systematic interpretation of qualitative information, such as text reports, websites, photos, maps, and field observations (Dey, 1993; Miles & Huberman, 1994). A review of work using qualitative data analysis on electronic government information, government information products, and government information use is provided by Yildiz (2007), with a multitude of examples published more recently in *Government Information Quarterly* outside of the domains of law enforcement and public safety. In the most robust form of QDA, the documents in the set are decomposed to their smallest unit of analysis and a series of codes are applied to the units by several independent coders, with the coding then compared across coders to ensure reliability in interpretation of the document set.

Transcription of the audio recordings was completed using Transana, with the transcripts then unitized at the statement level in Microsoft Excel for margin coding (Bertrand, Brown, & Ward, 1992). The above background review on the current state of science in spatiotemporal crime analysis was used to identify six key themes: geographic information (G), cartographic representation (R), cartographic interaction (I), spatial analysis (S), temporal analysis (T), and map and analysis use (U). These key themes are areas in which law enforcement agencies may have an unmet spatiotemporal analysis *need*, defined as a resource or feature required by the targeted end user to complete their work and thus represents a disconnect between the current states of science and practice in spatiotemporal crime analysis. Thirty-one individual codes then were identified from the above background review within these six needs; each code was marked during margin coding to distinguish needs that were met by existing software (+) from those that were not met (−) at the time of the interview. Table 3 lists the six higher level categories, each of the 31 codes across these categories, and the source of the individual code from the above background review; Table 4 lists the frequency of each code across the nine transcripts. A total of 515 codes identifying user needs were applied to the nine transcripts, an average of 57.2 codes per transcript.

Two coders with expertise in GIScience and training in crime analysis were hired to apply independently the same 31-part coding scheme used in the initial coding, with code reliability assessed using the inter-rater reliability score described by Robinson (2008). The two coders achieved inter-coder reliability scores of 93.2% and 87.6% against the initial margin coding, indicating a high degree of reliability in the interpretation and application of the coding scheme, particularly considering the large number of codes in the coding scheme. Differences in coding were reconciled for reporting through discussion among the coders and a third project member. Statements were sorted according to the assigned code and summarized using the synoptic style of reporting described by Monmonier and Gluck (1994) and Roth (2009). Crime analysis needs within the six higher level categories are summarized in the following section.

4. Results and discussion: current state of practice

4.1. Geographic information

Codes included in the geographic information (G) category indicate statements about the geographic information sets leveraged to support crime analysis. Five codes were included under the geographic information (G) category based upon the above background review: (G1) crime reports (incidents plus arrests), (G2) calls for service, (G3) vehicle recoveries, (G4) field interviews, and (G5) any external information sources not collected or maintained by the law enforcement agency itself. The most frequently discussed geographic information sets include crime reports (average = 6.6) and external information sources (average = 6.3), with participants identifying external information sources as an unmet need (average = 1.9) slightly more frequently than crime reports (average = 1.8). Participants rarely discussed calls for service (average = 1.6), vehicle recoveries (average = 0.8), and field interviews (average = 0.2).

Overall, participants indicated that crime reports are the primary geographically-referenced information collected and used at their law enforcement agencies. Discussion centered almost exclusively on crime reports describing incidents, rather than arrests. The number of crime incident records collected per year by the interviewed agencies ranges from approximately 7,000 to 2.5 million, indicating a need for user interfaces to scale to increasingly large and complex information sets. All participants described a similar set of core attributes captured in their crime incident reports: crime type (by UCR code), address, date and time (often with precision to the minute, except in the cases of burglary when a time range is given), MO, suspect and victim description, and a text narrative. Surprisingly, one participant noted that his/her agency did not regularly geocode (i.e., convert the listed address to spatial coordinates) their crime incident reports for mapping and analysis, instead geocoding only a small grouping of crime incident reports if an association is suspected. One participant also noted that his/her agency also captures information on *location type*, such as "parking lot, convenience store, restaurant, street, sidewalk"; while this information is not geographic in the sense of absolute coordinates, it is highly relevant to spatiotemporal crime analysis as it provides important geographic context for understanding the crime setting.

Most participants indicated that their agency leverages externally maintained geographic information sources. One participant stated that "we have gone out and tried to collect as many datasets as we can find that may or may not be useful to us, just so we know where they are at and what we have access to." Two important geographic information sets mentioned repeatedly were parole/probation records and registered sex offender records maintained at the state level, both of which include the home address of the offenders. Departments that have access to this information emphasized its utility and those that do not have access acknowledged their desire to acquire it. Other information sets include DMV (Department of Motor Vehicles) records and infrastructure information from the City's GIS department. One external geographic information set that is not used regularly is the federal census, with one participant stating that "I had to jump through hoops just to break it down by district and section within the police department" and a second stating that "the census data is about nine years old now [and] just isn't accurate…the census doesn't really mean much of anything to us." This contradicts descriptions of crime analysis in the

Table 3
The coding scheme applied for QDA of the needs assessment study. The categories of needs and individual codes were derived from the background review.

ID	Name	Source
Geographic information: Statements about the information sets used and their characteristics		
G1	Crime reports (incidents and arrests)	Harries (1999)
G2	Calls for service	Harries (1999)
G3	Vehicle recoveries	Harries (1999)
G4	Field interviews	Osborne & Wernicki (2003)
G5	External information sources	Harries (1999)
Cartographic representation: Statements about the way information sets are mapped		
R1	Push pin maps (i.e., one-to-one dot or single-symbol maps)	Boba (2005)
R2	Choropleth maps (i.e., graduated maps using color)	Boba (2005)
R3	Proportional symbol maps (i.e., graduated maps using size)	Boba (2005)
R4	Hot spot maps (i.e., density maps)	Boba (2005)
R5	Multivariate symbolization (i.e., chart maps)	Boba (2005)
R6	Buffer maps	Boba (2005)
R7	Maps representing time	Monmonier (1990)
R8	Reference or basemap symbolization	Dent (1999)
Cartographic interaction: Statements about the way in which maps are manipulated		
I1	Focusing/filtering	MacEachren et al. (1999)
I2	Viewpoint manipulation	MacEachren et al. (1999)
I3	Brushing	MacEachren et al. (1999)
I4	Sequencing	MacEachren et al. (1999)
I5	Colormap manipulation	MacEachren et al. (1999)
I6	Assignment	MacEachren et al. (1999)
Spatial analysis: Statements about applied spatial statistics and geocomputation		
S1	Spatial autocorrelation measures	Griffith (1987); Anselin (1995)
S2	Spatial scan statistics	Openshaw et al. (1987)
S3	Journey–to–crime analysis (i.e., geographic profiling)	Brantingham & Brantingham (1981)
Temporal Analysis: Statements about applied temporal transformations and models		
T1	Temporal and spatiotemporal cluster analysis	Zeng et al. (2004)
T2	Trend analysis	Ratcliffe (2004)
T3	Predictive analysis	Bowers (2004)
T4	Aoristic analysis	Ratcliffe & McCullagh (1998)
Map & analysis use: Statements about the use of maps and analysis to support law enforcement		
U1	Criminal investigative analysis	Boba (2005)
U2	Intelligence analysis	Boba (2005)
U3	Tactical crime analysis	Boba (2005)
U4	Strategic crime analysis	Boba (2005)
U5	Administrative analysis	Boba (2005)

Table 4
Frequency of codes applied for QDA of the needs assessment study. *Total* describes the total number of statements given the code, while *Avg* divides this total by the sample size (n = 9).

ID	Name	Have		Need		All	
		Total	Avg	Total	Avg	Total	Avg
G1	Crime reports	43	4.8	16	1.8	59	6.6
G2	Calls for service	9	1.0	5	0.6	14	1.6
G3	Vehicle recoveries	5	0.6	2	0.2	7	0.8
G4	Field interviews	2	0.2	0	0.0	2	0.2
G5	External information sources	40	4.4	17	1.9	57	6.3
Total geographic information (G)		**99**	**11.0**	**40**	**4.4**	**139**	**15.4**
R1	Push pin maps	26	2.9	6	0.7	32	3.6
R2	Choropleth maps	13	1.4	3	0.3	16	1.8
R3	Proportional symbol maps	7	0.8	2	0.2	9	1.0
R4	Hot spot maps	15	1.7	2	0.2	17	1.9
R5	Multivariate symbolization	1	0.1	1	0.1	2	0.2
R6	Buffer maps	11	1.2	0	0.0	11	1.2
R7	Maps representing time	33	3.7	9	1.0	42	4.7
R8	Reference or basemap symbolization	29	3.2	8	0.9	37	4.1
Total cartographic representation (R)		**135**	**15.0**	**31**	**3.4**	**166**	**18.4**
I1	Focusing/filtering	18	2.0	10	1.1	28	3.1
I2	Viewpoint manipulation	5	0.6	1	0.1	6	0.7
I3	Brushing	11	1.2	1	0.1	12	1.3
I4	Sequencing	3	0.3	2	0.2	5	0.6
I5	Colormap manipulation	1	0.1	0	0.0	1	0.1
I6	Assignment	3	0.3	0	0.0	3	0.3
Total cartographic interaction (I)		**41**	**4.6**	**14**	**1.6**	**55**	**6.1**
S1	Spatial autocorrelation measures	0	0.0	2	0.2	2	0.2
S2	Spatial scan statistics	4	0.4	1	0.1	5	0.6
S3	Journey-to-crime analysis	7	0.8	2	0.2	9	1.0
Total spatial analysis (S)		**11**	**1.2**	**5**	**0.6**	**16**	**1.8**
T1	Temporal & spatiotemporal cluster analysis	3	0.3	0	0.0	3	0.3
T2	Trend analysis	23	2.6	5	0.6	28	3.1
T3	Predictive analysis	4	0.4	0	0.0	4	0.4
T4	Aoristic analysis	3	0.3	0	0.0	3	0.3
Total temporal analyses (T)		**33**	**3.7**	**5**	**0.6**	**38**	**4.2**
U1	Criminal investigative analysis	3	0.3	2	0.2	5	0.6
U2	Intelligence analysis	8	0.9	3	0.3	11	1.2
U3	Tactical crime analysis	34	3.8	3	0.3	37	4.1
U4	Strategic crime analysis	19	2.1	7	0.8	26	2.9
U5	Administrative analysis	20	2.2	2	0.2	22	2.4
Total map & analysis uses (U)		**84**	**9.3**	**17**	**1.9**	**101**	**11.2**
Total		**403**	**44.8**	**112**	**12.4**	**515**	**57.2**

literature, where the integration of census information for strategic crime analysis is highly recommended (e.g., Cahill & Mulligan, 2007); the recent release of the 2010 census may alleviate the latter participant's concern, at least for the small window of time that the census is current enough for the purpose of crime analysis. Many of the participants stated that their agencies use volunteered geographic information collected from social networking websites such as Facebook and MySpace, primarily for link analysis; none of the participants recalled using volunteered geographic information posted to microblogging websites like Twitter, but stated that they would like to do so if simple methods were available. While most law enforcement agencies appear willing and able to synthesize a large amount of external information sources, it is important to note that participants knew little about if or how their internally maintained information are shared with other agencies within their municipality or other law enforcement agencies in neighboring cities.

4.2. Cartographic representation

Codes included in the cartographic representation (R) category indicate statements about the way in which the collected geographic information are represented in map form to support crime analysis. Eight codes are included under the cartographic representation (R) category based upon the above background review: (R1) push pin maps (i.e., one-to-one dot or single-symbol maps), (R2) choropleth maps (i.e., graduated maps by color), (R3) proportional symbol maps (i.e., graduated maps by size), (R4) hot spot maps (i.e., density or surface maps), (R5) multivariate maps, (R6) buffer maps, (R7) maps representing time, and (R8) aspects of the underlying basemap. On average, cartographic representation was the most mentioned of the six key themes, indicating that the design of the map remains as important—or more so, as suggested by the code frequency—as more technically complex spatial and temporal analyses. The most frequently discussed cartographic representation forms include maps representing time (average = 4.7), basemaps (average = 4.1), and push pin maps (average = 3.6). Less discussion was elicited concerning hot spot maps (average = 1.9) and choropleth maps (average = 1.8). Participants infrequently identified buffer maps (average = 1.2), proportional symbol maps (average = 1.0), and multivariate maps (average = 0.2) as key needs, either met or unmet.

Of the set of crime map types identified by Boba (2005), push pin maps were by far the most commonly identified as a core need by participants. Participants indicated that the primary explanation for the frequent employment of push pin maps was their simplicity, but their simplicity in interpretation by map users rather than their simplicity in creation by mapmakers. One participant stated that "for the most part, simpler is better for the [officers] on the street" and second stated that "a lot of times we want to do more fancy and sophisticated analytical maps, but [high ranking officials] want to see pin maps, so of course we have to do pin maps." Such statements suggest that a large number of the information users at law enforcement agencies currently are incapable or unwilling to utilize more complex cartographic representations in support of criminal investigation and resource allocation. Interestingly, at least one agency still commonly adds push pins to printed wall maps manually for serial tracking and collaborative decision-making.

Despite its relatively small amount of overall discussion within the cartographic representation (R) category, participants identified the hot spot map (i.e., density or surface map) as the preferred cartographic representation technique for large volume crimes where aggregation is necessary. Several participants noted that the generation of hot spot maps is growing in popularity, with one participant stating that "the latest and greatest thing that people like to see is a hot spot map" and a second stating that "it was something that my analysts saw at a conference, [so] we started making hot spot maps." Several of the participants responsible for producing hot spot maps indicated that they primarily use the *kernel density estimation* (KDE) function in the 3D Analyst extension to Esri's ArcGIS, which uses a moving window (i.e., a kernel) or multiple pixels to generate a crime estimation for the central pixel. However, several participants were cautious about inappropriately using hot spot maps, fearing that officers and detectives "don't understand them." One agency specifically avoided the use of KDE because they considered it misleading, as their map users did not realize that the shading is a smoothed result of a search window and not an aggregate of crime incidents within the specific pixel. Interestingly, several participants identified hexagons, rather than squares, as the preferred tessellation (i.e., cell shape), as the representation leads to naturally shaped hot spots that are easier to interpret and use for allocating patrol.

Participants generally considered choropleth maps as inferior to hot spot maps for crime analysis, as choropleth maps aggregate crime information to political or jurisdictional units that have little impact on patterns of criminal activity. Only one law enforcement agency regularly generated choropleths instead of hot spot maps for tactical crime analysis. However, choropleth maps are generated regularly for strategic and administrative crime analysis; one participant stated that "with more strategic or long-term maps, then we will do a choropleth map" and a second stated that choropleth mapping is "more of the administrative work" that he/she does. Other maps that were created on occasion by a subset of departments include graduated symbol maps (graduated by point locations and, at one department, by line segment), flow maps (primarily to connect the location of vehicle thefts versus recoveries, but also to connect crime incidents in a serial; never scaled to show the volume of flow), and buffer maps. Most mapmaking is completed using Esri's ArcMap, although one agency exclusively uses the Microsoft MapPoint software to produce push pin maps.

Surprisingly, cartographic representation of time was the most discussed of the themes included in the cartographic representation (R) category; maps representing time also were the most frequent cartographic representation form listed as a need that currently is unmet. Coloring the pins on a push pin map according to the date of the incident—a separable coincident technique—was identified by participants as the primary method for representing time on maps. One participant stated that "I will have 28 days in color, the previous 28 days in grey, and the current 7 days will be in purple", a second participant stated that he/she will use "a color ramp to show thirty incidents over a two month period…the initial incident may be a white dot and it progresses to red over time", and a third participant stated that he/she would produce "a map of the last 30 days, [with] halos around the different periods of times." Participants identified the last seven days, the last month, and the last two months as common time periods used for temporal colored push pin maps. Several departments also apply color to represent cyclical temporal patterns, such as different days of the week (e.g., 'Monday', 'Tuesday', 'Wednesday') or shifts ('8 am–4 pm', '4 pm–12 am', '6 pm–2 am', '12 am–8 am'). Several participants also remembered isolated times that they had created animations when specifically prompted in the interview, with frames typically forwarded manually using Microsoft Powerpoint; one participant did report using Adobe Flash once to generate an animation. While these participants stated that the animations were extremely well received, they also stated this was not an approach they typically completed due to the perceived time-consuming nature of constructing the animation, making animation a key unmet cartographic representation need.

Participants agreed that a street network with labels is the primary basemap or reference information included on their crime maps. Other infrastructure information like building footprints and parcels may be included for large scale maps. Participants noted that points of interest (e.g., schools, parks, police stations, bars, restaurants, bus stops) may be included, but typically only upon special request or for the generation of buffer maps. Most agencies have access to

aerial imagery, but generally only include it upon special request (particularly for court maps).

4.3. Cartographic interaction

Codes included in the cartographic interaction (I) category indicate statements about the way in which the generated cartographic representations are manipulated through user interfaces. Six codes are included under the cartographic interaction (I) category based upon the above background review: (I1) focusing/filtering, (I2) viewpoint manipulation, (I3) brushing, (I5) colormap manipulation, and (I6) assignment. Focusing/filtering was identified as the biggest need (average = 3.1) among the MacEachren et al. (1999) cartographic interaction operators, with brushing (average = 1.3) also garnering some discussion. Viewpoint manipulation (average = 0.7), sequencing (average = 0.6), assignment (average = 0.3), and colormap manipulation (average = 0.1) were discussed infrequently.

Participants stated that most of the exploratory crime analysis leveraging cartographic interaction is completed with desktop GIS software designed for other purposes. While some law enforcement agencies have generated customized interface widgets providing some cartographic interaction in real-time, this still is limited to a subset of interaction operators and a subset of agencies. Across the nine participants, cartographic interaction was performed only by the crime analysts responsible for producing crime maps and analyses. Participants from agencies that hold CompStat meetings noted that high-ranking officers may include interactive maps in their presentations, manipulating the cartographic representation in real-time; however, in all reported examples "analysts are in the back driving that." Thus, a transparently usable interface, providing a subset of core interaction operators through an intuitive interface design, may fill a key unmet need for the consumers of crime maps, such as administrators, detectives, supervisors, and other decision makers.

Participants identified focusing/filtering as the most needed cartographic interaction operator, with one participant acknowledging that crime analysts "filter continuously, every time they make a map they filter." Most participants have to perform focusing/filtering queries using a series of *nested dialog windows*, which hide interface features in a set of windows that must be activated in sequence by users, limiting the *usability* (i.e., ease-of-use) of the application and making exploratory crime analysis difficult. One recommended solution is the inclusion of *persistent dialog windows* housing focusing/filtering controls that remain visible until minimized by the user; this was deemed particularly appropriate for common filtering attributes such as UCR and MO. Participants also identified brushing as a commonly employed cartographic interaction operator, noting that brushing typically is provided on digital push pin maps in order to retrieve additional information about the selected crime incident. Only one participant described an application currently in use at his/her agency that uses brushing for linked highlighting across multiple information graphics (a desktop mashup between ArcGIS and the ATAC system).

Participants indicated that the other cartographic interaction operators are employed infrequently. Participants from agencies that hold CompStat meetings noted that the sequence operator sometimes is applied during the meeting, as they have to bin their crime incident information across multiple attributes for each weekly or bimonthly meeting. However, the generated maps almost never are animated across the temporal bins generated by the sequence operator. Participants stated that viewpoint manipulation often is available only in a discrete fashion (i.e., no continuous panning across the map extent or zooming across scales), as analysts have the extent of each district preset in ArcMap and toggle between individual districts and the full extent. Participants indicated that assignment and colormap manipulation are applied rarely.

Interestingly, participant discussion on cartographic interaction revealed a split on the potential utility of web maps and web mapping services, such as Google Maps. One participant was excited about the potential of such services, stating that "a lot of folks are looking at the Google Maps…I think Google Maps has been really useful in getting law-enforcement to use these types of things because prior to Google Maps, agencies didn't even know these things were accessible" while a second was concerned not about the interactivity of these services, but about the underlying information quality, stating that "if you go to Google Maps, or something like that, you don't know how old those maps are or what actually changed, so some of that information when you physically get out there could be bad information." Interestingly, one participant stated that he/she had experimented with using Google Earth because of the potential for sharing interactive maps via .kml files; this participant stopped doing this, however, because the intended users did not have Google Earth installed on their work machine (and did not have security permission to do so) and therefore could not access the maps.

4.4. Spatial analysis

Codes included in the spatial analysis (S) category indicate statements about the spatial statistics and geocomputational routines that are applied in support of crime analysis. Three broad codes were included under the spatial analysis (S) category based upon the above background review: (S1) measures of spatial autocorrelation, (S2) spatial scan statistics, and (S3) journey-to-crime analysis; other spatial analyses were considered in the original coding scheme, but were dropped because they were not identified during the interviews. Spatial analysis was by far the least identified need during the transcript analysis, with an overall average of only 4.2 statements per transcript; journey-to-crime analysis was identified most frequently as a spatial analysis need (average = 1.0), followed by spatial scan statistics (average = 0.6) and spatial autocorrelation measures (average = 0.2).

The overall low amount of discussion on spatial analysis needs revealed a large and unexpected disconnect between practice and science, where the application of spatial transformations and spatial models is frequently reported and highly recommended. One possible explanation for this disconnect that came up in several of the interviews was that there is a lack of relevant expertise within the municipal law enforcement agencies in terms of both understanding how to apply the spatial analysis techniques and how to interpret their results. There were only two total references to spatial analysis by sworn officers, with one participant stating that he/she "leave[s] that up to the analyst to do because they are a lot more familiar with those type of things that I am." There also was a general notion communicated by many of the participants that crime analyst units are undermanned and even misused, forcing the crime analysts to respond to specific, often basic requests rather than providing them with the autonomy to complete more advanced spatial analyses. One participant noted his/her agency "is very short on manpower… so it becomes very difficult [to complete such analyses]." This participant went on to add that "I don't want to say it's a waste of time, but they just don't have the time to focus on this."

The two most commonly applied spatial analyses—kernel density estimation and buffering—are completed to generate an output map. Interestingly, participants from two different agencies commonly apply these two spatial analyses upon request during their CompStat meetings, illustrating the potential of providing cartographic interfaces to computational processes in support of exploration and reasoning in real-time. Several participants stated that they occasionally conduct journey-to-crime analyses using the geographic mean calculation in the CrimeStat application (Levine, 2006) or the animal movement extension in ArcGIS. None of the participants calculated spatial autocorrelation statistics, with most participants seeing limited value in metrics that do not provide local indicators of crime clusters. One participant had applied the spatial scan statistic routines provided in SaTScan

(Kulldorff, 2010) and the LISA statistics provided in GeoDa (Anselin et al., 2006) several times, but noted that such application "is unusual… [crime analysts] don't use GeoDa, they don't even use SaTScan." While other participants hinted at the need to automate the identification of crime incident clusters, no other participants were aware of spatial scan statistics or other methods of spatial cluster analysis, stating that cluster identification is completed visually at their agencies.

4.5. Temporal analysis

Codes included in the temporal analysis (T) category indicate statements about the temporal transformations and models that are applied in support of crime analysis. The temporal analysis (T) category includes four codes drawn from the above review: (T1) temporal and spatiotemporal cluster analysis, (T2) trend analysis, (T3) predictive analysis, and (T4) aoristic analysis. Trend analysis using information graphics was identified by participants as the largest temporal analysis need (average = 3.1). Predictive analysis (average = 0.4), temporal and spatiotemporal cluster analysis (average = 0.3), and aoristic analysis (average = 3.1) were discussed infrequently.

There was extreme variation in the temporal analyses applied to the crime information across the participating law enforcement agencies. Most of the participants indicated that their agencies apply very little temporal analyses in support of crime analysis. These participants primarily rely upon trend analysis that does not have a linked cartographic component to it. Participants stated that the temporal information graphics used for trend analysis typically are generated as one-offs in Microsoft Excel or Crystal Reports and are restricted to incidents in the past one or two months, indicating an emphasis on tactical over strategic crime analysis (see the following subsection). Interactive integration of these temporal graphics with map views was identified as a key unmet need. Participants noted that there is little systematic interpretation of the time-series graphics, with one participant saying "there is quite a bit of qualitative feel to it." Most participants stated that their agencies generate temporal composite graphics that show cyclical crime patterns, such as peaks by day of the week or time of the day; at several agencies, however, these graphics only are created when a crime analyst notices a potential cyclical pattern when reading the individual crime reports, rather than generating the composite graphics regularly to identify patterns without sifting through the full incident narrative. For these agencies, the application of temporal or spatiotemporal clustering analysis, temporal prediction algorithms, and aoristic analysis was minimal.

In contrast, participants from two of the law enforcement agencies indicated that they regularly apply sophisticated temporal analyses, offering initial insight into potential use case scenarios of spatiotemporal crime analysis. One agency takes full advantage of the ATAC software (Bair, 2000). ATAC automates the generation of temporal information graphics for trend analysis and provides a suite of statistics to help with the interpretation of these graphics. ATAC also has an aoristic analysis feature that provides a time split for incidents logged with a time interval. Further, ATAC includes a temporal prediction feature, although the participant using ATAC stated "I don't understand the [technique] very well" and that "this technique didn't seem to work very well for me." Finally, as mentioned above, the ATAC software is fully linked with ArcMap, allowing for real-time, interactive exploration of criminal activity in both space and time.

A second agency employs the HunchlLab/Crime Spike Detector software developed by Azavea (Cheetham, 2010), referred to internally as SpikeStat. The purpose of SpikeStat is to automate the identification of spatiotemporal hot spots of criminal activity. The participant reported that the software uses a spatiotemporal scan statistic to identify crimes spikes in both space and time. Interestingly, the implemented spatiotemporal scan statistic uses different window sizes based upon the type of crime under investigation because, as the participant noted, "you want to have a smaller search radius for thefts and a bigger one for homicide." The locations of these spikes then are compared against the set of police jurisdictions in order to send an alert to the appropriate commanding officer. The crime analyst using SpikeStat was pleased with the results, stating that it works "like an early warning system."

4.6. Map and analysis use

Codes included in the map and analysis use (U) category indicate statements about the way in which spatiotemporal mapping and analysis techniques are used in support of crime analysis. The map and analysis use (U) category includes a code for each of Boba's (2005) five types of crime analyses reviewed above: (U1) criminal investigative analysis, (U2) intelligence analysis, (U3) tactical crime analysis (U4) strategic crime analysis, and (U5) administrative analysis. Support for tactical crime analysis was identified by participants as the overall largest need (average = 4.1), but strategic crime analysis was identified as the largest unmet need (overall average = 2.9; unmet average = 0.8). Participants also identified support for administrative analysis as an important need (average = 2.4); support for intelligence analysis (average = 1.2) and criminal investigate analysis (average = 0.6) were less frequently discussed.

An interesting characteristic of the current practice of crime analysis was revealed when comparing discussion on tactical versus strategic crime analysis. Most of the participating law enforcement agencies primarily conduct tactical crime analysis, applying spatiotemporal analyses and producing output crime maps in order to react to the most recent crime spikes. One participant stated that "tactically is probably how we most often use our maps" and went on to say "a lot of what we do is more just support of day-to-day functions of the police department, so you don't see [strategic crime analysis] a lot…we are more tactical." A second participant stated that "currently the way that we are utilizing [crime analysis] is to attack specific problems" and went on to say "by the time we start gathering the clustering, we have attacked the clustering, and once that clustering is eradicated, we then move on; we do not have the time or luxury to see if the clustering started nine months ago." Participants indicated that the purpose of the tactical analysis is to adjust blue force patrolling in response to recent criminal activity; one participant stated that the "most common use [of tactical crime analysis] is for patrol deployment" while a second stated "we do tactical analysis for patrol." However, at least one participant lamented this focus of crime analysis, stating "that is why we play catch-up most of the time, because it is all reactionary."

While all but one participant could think of at least a single example of strategic mapping completed at their agency, only three of the interviewed agencies considered themselves positioned to complete strategic crime analysis on a regular basis. One key barrier to strategic analysis identified by the participants is that many agencies are undermanned, a similar barrier preventing more sophisticated spatial analyses. With regard to conducting strategic crime analysis, one participant stated that "allocating resources is a big deal" while a second participant stated "if they are overloaded with requests they won't have time to do [strategic analysis]", and a third stating that his/her agency conducts strategic analysis "whenever they get grant money or extra money." Thus, it appears as though law enforcement agencies currently are in need of a cartographic interface that supports rapid and straightforward strategic analyses in addition to tactical analyses.

A second, and likely related, barrier to strategic analysis identified by several of the participants is accountability, as enacting intervention programs based on strategic analysis requires a long-term commitment from decision makers. As one participant noted "with strategic projects, it is a lot easier to ignore them than actually go out and initiate projects with them because it takes a lot of time and effort to do a [strategic] project … if we are not being held accountable, it is a lot easier for some people not to do it." The best examples of cartographic representations, and associated user interfaces,

that support strategic crime analysis were provided by participants whose agencies hold regular CompStat meetings; four of the seven participating agencies hold CompStat meetings, with three of these four agencies conducting regular or semi-regular strategic crime analysis. These participants noted that the primary purpose of the CompStat process is to increase accountability, which has a direct tactical goal of reacting to recent crime spikes, but also should support the long-term strategic goal of improving the quality of life of a community.

Beyond Boba's (2005) tactical and strategic forms of crime analysis, participants also were able to provide numerous examples of administrative analysis. Almost all participants indicated that their agencies generate maps for court proceedings, with several agencies balancing their budgets by charging for the preparation of these maps. Many participants also described map-based reports or simple online mapping websites that are maintained by their agency for the public consumption of spatiotemporal crime information. Participants provided few examples of criminal investigative analysis and intelligence analysis during the interviews. Examples of criminal investigative analysis primarily referenced applications of the journey-to-crime analysis described above. Examples of intelligence analysis primarily referenced the extraction of geographic information from social networking applications to identify the location and relationships of a gang or other organized crime syndicate.

5. Conclusion and outlook: unmet needs for spatiotemporal crime analysis

This article provides a snapshot and comparison of spatiotemporal crime analysis science and practice, with the aim of revealing major disconnects and currently unmet needs. This research contributes to our knowledge of spatiotemporal crime analysis in two ways. We first completed a comprehensive background review across the domains of Criminology/Crime Analysis and GIScience/Cartography in order to characterize the current science of spatiotemporal crime analysis. We then conducted a set of interviews with seven law enforcement agencies in order to compare our background review to the current practice of spatiotemporal crime analysis; again, the insights elicited from the interviews are specific to intermediate- to large-size law enforcement agencies located in the Northeastern United States. The comparison between science and practice was completed across six themes relevant to spatiotemporal crime analysis: (1) geographic information, (2) cartographic representation, (3) cartographic interaction, (4) spatial analysis, (5) temporal analysis, and (6) map and analysis use. Importantly, the comparison between the background review and interview responses revealed several broad, unmet needs for spatiotemporal crime analysis in United States law enforcement agencies, each of which span across several or all of these six themes:

(1) *Expand and combine geographic information sources*: All of the participating law enforcement agencies indicated the need to acquire geographic information from additional sources. Participants noted two internal or government information sources, which are compiled in a consistent and top-down manner: parole/probation records and registered sex offender records. However, many intriguing comments were offered from participants with regards to external information sources. Law enforcement personnel need applications that allow for fast and flexible combination of internal and external information sources, an approach described in information science as a *mashup* or, regarding online applications, *Web 2.0* technologies (O'Reilly, 2007; Roth et al., 2008). They also require cartographic representations and cartographic interactions that scale to the growing size of these information sets, particularly volunteered geographic information sources such as Facebook and Twitter. Further, participants indicated that they are not fully aware how their internally maintained information sets are used by other agencies at the municipal, state, or federal level. Greater coordination across agencies involved in law enforcement and public safety would act to refine the database schema to better support diverse information uses, promote transparency and collaboration across agencies, and remove overlap in collection and maintenance efforts.

(2) *Improve the usability of crime mapping and analysis tools*: Acquisition of additional information sources means little if this information cannot be made usable through mapping and analysis techniques that are both easy to perform and comprehend. The above background review yielded a large number of techniques regarding crime mapping and analysis that support the mission of law enforcement. However, based on our empirical results, only a portion of this spatiotemporal crime analysis toolkit regularly is put to use in intermediate- to large-size law enforcement agencies; cartographic representation is limited primarily to pushpin and hotspot maps, cartographic interaction is limited primarily to the focusing/filtering and brushing, spatial analysis is surprisingly limited altogether, and temporal analysis exhibits a large amount of variation across agencies. Rather than continuing the pursuit of novel spatiotemporal crime mapping and analysis tools and techniques to add to those reviewed above, researchers perhaps instead should be investigating how to make existing tools and techniques *transparently usable* (i.e., immediately can be used by law enforcement with little training) (Robinson, Roth, & MacEachren, 2011). The topic of usability is one that has received minimal attention within crime analysis and spatial criminology, but one that is of fundamental importance considering that most-to-all law enforcement personnel are not formally trained in spatiotemporal crime analysis and have little experience interpreting complex cartographic representations and spatiotemporal analytical results. By placing an emphasis on the design of the user interfaces to the mapping and analysis techniques—rather than the techniques themselves, and perhaps even limiting their sophistication depending on the use case scenario—a greater number of law enforcement personnel can integrate spatiotemporal analysis into their workflows. Pervasive use of highly usable interfaces to simplified spatiotemporal mapping and analysis techniques also may promote buy-in within the agency to allow dedicated crime analysts to spend their time performing more sophisticated mapping and analysis techniques.

(3) *Integrate geographic and temporal representations and analyses*: Criminal activity has prominent spatial and temporal components that must be treated in concert during the analysis of crime information in order to glean the maximum amount of insight in to the identified pattern. Feedback elicited from the interview study supports Ratcliffe's (2009: 12) assessment that "At present, the most under-researched area of spatial criminology is that of spatio-temporal crime patterns." Across the seven participating law enforcement agencies, there were numerous positive examples of crime analysis treating the geographic (at least regarding mapping; less so for spatial analysis) or the temporal component individually. However, the representation of time on maps was identified as the primary unmet need regarding cartographic representation, with participants indicating the need for cartographic animation in particular. Further, only a single law enforcement agency described a use case scenario that considered both space and time together (the ATAC-ArcGIS desktop mashup). Research on representation, interaction, and analysis techniques that are explicitly spatiotemporal appears to be the most fruitful avenue for crime analysis moving forward, again with a mind towards development of transparently usable interfaces to these spatiotemporal techniques.

(4) Improve support for strategic crime analysis: An emphasis on tactical crime analysis of recent criminal activity, while

understandable from a practical perspective, privileges the victim of the crime, as the goal is to ameliorate the damages incurred quickly and ensure that justice is served. However, strategic crime analysis across longer time periods is needed to better understand the offenders participating in the criminal activity, the second condition under routine activity theory required for the occurrence of a crime (the law enforcement guardian being the third). All participating law enforcement agencies emphasized that it only is through such long-term, strategic spatiotemporal analysis of criminal activity that institutionalized criminal activity may be mitigated and blighted communities may be revitalized. Such an emphasis ultimately requires and reinforces better public safety policymaking and administration as well. Yet, participants noted that resources, tools, and training for strategic spatiotemporal analysis are lacking. In a period during which resources towards law enforcement and public safety are in decline, it is increasingly important to provide law enforcement agencies with crime mapping and analysis tools that are affordable, intuitive, and useful so that these agencies can improve the efficiency of their spatiotemporal crime analysis work and therefore dedicate additional time towards strategic crime analysis.

As stated in the Introduction, the interview study acted as the needs assessment stage for design and development of a spatiotemporal crime mapping application called *GeoVISTA CrimeViz* (http://www.geovista.psu.edu/CrimeViz), a project completed in collaboration between the Penn State GeoVISTA Center and the Harrisburg (PA, USA) Bureau of Police. The key unmet needs identified through the interviews directly informed the conceptual design of the *GeoVISTA CrimeViz* application, providing positive evidence for speaking with the targeted end users about their core needs prior to development. Important design elements of *GeoVISTA CrimeViz* drawn from the interviews include: a web-based architecture for real-time loading of internal and external information sets, persistent interface controls and help documentation to improve the transparent usability of the application for use by all personnel within the Harrisburg Bureau of Police, multiple geographic and temporal representations that are live-linked for coordinated interaction, and contextual geographic information layers and advanced spatiotemporal analyses oriented towards strategic crime analysis. Since the initial needs assessment interview study reported here, we have completed several interface evaluation-refinement loops with the Harrisburg Bureau of Police following a user-centered design approach. The application was transitioned into use by the Harrisburg Bureau of Police for spatiotemporal crime analysis in 2012.

Acknowledgments

This research was funded in part by the Visual Analytics for Command, Control, and Interoperability Environments (VACCINE) project, a center of excellence of the Department of Homeland Security. We also would like extend our thanks to the seven law enforcement agencies that participated in the interview study.

References

Andrienko, G. L., & Andrienko, N. V. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, *13*(4), 355–374.
Andrienko, N., Andrienko, G., & Gatalsky, P. (2003). Exploratory spatio-temporal visualization: An analytical review. *Journal of Visual Languages and Computing*, *14*, 503–541.
Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, *27*, 93–115.
Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, *38*, 5–22.
Bair, S. (2000). ATAC: A tool for tactical crime analysis. *Crime Mapping News*, *2*, 9.
Bertin, J. (1983). Semiology of graphics: Diagrams, networks, maps. Madison, WI: University of Wisconsin Press.
Bertrand, J. T., Brown, J. E., & Ward, V. M. (1992). Techniques for analyzing focus group data. *Evaluation Review*, *16*(2), 198–209.
Block, C. R. (1995). *STAC hot spot areas: A statistical tool for law enforcement decisions.* Washington, D.C.: National Institute of Justice.
Boba, R. (2001). *Introductory guide to crime analysis and mapping.* Washington, DC: Office of Community Oriented Policing Services.
Boba, R. (2005). *Crime analysis and crime mapping.* Thousand Oaks, CA: Sage.
Bowers, K. J., Johnson, S. D., & Pease, K. (2004). Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology*, *44*, 641–658.
Brantingham, P. L., & Brantingham, P. J. (1981). Notes on the geometry of crime. In P. J. Brantingham, & P. L. Brantingham (Eds.), *Environmental Criminology* (pp. 27–54). Prospect Heights, IL: Waveland Press.
Brown, D. (1998). The Regional Crime Analysis Program (ReCAP): A framework for mining data to catch criminals. *Paper presented at the International Conference on Systems, Man, and Cybernetics, San Diego, CA.*
Bruce, C. W. (2008). *Police strategies and tactics: What every analyst shoud know.* International Association of Crime Analysts.
Bruce, C., & Ouellette, N. (2008). Closing the Gap Between Analysis and Response. *The Police Chief*, *75*(9), 30–32.
Brunsdon, C., Carcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, *31*, 52–75.
Buetow, T., Chaboya, L., O'Toole, C., Cushna, T., Daspit, D., Petersen, T., et al. (2003). A spatio temporal visualizer for law enforcement. *Lecture Notes in Computer Science*, *181–194*.
Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimension data visualization. *Journal of Computational and Graphical Statistics*, *5*(1), 78–99.
Cahill, M., & Mulligan, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review*, *25*(2), 174–193.
Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, *21*, 4–28.
Cheetham, R. (2010). HunchLab: Spatial data mining for intelligence-driven policing. *Paper presented at the Annual Meeting of the Association of American Geographers, Washington, DC.*
Chen, J. (2009). *Exploratory learning from space-attribute aggregated data: A geovisual analytics approach.* (Ph.D.). University Park, PA: The Pennsylvania State University.
Chen, J., Roth, R. E., Naito, A. T., Lengerich, E. J., & MacEachren, A. M. (2008). Enhancing spatial scan statistic interpretation with geovisual analytics: An analysis of US cervical cancer mortality. *International Journal of Health Geographics*, *7*, 57.
Chung, W., Chen, H., Chaboya, L. G., O'Toole, C. D., & Atabakhsh, H. (2005). Evaluating event visualization: A usability study of COPLINK spatio-temporal visualizer. *International Journal of Human-Computer Studies*, *62*, 127–157.
Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*, 588–608.
Conley, J., Gahegan, M., & Macgill, J. (2005). A genetic approach to detecting clusters in point data sets. *Geographic Analysis*, *37*, 286–314.
Dent, B. D. (1999). *Cartography: Thematic map design.* Boston, MA: McGraw-Hill.
Dey, I. (1993). *Qualitative data analysis: A user-friendly guide for social scientists.* London, England: Routledge.
DiBiase, D., MacEachren, A. M., Krygier, J. B., & Reeves, C. (1992). Animation and the role of map design in scientific visualization. *Cartographic and Geographic Information Systems*, *19(*(4), 201–214 (265-266).
Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping Crime: Understanding Hot Spots.* Washington, D.C.: National Institute of Justice.
Eck, J., & Spelman, W. (1987). *Problem-solving: problem-oriented policing in Newport News.* Police Executive Research Forum.
Emig, M. N., Heck, R. O., & Kravitz, M. (1980). *Crime analysis–A selected bibliography.* Rockville, MD: National Institute of Justice.
Getis, A., Drummy, P., Gartin, J., Gorr, W., Harries, K., Rogerson, P., et al. (2000). Geographic information science and crime analysis. *URISA Journal*, *12*(2), 7–14.
Goldstein, H. (1979). Improving policing: A problem-oriented approach. *Crime & Delinquency*, *25*(2), 236–258.
Goodchild, M. F. (1992). Geographic information science. *International Journal of Geographical Information Science*, *6*(1), 31–45.
Gottlieb, S., Arenberg, S., & Singh, R. (1994). *Crime analysis: From first report to final arrest.* Montclair, CA: Alpha Publishing.
Griffith, D. A. (1987). *Spatial Autocorrelation: A Primer.* Washington, D.C.: The Association of American Geographers.
Grubesic, T. H., Mack, E., & Murray, A. T. (2007). Geographic Exclusion: Spatial analysis for evaluating the implications of Megan's Law. *Social Science Computer Review*, *25*(2), 143–152.
Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science Association*, *24*(1), 155–168.
Harries, K. (1999). *Mapping crime: Principle and practice.* Washington, D.C.: National Institute of Justice, Crime Mapping Research Center.
Hutwagner, L. C., Thompson, W. W., & Seeman, G. M. (2003). The bioterrorism preparedness and response early aberration reporting system. *Journal of Urban Health*, *80*(2), 89–96.
Innes, M., Fielding, N., & Cope, N. (2005). 'The appliance of science?': The theory and practice of crime intelligence analysis. *British Journal of Criminology*, *45*(1), 39.
Jackson, J. L., & Bekerian, D. A. (1997). Does offender profiling have a role to play? In J. L. Jackson, & D. A. Bekerian (Eds.), *Offender profiling: theory, research and practice*. West Sussex: John Wiley & Sons.
Jefferis, E. S. (1998). *A multi-method exploration of crime hot spots: SaTScan results.* Washington, D.C.: National Institute of Justice, Crime Mapping Research Center.
Klippel, A., Hardisty, F., & Weaver, C. (2009). Star Plots: How shape characteristics influence classification tasks. *Cartography and Geographic Information Science*, *36*(2), 149–163.

Krygier, J., & Wood, D. (2005). *Making maps: A visual guide to map design for GIS.* New York, NY, USA: The Guilford Press.
Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*, 1481–1496.
Kulldorff, M. (2010). SaTScan v9.0: Software for the spatial and space-time scan statistics. Information Management Services.
Langran, G. E. (1992). *Time in geographic information systems.* Bristol, PA: Taylor & Francis.
LeBeau, J. L. (2000). *Demonstrating the analytical utility of GIS for police operations.* Rockville, MD: US Department of Justice, National Institute of Justice.
Levine, N. (2006). Crime mapping and the CrimeStat program. *Geographic Analysis*, *38*, 41–56.
Lodha, S. K., & Verma, A. (1999). Animations of crime maps using Virtual Reality Modeling Language. *Western Criminology Review*, *1*(2).
Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science.* West Sussex, England: John Wiley & Sons.
MacEachren, A. M. (1994). Visualization in modern cartography: Setting the agenda. In A. M. MacEachren, & D. R. F. Taylor (Eds.), *Visualization in modern cartography* (pp. 1–12). Oxford, England: Pergamon.
MacEachren, A. M., Brewer, C. A., & Pickle, L. W. (1998). Visualizing georeferenced data: Representing reliability of health statistics. *Environment and Planning A*, *30*, 1547–1561.
MacEachren, A. M., & DiBiase, D. (1991). Animated maps of aggregate data: Conceptual and practical problems. *Cartographic and Geographic Information Science*, *18*(4), 221–229.
MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D., & Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, *13*(4), 311–334.
Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., et al. (2010). A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, *16*(2), 205–220.
Mamalian, C. A., & La Vigne, N. G. (1999). *Research preview: The use of computerized crime mapping by law enforcement: Survey results.* Washington, DC: U.S. Department of Justice, National Institute of Justice.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.
Monmonier, M. (1990). Strategies for the visualization of geographic time-series data. *Cartographica*, *27*(1), 30–45.
Monmonier, M., & Gluck, M. (1994). Focus groups for design improvement in dynamic cartography. *Cartography and Geographic Information Science*, *21*(1), 37–47.
Nakaya, T., & Yano, K. (2010). Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density esimtation and scan statistics. *Transactions in GIS*, *14*(3), 223–239.
Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Science*, *1*, 335–358.
O'Reilly, T. (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, *17*.
Osborne, D., & Wernicke, S. (2003). *Introduction to crime analysis: Basic resources for criminal justice practice.* Binghampton, NY: Haworth Press.
O'Shea, T. C., & Nicholls, K. (2003). *Crime analysis in America: Findings and recommendations.* Washington, D.C.: Office of Community Oriented Policing Services, U.S. Department of Justice.
Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, *84*(3), 441–461.
Ratcliffe, J. (2000). Implementing and integrating crime mapping into a police intelligence environment. *International Journal of Police Science & Management*, *2*(4), 313–323.
Ratcliffe, J. H. (2004). The Hotspot Matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, *5*(1), 5–23.
Ratcliffe, J. H. (2009). Crime Mapping: Spatial and temporal challenges. In A. R. Piquero, & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 5–24). New York City, NY: Springer Science.
Ratcliffe, J. H., & McCullagh, M. J. (1998). Aoristic crime analysis. *International Journal of Geographic Information Science*, *12*(7), 751–764.
Robinson, A. C. (2008). Collaborative synthesis of visual analytic results. *Paper presented at the Visual Analytics Science and Techology, Columbus, OH.*
Robinson, A. C. (2009). *Needs assessment for the design of information synthesis visual analytics tools.*
Robinson, A. C., Roth, R. E., & MacEachren, A. M. (2011). Designing a web-based learning portal for geographic visualization and analysis in public health. *Health Informatics*, *17*, 191–208.
Rose, S. (2008). Community Safety Mapping Online System: Mapping reassurance using survey data. In Chainey, S. Tompson, L. (Eds.), *Crime Mapping Case Studies: Practice and Research* (pp. 93–102). John Wiley & Sons, Ltd.
Rossmo, D. K., & Velarde, L. (2008). Geographic profiling analysis: principles, methods and applications. In S. Chainey, & L. Tompson (Eds.), *Crime Mapping Case Studies: Practice and Research* (pp. 35–43). John Wiley & Sons, Ltd.
Roth, R. E. (2009). A qualitative approach to understanding the role of geographic information uncertainty during decision making. *Cartographic and Geographic Information Science*, *36*(4), 315–330.
Roth, R. E. (2010). Dot density maps The Encyclopedia of Geography. Thousand Oaks, CA: SAGE, 787–790.
Roth, R. E. (2011). *Interacting with Maps: The science and practice of cartographic interaction.* (PhD). University Park: The Pennsylvania State University.
Roth, R. E. (2012). Cartographic interaction primitives: Framework and synthesis. *The Cartographic Journal*, *49*(4), 376–395.
Roth, R. E., Robinson, A., Stryker, M., MacEachren, A. M., Lengerich, E. J., & Koua, E. (2008). Web-based geovisualizaiton and geocollaboration: Applications to public health. *Paper presented at the Joint Statistical Meeting, Invited Session on Web Mapping, Denver, CO.*
Roth, R. E., & Ross, K. S. (2009). Extending the Google Maps API for event animation mashups. *Cartographic Perspectives*, *64*, 21–40.
Roth, R. E., Ross, K. S., Finch, B. G., Luo, W., & MacEachren, A. M. (2010). A user-centered approach for designing and developing spatiotemporal crime analysis tools. *Paper presented at GIScience 2010, Zurich, Switzerland.*
Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology*, *94*(4), 774–802.
Shaw, C., & McKay, H. (1942). *Juvenile deliquency and urban areas.* Chicago, IL: University of Chicago Press.
Sinton, D. (1978). The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Paper presented at the Harvard papers in GIS #7, Cambridge, Massachusetts.*
Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2005). *Thematic cartography and geographic visualization* (2nd ed.). Upper Saddle River, NJ, USA: Pearson Prentice Hall.
Spelman, W. (1995). Criminal careers of public places. In J. E. Eck, & D. Weisburd (Eds.), *Crime and Place.* Monsey, NY: Criminal Justice Press.
Sutherland, E. (1934). *Principles of Criminology* (2nd ed.). Philadelphia, PA: J.B. Lippincott.
Sutherland, E. H., Cressey, D. R., & Luckenbill, D. F. (Eds.). (1992). *Principles of Criminology* (11th ed.). Landhan, MD: AltaMira Press.
Townsley, M. (2008). Visualising space time patterns in crime: The Hotspot Plot. *Crime Patterns and Analysis*, *1*(1), 61–74.
Tufte, E. (1983). *The visual display of quantitative information* (2nd ed.). Cheshire, Connecticut: Graphics Press LLC.
Wallace, T. R. (2011). A new map sign typology for the GeoWeb. *Paper presented at the International Cartographic Conference, Paris, France.*
Walsh, W. (2001). COMPSTAT: An analysis of an emerging police managerial paradigm. *Policing: An International Journal of Police Strategies & Management*, *24*(3), 347–362.
Weisburd, D., Mastrofski, S., McNally, A., & Greenspan, R. (2002). Reforming to preserve: COMPSTAT and strategic problem solving in American policing. *Criminology & Public Policy*, *2*, 421–456.
White, J. J. D., & Roth, R. E. (2010). TwitterHitter: Geovisual analytics for harvesting insight from volunteered geographic information. Paper presented at GIScience 2010, Zurich, Switzerland.
Wilcox, P., Land, K. C., & Hunt, S. (2003). *Criminal circumstance: A dynamic multi-contextual criminal opportunity theory.* New York, NY: Walter de Gruyter.
Wilson, R. E. (2007). The impact of software on crime mapping: An introduction to a special journal issue of Social Science Computing Review on crime mapping. *Social Science Computer Review*, *25*(2), 135–142.
Wilson, O. W., & McLaren, R. C. (1977). *Police Adminstration* (4th ed.). New York, NY: McGraw-Hill.
Wolff, M., & Asche, H. (2009). Geovisualization approaches for spatio-temporal crime scene analysis - Towards 4D crime mapping. *Lecture Notes in Computer Science*, *5718*, 78–89.
Yildiz, M. (2007). E-Government research: Reviewing the literature, limitations, and ways forward. *Government Information Quarterly*, *24*(3), 646–665.
Zeng, D., Chang, W., & Chen, H. (2004). A comparative study of spatio-temporal hotspot analysis techniques in security informatics. *Paper presented at the Intelligent Transportation Systems Conference, Washington, DC.*
Zhao, J. L., Bi, H. H., Chen, H., Zeng, D. D., Lin, C., & Chau, M. (2006). Process-driven collaboration support for intra-agency crime analysis. *Decision Support Systems*, *41*, 616–633.

Robert E. Roth is an Assistant Professor of Geography in the University of Wisconsin–Madison, Department of Geography. Additional information about Roth's teaching and research interest can be viewed at http://www.geography.wisc.edu/faculty/roth.

Kevin S. Ross is a Systems Design Architect at Nsite LLC. Additional information about Ross can be viewed at http://ksross.com.

Benjamin G. Finch is an undergraduate student majoring in Geography, with a specialization in GIScience, and a student intern in the Penn State GeoVISTA Center.

Wei Luo is a PhD candidate in the Penn State Department of Geography and a researcher in the Penn State GeoVISTA Center. Additional information about Luo can be viewed at http://www.personal.psu.edu/wul132/.

Alan M. MacEachren directs the GeoVISTA Center (http://www.geovista.psu.edu) and is a Professor of Geography and Affiliate Professor of Information Sciences and Technology at Penn State University. For details, see http://www.geovista.psu.edu/members/maceachren/.

Symbol Store Reviewer Report (July, 1014)

Scott Pezanowski and Alan M. MacEachren

Introduction

The development of Symbol Store www.symbolstore.org necessitated a standardized way for map symbol users and developers to contribute their symbols for the mapping community. The Symbol Reviewer is a website that allows users to contribute symbols in various formats, rate those symbols, and add descriptive metadata to allow them to be retrieved through Symbol Store.

Symbol Reviewer Functionality and Implementation

The Symbol Reviewer is designed to be an extensible system that can support both simple upload of map symbols and structured assessment of symbol sets. The primary components of the base application is a web interface that enables users to upload symbol sets in either SVG or PNG format, have thumbnail images generated for display in Symbol Store, add metadata and ratings, and publish the symbols to the Symbol Store.

The Symbol Reviewer architecture takes advantage of the web content management system, Drupal. In addition, multiple third party Drupal modules were used and modified to support functionality. One advantage provided by Drupal is its default user authentication, which allows only credentialed users to contribute symbols. The Symbol Reviewer user interface uses a custom theme developed to match the look and feel of the Symbol Store, and custom code allowing for communication with the Symbol Store API and to display results.

Once logged into the Symbol Reviewer, the user is led through a step by step process to upload symbols, add descriptive metadata, and publish their symbols to the Symbol Store. The first step is the upload of either raster image symbols (in PNG or JPG format) or vector symbols (in SVG format). These symbols can be uploaded one by one or with multiple symbols contained in a ZIP file. The symbols are uploaded to the Symbol Store server (accessible only to the contributor until published), and a response is sent back to the Reviewer with the uploaded symbol URLs. The symbol file name is used as the default symbol name, but the contributor is free to edit that name.

After the results of the upload are returned, the user may change the symbol name (if desired) and add metadata, such as a description and keywords, to each of the symbols through standard HTML form input elements dynamically generated by custom PHP code. The PHP code primarily parses the results from the upload and constructs the HTML needed for the display of the symbols and the metadata form fields.

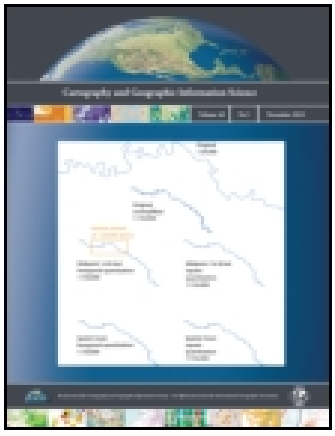
Once the user is satisfied with the metadata for each symbol, they click a button in the Reviewer that publishes the symbols to the Symbol Store so that they are then directly available to Symbol Store users.

The overall process from symbol upload to publication is described in step by step instructions presented in the Reviewer.

Currently, the Reviewer accepts symbols in PNG, JPG, and SVG formats. There is a plan to also support Esri Style files. However, do to limitations in Esri ArcObjects programming API, along with the challenge of dealing with ownership of fonts that are often used in Style files, the Reviewer does not currently support the upload of Esri Style Files. Symbol Store, however, does support display and download of Style Files (that have been manually added at Penn State). Supporting Style File upload generally will require development of a stand-alone tool that contributors can download and use to preprocess the Style Files in order to create the thumbnail images needed for browsing the symbols in Symbol Store. More details are provided in the Future Work section below.

Future work

Esri Style Files is a widely used proprietary format due to the prominence of ArcGIS for geospatial analysis and mapping in both industry and government. In order to extract symbols and generate the raster thumbnails needed by Symbol Store, Esri's ArcObject programming API is needed. Through much research and development, and through work with Esri support, we identified a system flaw in the Esri API that prevents certain individual symbols in Style Files from being converted to raster images when processed by the ArcObjects code. The failure typically causes the whole system to hang, requiring a system restart. Also, because symbols in Style Files are often based on fonts, it is necessary for those fonts to be installed on the Symbol Store server for processing. It is impractical for Symbol Store to support all fonts that could be used. These two issues with processing Style Files have made their support in the Symbol Store and Symbol Reviewer problematic. The solution we have identified for future development (subject to availability of resources) is to create a standalone distributable application that can be downloaded from the Reviewer site and used on the Reviewer user's own computer. This application could preprocess the symbols, generating the thumbnail images using the fonts installed on the contributor's computer (thus avoiding the need for Symbol Store to have the font installed). Once processed, the symbols could be automatically uploaded to the Reviewer if an Internet connection is available. In addition, it would be possible to output the files required by the Reviewer for later upload by the user.



Cartography and Geographic Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcag20

Symbol Store: sharing map symbols for emergency management

Anthony C. Robinson [a], Scott Pezanowski [a], Sarah Troedson [a], Raechel Bianchetti [a], Justine Blanford [a], Joshua Stevens [a], Elaine Guidero [a], Robert E. Roth [b] & Alan M. MacEachren [a]

[a] Department of Geography, GeoVISTA Center, The Pennsylvania State University, University Park, PA, 16802, USA

[b] Department of Geography, University of Wisconsin-Madison, Madison, WI, 53706, USA

Published online: 03 Jun 2013.

To cite this article: Anthony C. Robinson, Scott Pezanowski, Sarah Troedson, Raechel Bianchetti, Justine Blanford, Joshua Stevens, Elaine Guidero, Robert E. Roth & Alan M. MacEachren (2013) Symbol Store: sharing map symbols for emergency management, Cartography and Geographic Information Science, 40:5, 415-426, DOI: 10.1080/15230406.2013.803833

To link to this article: http://dx.doi.org/10.1080/15230406.2013.803833

Cartography and Geographic Information Science, 2013
Vol. 40, No. 5, 415–426, http://dx.doi.org/10.1080/15230406.2013.803833

Taylor & Francis
Taylor & Francis Group

Symbol Store: sharing map symbols for emergency management

Anthony C. Robinson[a]*, Scott Pezanowski[a], Sarah Troedson[a], Raechel Bianchetti[a], Justine Blanford[a], Joshua Stevens[a], Elaine Guidero[a], Robert E. Roth[b] and Alan M. MacEachren[a]

[a]*Department of Geography, GeoVISTA Center, The Pennsylvania State University, University Park, PA 16802, USA;* [b]*Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA*

(*Received 21 December 2012; accepted 23 April 2013*)

Maps are a primary means for supporting information sharing and collaboration in emergency management and crisis situations. While a variety of formalized map symbol standards for emergency contexts exist, they have not been widely adopted by mapmakers. Informal symbol conventions are commonly used within emergency management stakeholder groups, but until now there has not been a flexible mechanism for discovering, sharing, and previewing these symbol sets among mapmakers. In this paper, we describe the design and development of the Symbol Store, a visually enabled, web-based interactive tool intended to help mapmakers share point symbols. The Symbol Store allows users to browse symbols by keyword, category tags, and contributors. It also allows for symbols to be previewed on realistic maps prior to download. An initial prototype of the Symbol Store was evaluated by flood mapping experts from the State of California, and the results of this user study led to multiple refinements now implemented in the public version of Symbol Store located at www.symbolstore.org.

Keywords: map symbology; emergency management; web applications; standards

1. Introduction

Maps offer a critical form of communication and function as important analytical tools for distributing information in emergencies when it becomes necessary for stakeholders to develop a common operational picture (COP). Maps are also essential in other phases of emergency management that include preparedness, recovery, and mitigation activities. During emergencies, all tasks are time sensitive, and the speed with which one can read and interpret a map can signal the difference between lives and property saved or lost. Supporting rapid and efficient consumption of geographic data can be significantly assisted by having map design and symbology standards that users can learn and apply in advance of an emergency event. Additionally, when teams of collaborators from different organizations come together in an emergency operations center (EOC), they can quickly learn how to read and interpret maps by learning from an existing map design and symbol standard.

This paper describes the design, development, and initial evaluation of a new, web-based mechanism for searching for and retrieving point symbols to support digital cartography. We present a novel system called the Symbol Store, based on a long-term, iterative study of mapmakers' symbol needs at the US Department of Homeland Security (DHS) (see previous work in Robinson, Roth, and MacEachren, 2010, 2011; Robinson et al. 2012). The Symbol Store can help aid interoperability in emergency management situations where multiple local, state, and federal authorities collaborate on the development of a wide range of mapping products to develop situational awareness and marshal resources. Emergency management mapping tasks frequently include pre-disaster planning activities and strategic recovery efforts which can also stand to benefit from mechanisms that improve symbol interoperability. Currently, there are few elegant or efficient means for such stakeholders to access and share each other's methods for representing features on maps.

In the following sections, we describe relevant prior work on symbol standardization and previous means for sharing map symbology. Then we describe the key design goals and features of the Symbol Store. This is followed by a section describing an initial user evaluation conducted with flood mapping experts to test core features of the Symbol Store. Finally, we conclude with planned next steps to further refine and exploit the unique capabilities of the Symbol Store.

2. Background

Our focus in this research is on sharing point symbols for maps. While line and polygon features are also quite important, we have focused on symbol interoperability first with respect to point features, which are among the most common types of symbols used in emergency management tasks. These tasks can include disaster mitigation and response planning activities, as well as direct response and situational assessment mapping in the immediate aftermath of a disaster (Cutter 2003). They can also include long-term planning and

*Corresponding author. Email: arobinson@psu.edu

remediation mapping efforts as communities engage in recovery efforts. As we discuss below, the majority of existing symbol standards focus on point symbols entirely or in large part. In addition, funding support for this research from DHS was directed specifically toward the task of exploring point symbol interoperability since it represents their most pressing area of need.

Two primary influences for our work include past efforts to develop map symbol standards and existing approaches for sharing map symbols.

2.1. Symbol standards

In recent decades there have been a wide range of symbol standardization efforts that have resulted in the design and dissemination of point symbol references for use in common mapping contexts. These efforts include the development of the US Military Standard MIL-STD-2525 (Department of Defense 2008), which prescribes a set of representations and modifiers to support map interoperability in military planning and combat situations. As a result of its adoption by NATO, this standard is used and modified by other stakeholders as well in a wide range of countries and contexts. For example, NATO countries frequently support humanitarian relief missions with military-led assistance, and non-governmental agencies engaged in relief efforts will often make use of the same symbology to ensure interoperability.

Symbol standards for non-military purposes include the US ANSI INCITS 41-2006 (ANSI 2006) Homeland Security Mapping Standard which prescribes a set of point symbols, symbol outlines, and graphical modifiers to support emergency management interoperability. The ANSI set was designed to function as a single standard used by local, state, and federal stakeholders in US domestic emergency management situations. Despite concerted efforts to develop the ANSI standard for point-symbol symbology for exclusive use across DHS and with DHS partners (Dymon 2003; Dymon and Mbobi 2005), the standard has not achieved widespread adoption within DHS, or outside DHS in commercial tools like WebEOC (www.esi911.com) that are commonly used by state and local emergency management organizations. In previous research, we found that DHS mapmakers had symbolization needs that were unmet by the ANSI set, that many ANSI symbols were too graphically intricate to use in many situations, and that task-specific (but non-official) symbol sets already existed that were more commonly used to support emergency management mapmaking (Robinson, Roth, and MacEachren, 2011). Furthermore, recent usability and utility evaluation of these symbols revealed that they have serious weaknesses when they are used to support emergency management tasks (Akella 2009).

Other symbol standards exist for emergency mapping and related contexts, including an international standard for humanitarian demining symbols (Kostelnick et al. 2008) and a recent homeland security symbol standardization project in Canada (Sondheim, Charmley, and Leeming 2010), which led to the development of a web repository (www.EMSymbology.org) where those symbols could be downloaded. Despite the availability of these official and many non-official but widely utilized symbol sets (such as those included with common Geographic Information Systems (GIS) software packages) no single symbol standard has been widely adopted for general use across the full range of emergency situations in the US or elsewhere.

One potential reason for lack of widespread adoption of symbol standards is that their development tends to receive a great deal of attention and resources, while their dissemination, implementation, and revision does not. Emerging needs for new and modified representations are not easily met through official standardization processes. In discussing the need for symbol standardization in the 1970s, Arthur Robinson suggested that while standards are necessary and useful, they must remain open-ended so that evolution can occur as mapping requirements change (Robinson 1973). New efforts to design symbols by MapBox (www.mapbox.com/maki) and the Noun Project (www.thenounproject.com) are responding to evolutionary needs to develop simple point-of-interest symbols for web maps for the former project, and to develop a common visual language to cross language barriers for the latter. The Map Icons Collection (www.mapicons.nicolasmollet.com) allows users to contribute point symbol designs for use in Google Map mashups. New approaches for supporting symbol sharing and interoperability need to be flexible enough to support new symbol sets like these and integrate them with existing methods to ensure the best representations are used for any given mapping context.

2.2. Current methods for sharing symbols

Map symbols are most commonly shared today via digital means through common GIS software packages. While mapmakers are always able to create their own symbols, the default palettes available in GIS software see wide adoption, although the original meaning of a particular symbol may be ignored in part or whole as mapmakers reinterpret symbols for use in new representational contexts (Robinson, Roth, and MacEachren, 2011). In terms of their formal implementation, outside of the MIL-STD-2525 example noted above, where military units require its use and mapmakers as well as map readers are trained to use the standard, the other standards we reviewed may have formal endorsement but no firm requirements for mapmakers to actually use them.

In terms of tools for sharing symbols, Esri's ArcGIS supports the use of *.style files which can catalog and describe the representations used on maps in ArcGIS. We found when studying the use of ANSI symbols and other symbology at DHS that DHS mapmakers frequently made use of *.style files to curate their own task-specific (but not official) symbol collections, and in some cases would share these with other

mapmakers. Recent development by the makers of Ortelius, a cartography software package for use in Apple iOS, includes the ability to share Ortelius-compatible map symbols online via a dedicated symbol management tool (Saligoe-Simmel 2010). Most officially sanctioned map symbol standards are disseminated by government websites and shared in Esri-compatible formats. Our work with mapmakers at DHS indicated that the primary means for discovering symbols was by using web search engines to find new *.style files and fonts that include symbol markers.

We began our work on map symbology for emergency management in 2009 by researching the use and adoption of the ANSI symbol standard, identifying other symbols in use, and developing key user requirements for map symbology standards (Robinson, Roth, and MacEachren, 2011) at DHS. We followed this work with the development and evaluation of a new, more flexible and iterative process for creating map symbol standards for DHS (Robinson et al. 2012). A primary result discovered through this prior research was the need for a new platform and model for supporting symbol interoperability. While current tools like ArcGIS and Ortelius allow users to share pre-defined symbol sets using Esri *.style files or other special formats, and web repositories like the Noun Project and EMSymbology.org exist for individual symbol sets there remains a need for flexible, visually enabled tools to support dynamic symbol sharing and discovery. We believe the next step involves moving beyond simply retrieving symbols via the web, to support users who wish to contribute symbols to an evolving repository, to support users who want to search for appropriate symbols using keywords and other metadata (something not supported on sites like www.EMSymbology.org for example), to preview symbols on realistic maps, and to allow communities of mapmakers to iteratively rate and refine symbol collections to create new *de facto* map symbol standards. To fill this gap, we have developed the web-based Symbol Store to provide a usable and useful tool for contributing, browsing, rating, and assembling customized symbol palettes to support mapmakers at DHS and beyond. The following sections describe our progress so far in meeting that goal.

3. Symbol store

To support map symbol interoperability we have designed a web-based prototype tool for discovering, sharing, and retrieving map symbols called the Symbol Store. The following sections describe Symbol Store's core design objectives, its technical underpinnings, our demonstrated progress toward implementing a working prototype, and the results of a user evaluation to test the first working prototype to suggest future refinements.

3.1. Design objectives

Based on the results of our prior work to study the utility of the ANSI standard for DHS mapmakers and to design a new process to developing more flexible symbol standards, we developed four core system functionality targets for shaping the design of our first Symbol Store prototype:

3.1.1. Search for and retrieve symbols

The most basic design goal for Symbol Store is to support keyword searches for symbols in use by agencies across DHS (and potentially wider audiences as the tool becomes open to other groups as well). Symbols retrieved via keyword search can be collected and downloaded as an ESRI *.style file or in other common formats for immediate application.

3.1.2. Preview symbols on realistic maps

After selecting a subset of symbols from the Symbol Store search interface, users can preview their symbols on a variety of realistic base maps. The map preview feature in Symbol Store provides a range of basic design controls to allow users to change the map scale, feature density, labeling, coloring, and other common map design aspects in order to preview the suitability of their symbols prior to downloading them.

3.1.3. Browse for symbols

Apart from searching for specific symbols, users can browse symbols by time (most recent uploads, for example), contributor (symbols from a specific agency, for example), and symbol categories (all symbols corresponding to infrastructure, for example). This supports flexible means for discovery, for instances in which keyword searches are not as efficient or effective. It also allows popularity measures such as subjective ratings to become one of the means through which new symbols can be discovered and disseminated.

3.1.4. Share symbols

Users can contribute symbols to Symbol Store by uploading an Esri *.style file and associated fonts through the Symbol Store interface. After uploading symbols, users are able to tag individual symbols or groups of symbols to assign keywords, category names, and other important metadata information.

3.2. System architecture

To accomplish our design objectives we designed an interactive web interface to encourage use by a wide range of users across multiple platforms. Simple and effective interoperability is essential to support DHS users coming from a diverse set of organizations, and to ensure more

widespread adoption of formal standards by other related emergency management communities in the longer term.

The Symbol Store interface (Figure 1) runs in a standard web browser using the Adobe Flash plugin. The interface itself was constructed using Adobe Flash Catalyst, an interface development environment that integrates with Adobe Flex, which we used to connect the Symbol Store interface to server-side components. Symbol Store is comprised of four main components, illustrated in Figure 2 and includes: 1) the Flex and ActionScript User Interface (UI); 2) the .NET CSharp web service middleware; 3) the storage system of an Apache Lucene Index (apache.lucene.org) and a Structured Query Language (SQL) Server Relational Database Management System (DB); 4) and an instance of Esri ArcGIS Server to produce live, interactive map previews with selected symbols.

A Lucene index is used to store the text metadata about symbols in the Symbol Store, including Symbol name, symbol description, keywords, user, symbol set, and other features. This allows for text searches to be performed when searching for information. When a text based search is performed, the Lucene index is queried and pointers to the symbols are returned. Next the DB is queried to retrieve those symbols that matched the Lucene index query. Users are then able to create a customized symbol set by selecting individual symbols, adding them to their symbol cart and downloading the newly created style file containing the symbols.

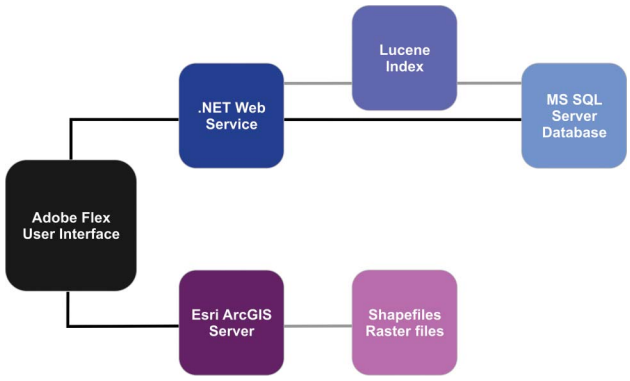


Figure 2. Architecture of Symbol Store. Symbols with metadata (e.g., keywords, set, categories, date of upload, ratings) are stored in a database and indexed using the Lucene Index. Map previews are stored and served using Esri ArcGIS Server 10.

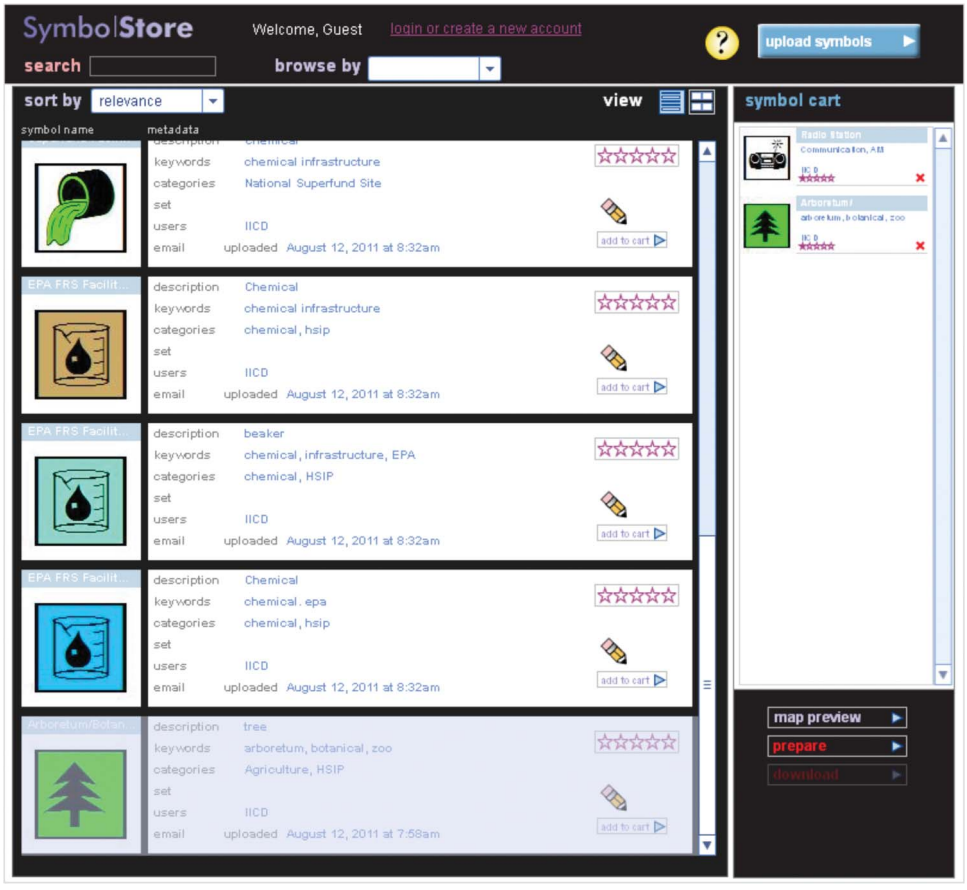


Figure 1. This screen capture shows the initial prototype Symbol Store interface with symbols contributed from Infrastructure Information Collection Division (IICD).

In addition to obtaining symbols to match their needs, users can contribute symbols to the store by uploading either Esri *.style files or TrueType font files. When these files are uploaded, symbols are stored in a SQL Server Database table record and an image thumbnail of the symbol is created using Esri's ArcObjects tools. Once symbols have been uploaded to the system and stored in the database, the user is presented with a metadata editing interface where they have the option to add metadata for the symbol collection as a whole and/or for individual symbols. Symbols are easily retrieved through an SQL query when the user decides to download symbols from their symbol cart.

The final component of the Symbol Store allows for the preview of symbols on a map. The Symbol Store map preview tool allows users to preview the symbols that they have placed in their shopping cart. The map preview has been crafted to allow users to choose a base map style and scale, and then adjust many of the symbol properties, such as size, density, and labeling. This allows users to compare different potential symbol sets and choose the set that is the most legible given their choice in cartographic design. Three main control types have been designed to support on-the-fly visualization of symbol properties. The base map type can be chosen using a button feature. Horizontal sliders are used to adjust for symbol size, map scale, point density, and label size. Finally, a user's choice in labeling or not labeling the features uses a checkbox control.

The base map is comprised of several ArcGIS shapefiles representing transportation, political boundaries, and water bodies. The features were acquired from the US National Atlas for national and regional scales, and several state-level agencies for the state, county, and city level data. For state-level data, we chose Washington State as the example, and have used data from the Washington Department of Transportation, the Washington Department of Environmental Quality, and King County's GIS Center. To display map point symbols, we have created several levels of point density for each of the scales by generating random point symbols.

3.3. Symbol store prototype features

The initial Symbol Store prototype underwent multiple rounds of internal evaluation and refinement to satisfy our stated goals to support symbol search and retrieval, symbol browsing, symbol sharing, and preview symbols on realistic maps.

The initial prototype Symbol Store search interface is shown in Figure 2. Each symbol includes metadata to show its description, a list of relevant keywords, relevant thematic categories, membership in a formal symbol standard (e.g., ANSI), users of this symbol (e.g., different divisions within DHS), contact information for the person responsible for uploading the symbol, and the date the symbol was uploaded. Users can select symbols for download or preview by adding them to the cart shown on the right of the screen in Figure 2. Once users are satisfied with the selected set of symbols, they can prepare and download the set as an Esri *.style file.

Symbols can be easily contributed to the initial prototype Symbol Store by way of a specialized upload and metadata creation interface we have implemented (Figure 3). Users can select an Esri *.style file and associated font files and upload these to the Symbol Store database. Once the Symbol Store has processed these files, a secondary interface appears (shown at the bottom of Figure 3) so that users can add metadata to the contributed symbols. For the entire collection, users can edit the contributing agency, assign symbol categories associated with the set, identify the username of the person contributing the set, and name the official symbol standard to which the symbols belong (if applicable). For individual symbols (or smaller groups of multiple symbols, selected by clicking checkboxes in the metadata editing interface), users can add descriptions, keywords, categories, users (such as agencies or departments), and ratings (from 1 star to 5 stars). Users can also edit symbol metadata once the symbols have been published in the Symbol Store.

The map preview feature (Figure 4), inspired by popular web tools for cartographic design like ColorBrewer (Harrower and Brewer 2003) and TypeBrewer (Sheesley 2007), which provide example maps to help users choose among design options, will allow users to visually evaluate the symbols that are currently in their symbol cart in a realistic map design context. The map preview tool in Symbol Store allows users to change the type of map background between common base map options, to change the color of the base map between color or black and white, and to adjust symbol sizes, the map scale, feature density, and labeling. These common design parameters are adjustable interactively to help users rapidly evaluate the symbols they have chosen against a variety of realistic map design constraints. One goal here is to decrease the amount of time mapmakers spend creating multiple design iterations to develop a particular map.

4. Evaluation

In addition to multiple rounds of internal refinement of the Symbol Store prototype, including substantial informal feedback from DHS stakeholders in monthly project meetings, we have completed an initial formative user evaluation to inform our next steps.

We recruited six mapmakers from California's Department of Water Resources (DWR) who regularly engage in emergency management mapping centered on floods and other water-related emergency events. DWR staff are also engaged in an effort to develop an internal standard for symbology, so this user group would potentially benefit from a tool like the Symbol Store when identifying candidate symbols from other sets. Our goals in this study

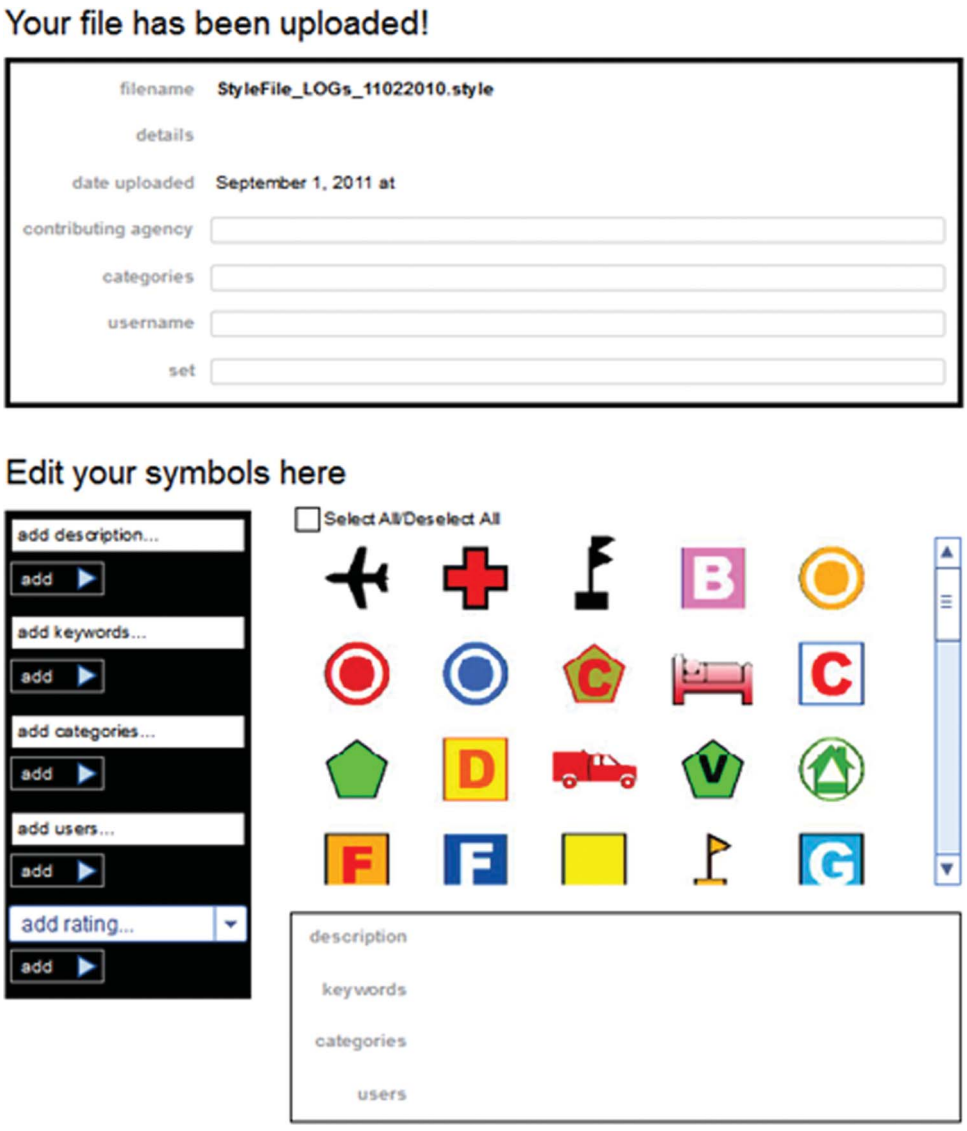


Figure 3. The Symbol Store symbol upload and metadata editing interface.

were to evaluate the utility of the Symbol Store for daily DWR mapmaking needs, to help develop their own standard set of symbols, and to characterize how well the Symbol Store could support symbol sharing for DWR once they have created a standard that could be shared.

4.1. *Evaluation procedures*

Our evaluation activities were composed of a task analysis to test key functions in the Symbol Store and an online survey to capture and elicit user feedback. We developed six tasks for DWR mapmakers to complete. In general terms, the six tasks included: uploading symbols, searching for symbols, selecting symbols to preview, previewing symbols on realistic maps, downloading symbols, and testing a downloaded Esri *.style file. After completing these tasks with the Symbol Store prototype, users were asked to rate the usability and utility of the Symbol Store in an online survey. Many of the survey questions also prompted users to provide qualitative feedback to elaborate upon their opinions and suggest specific new features or bugs to fix.

This user study was aimed at formative assessment (Buttenfield 1999) of the Symbol Store prototype in order to establish which steps to take going forward to add/remove key features and to develop a baseline understanding of the tool's usability.

4.2. *Task analysis results*

The following sections describe each of the six tasks our study participants completed and discuss survey feedback gained from questions related to each task. All

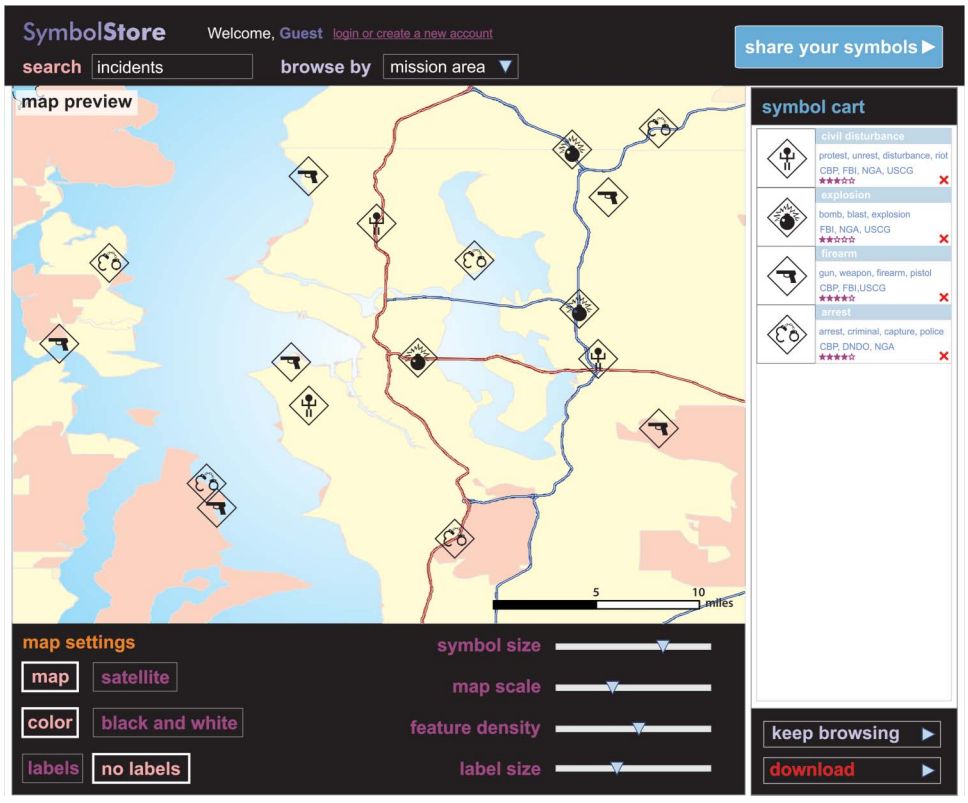


Figure 4. The map preview function (shown here in a design mockup) in the Symbol Store allows users to select from different map types and to control other symbol and map design attributes to preview symbols placed in the symbol cart.

survey questions aside from two multiple choice prompts used a five-point agreement scale; strongly disagree – 1, disagree – 2, neither agree nor disagree – 3, agree – 4, and strongly agree – 5. Average ratings are used in the following sections according to this scale to note participant agreement level with prompts about the usability and utility of features and functions in the Symbol Store (Figure 5). All questions were crafted in a manner to allow participants to agree or disagree with the premise of the prompt. A reference document showing the full task descriptions as well as the complete set of survey questions is available in Appendix A (www.personal.psu.edu/acr181/SymbolStore_Appendix_A.pdf).

4.2.1. *Task one: upload symbols*

Participants were asked to upload an Esri *.style file containing eight point symbols and to create metadata for those symbols by tagging them with keywords, categories, and other relevant source information. When asked to rate how easy it was to upload a .style file, the average of user

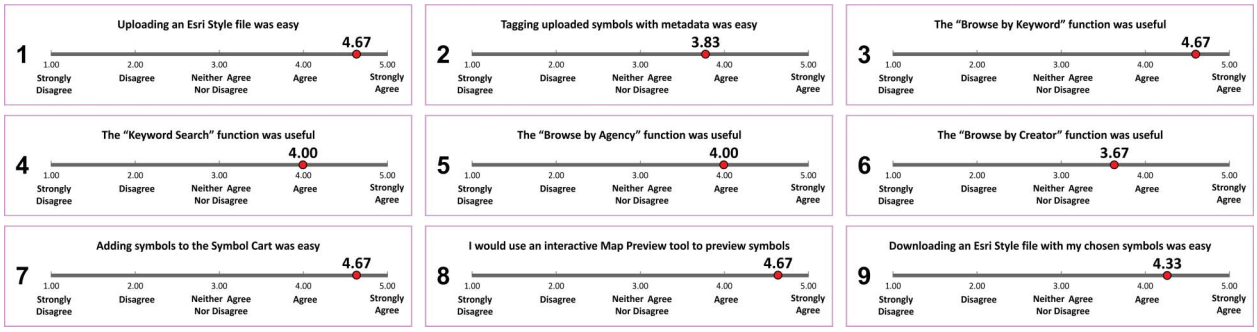


Figure 5. User rating results for key aspects of Symbol Store usability and utility.

ratings was 4.67 (Q1 in Figure 5) on the five-point agreement scale described above, indicating that users found this task easy to accomplish.

While participants generally agreed that the process of adding metadata to their symbols was easy (3.83 rating – Q2 in Figure 5), many suggestions on how to improve this part of the interface were offered. The names assigned to each symbol were not visible during the tagging process which led to much confusion over what each symbol was intended to be used for. Also, the text fields intended to contain the keywords and categories were not large enough. Other comments mentioned the need to apply changes in metadata to multiple symbols at one time (when assigning category names, for example).

4.2.2. *Task two: search for symbols*

In the second task, participants were asked to search for the symbols to which they had just added metadata via both keyword search and basic browsing methods. In addition to typing search terms into the keyword search box, participants used the "Browse By" option to select from a list of categories and keywords, and they could also search by clicking on keywords associated with a symbol in the symbol list itself. Additionally, during this task, participants were given the option to update or change metadata they had previously entered for any given symbol. Our users found the "Browse By" option to explore existing keywords most useful (4.67 rating), followed by traditional keyword search (4.0 rating). Browsing by agency (4.0 rating) and creator (3.67 rating) metadata was also judged to be useful (Q3–Q6 in Figure 5).

Qualitative feedback from this task revealed that the most common concern with the search and browse options was there was no obvious way to clear a search and return to a previous view. Additionally, there was no way to show an overview of all symbols. Some study participants also found a bug that caused keywords and categories in the list to appear without associated symbols. A small interface design issue regarding the size of the symbol name was also highlighted by several users.

4.2.3. *Task three: select symbols to preview*

In the third task, participants were asked to select approximately a dozen symbols and add them to their Symbol Cart to preview prior to download. This was a very simple task that required clicking icons on symbols of interest to add it to a "shopping cart" style interface. This feature's usability received strong support (4.67 rating, Q7 in Figure 5) and none of the participants provided comments or suggestions to improve this part of the interface.

4.2.4. *Task four: preview symbols*

Our intention for the fourth task was to have participants evaluate the map preview feature to self-assess a set of symbols added to the symbol cart. The development work to enable this feature was not completed in time for the user study, and instead our participants were given a mockup preview screen (shown in Figure 4) and asked to comment on the likelihood that they would use such a tool (4.67 rating, Q8 in Figure 5), and to rate the relative importance of each proposed map preview function.

Participants were unanimously in favor of all of the proposed functions suggested by the mockup concept. One specific additional suggestion was offered during the focus group discussion, which is to add the option to preview their symbols over a United States Geological Survey (USGS) Topographic base map.

4.2.5. *Task five: download symbols*

Once study participants had chosen their symbols they were asked to download them. The evaluated prototype supported downloads using an Esri *.style file, which is one of several common means for formalizing the look of a digital map product. This task was very straightforward and widely viewed as easy (4.33 rating, Q9 in Figure 5) and the only suggestion from participants was to implement a way to combine the "Prepare" and "Download" steps to make downloading possible with a single click.

4.2.6. *Task six: test downloaded *.style file*

The final task involved launching a new Esri ArcGIS project, adding the downloaded Esri *.style file to the map document and viewing the point symbols in the ArcGIS Style Manager or assigning those symbols to point features in the map document.

One study participant had trouble downloading the file on the first try (their subsequent attempt was successful), and another study participant found two of their chosen symbols did not download properly. The latter error is a known issue associated with having all of the necessary fonts installed locally where the *.style file is downloaded. A large proportion of symbols in Esri ArcGIS are drawn directly from custom fonts that use symbols in place of alphanumeric characters. The *.style file then connects to these fonts and draws them in specific ways based on chosen design attributes. In this case, the study participant was missing the font necessary to view two of their downloaded symbols. All other downloaded symbols worked properly for all of the study participants.

An improvement that can be made to the downloaded *. style file was suggested to incorporate the keywords associated with a symbol in the Symbol Store into the "tag" field associated with the *.style file so that those imported symbols can then be searched for using the ArcGIS symbol browsing tool. The evaluated Symbol Store prototype downloaded *.style file only contained the symbol name and Esri-software assigned default design

attributes present when the symbol was originally uploaded to the Symbol Store. Study participants expressed concern that they spent a good deal of time adding keywords and metadata to symbols and then were not able to use that information once the *.style file was downloaded to their local machine.

4.3. Evaluation summary

Our evaluation of the initial Symbol Store prototype with flood mapping users from the state of California yielded useful feedback on usability and utility. Approximately 20 small interface improvements were identified by the study participants, and a variety of major improvements were suggested:

- Implement the map preview as designed in the mockup
- Improve search behavior to retrieve additional relevant results beyond exact matching keywords
- Improve the interface look and feel to make functionality easier to find and visually distinct from search results
- Support a wider range of export formats to avoid problems with fonts
- Use tabbed pages to view symbol results incrementally
- Add a USGS topographic base map option to the map preview tool
- Develop a grid view to show a larger overview of available symbols when browsing
- Support import and retrieval of line and polygon symbols
- Implement user accounts so that draft symbol standards can be shared among a working group before being released to a wider audience
- Import/export more metadata when contributing or downloading *.style files

Overall, our evaluation results with this small group of real-world users demonstrate the potential of a tool like the Symbol Store to support usable and useful symbol discovery, retrieval, and contribution. User feedback after the survey indicated that our participants were eager to see future revisions to the prototype and for the tool to be robust enough to support regular, widespread use in their agency.

5. SymbolStore.org

Following the results of our initial evaluation with flood mapping users in California, we implemented many of their suggested changes and began transitioning the prototype Symbol Store to a public-facing site for widespread use. The Symbol Store is now available at www.symbolstore.org, and currently hosts over 2400 symbols that can be easily discovered, previewed on realistic maps, and downloaded in a variety of useful formats. Some of the major improvements made from the evaluated prototype include; a fully functioning map preview tool, a redesigned user interface to improve clarity and offer a standardized look and feel, pagination of search results to improve usability, and two new export formats (PNG and SVG) to avoid problems with sharing fonts and improve interoperability.

The improved primary interface is shown in Figure 6. In addition to the major improvements already listed, we fixed the bug that caused keywords and categories to appear that did not link to search results, and we have improved connections to Lucene to support more flexible search behavior to retrieve more results with single keywords. For example, a search for “fire” will now return anything that includes the stem of that term, so “firing range” will appear as a result rather than only exact matches for “fire.”

The map preview tool (Figure 7) now allows users to interactively assess the symbols in their symbol cart using a range of common map design controls to change the base map design, alter the size/density of symbols on the map, and explore the symbols when used at three common scales. The map preview tool leverages ArcGIS Server to generate and manipulate real-time map previews on the web client.

A major step toward supporting wider symbol interoperability is the inclusion of new symbol formats with every Symbol Store download. Users now can retrieve a single .zip file archive which includes PNG symbols at a range of useful icon sizes, an Esri *.style file, and SVG vector graphics for use in graphic design software. SVG symbol export is made possible by tracing PNG images of symbols using an automated back-end routine that leverages the Inkscape open source graphic design software (www.inkscape.org).

Symbol contributions are supported in the public www.symbolstore.org site through a new, simplified interface shown in Figure 8. Currently, we do not allow public contributions to be processed and appear automatically to prevent potential abuse, and we are exploring ways to support metadata creation and editing for public users while ensuring that contribution quality will remain high.

6. Conclusions and future work

Following our initial evaluation and refinement effort, we will focus on collecting feedback on the second-generation prototype available at www.symbolstore.org from DHS mapmakers through a series of planned practitioner workshops. In these workshops, DHS users will complete realistic symbol-related tasks using the Symbol Store and we

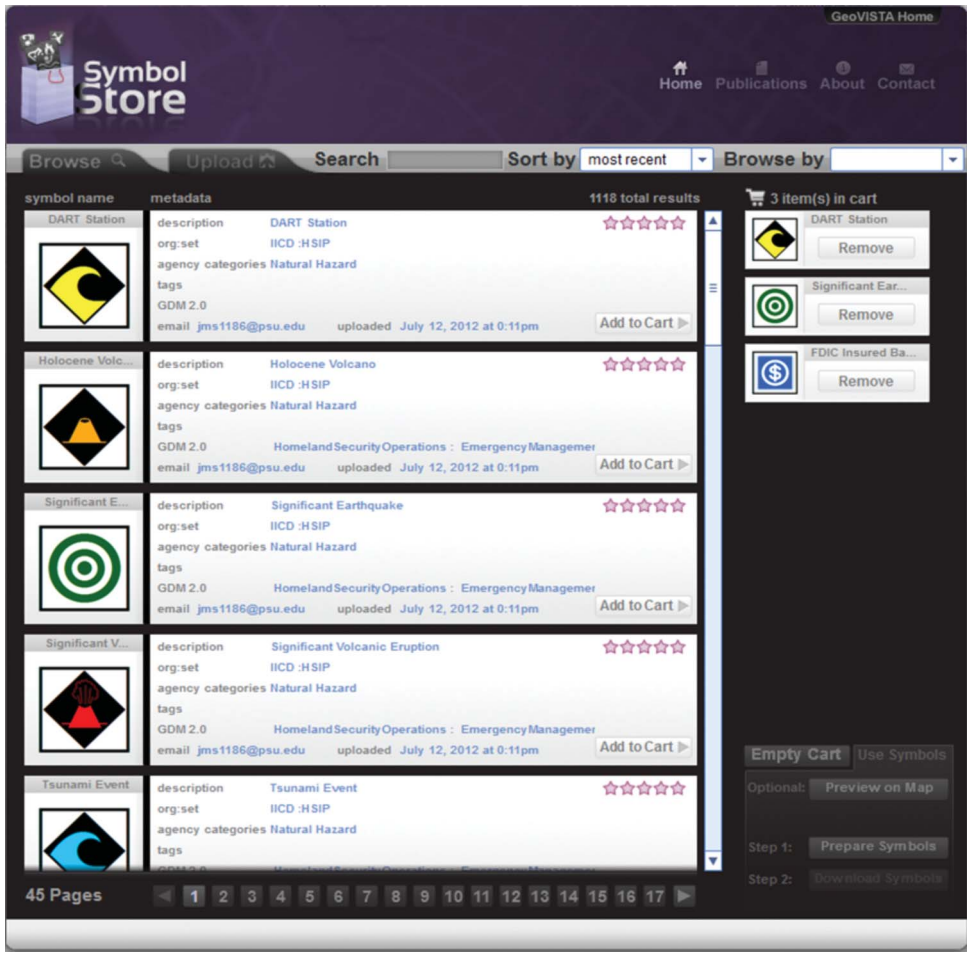


Figure 6. The redesigned Symbol Store interface is now available for public use at www.symbolstore.org.

will engage these users in group discussions to identify next steps for Symbol Store development and integration into existing DHS work practices. We anticipate that the Symbol Store will be useful for a wide range of emergency management related tasks, including pre-emergency planning, post-disaster situational assessment mapping, and long-term recovery efforts. In terms of its immediate utility during the response phase of an emergency, we would anticipate that the Symbol Store would be a quicker way to discover, download, and use symbols than their current affordances, which rely on manual web searches and informal relationships with other mapmakers (Robinson, Roth, and MacEachren, 2011) allow.

A key focus for Symbol Store development going forward will be on integrating what we have learned from symbol standard development through the standardization process we developed and the e-Symbology Portal tools we used to conduct process tests with Customs and Border Protection (CBP), Federal Emergency Management Agency (FEMA), and Infrastructure Information Collection Division (IICD) (Robinson et al. 2012). We will develop methods and techniques to move components from our standardization process into the Symbol Store interface to combine the two efforts in an elegant and effective unified environment. A key goal is to provide access to more sophisticated metadata and category standardization tools and procedures for small groups of motivated users and symbol set curators. These higher-level functions will require accounts and log-in permissions so as not to interfere with basic use by members of the public and mapmakers who simply want to quickly find and retrieve symbols.

Other challenges for future development include new methods to expand searches to return relevant results. One strategy we are currently implementing is to leverage WordNet (Miller 1995) measures of similarity between words in the English language to find relevant terms beyond an initial keyword and retrieve a wider set of relevant symbols to the user. Searching for symbols by visual similarity also remains an important, but difficult to achieve goal. Ideally, cartographers should be able to find

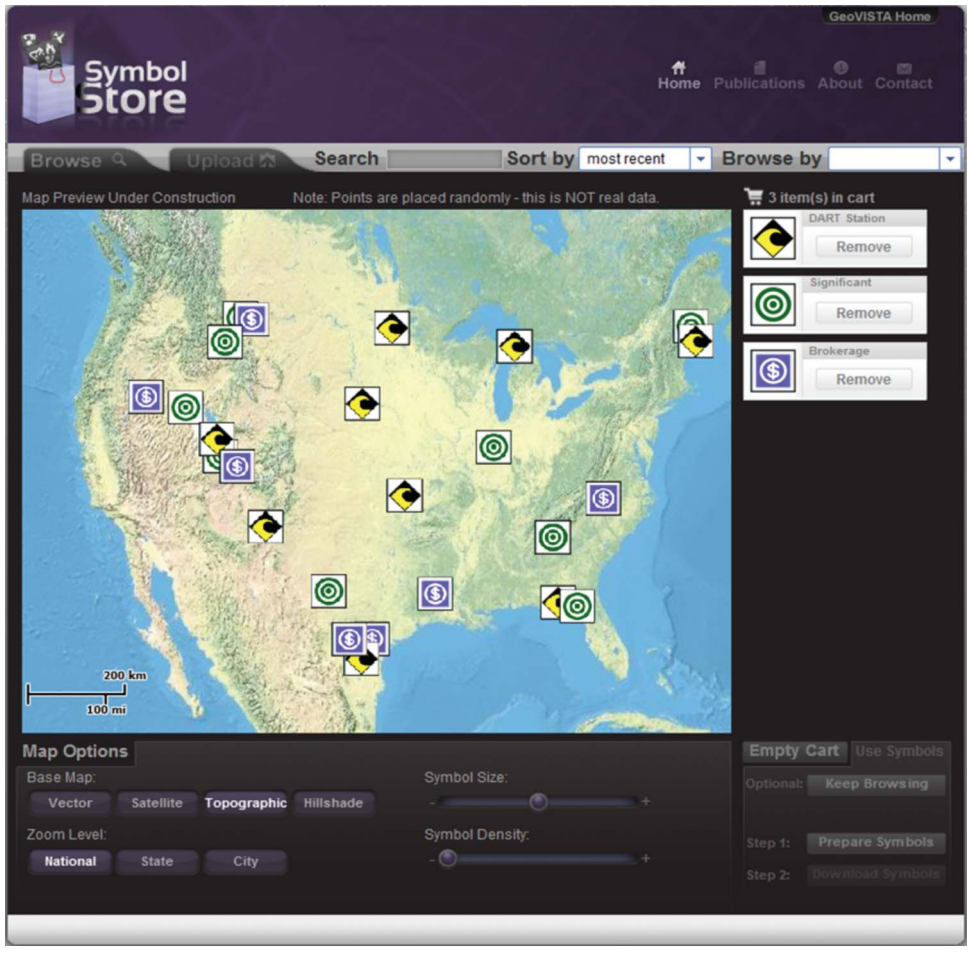


Figure 7. The map preview function of the Symbol Store allows users to visually evaluate symbols prior to download.

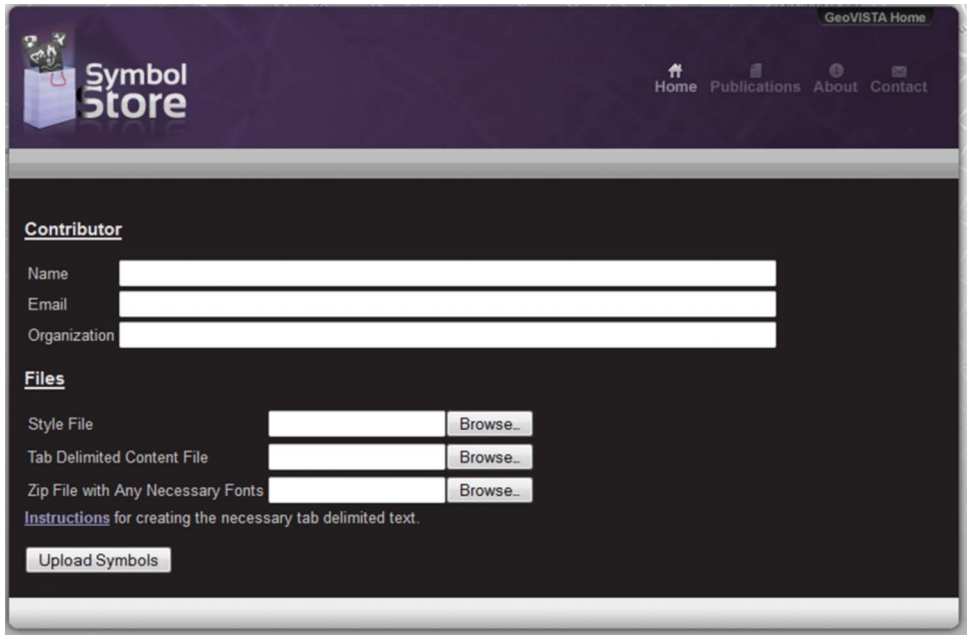


Figure 8. Public contributions of symbols to the Symbol Store can make use of the new, simplified uploader shown here.

"more like this" when viewing a particular symbol. Along those lines, it may be possible for us to develop a mechanism to crawl the web to automatically collect symbols that already exist in a wide range of formal and informal repositories.

Additionally, many issues exist when it comes to how users define, assign, and interpret categories of symbols, and definitions associated with specific symbols. Multiple examples of the same graphical symbol carrying different meanings have been noted in our prior work with DHS mapmakers (Robinson , Roth, and MacEachren, 2011), and we remain in need of better strategies for highlighting such differences in search results in the Symbol Store to make users aware of different interpretations. Card-sorting and other methods for developing and evaluating categories associated with symbols (Roth et al. 2011) need to be adapted for individuals to use in web-based tools like the Symbol Store.

While supporting map interoperability involves challenges that extend well beyond the common and consistent representation of features using symbols, we believe our work to design and develop the Symbol Store contributes a novel web-based approach that has the potential to significantly help cartographers discover, retrieve, and share symbols beyond the means afforded by current GIS software and informal personal symbol collections.

Acknowledgments

This work is supported by a contract from the US Department of Homeland Security Science and Technology Directorate, Command, Control and Interoperability Division. The views and opinions expressed here are of the authors, and do not reflect the official positions of the Department of Homeland Security or the Federal Government.

References

Akella, M. K. 2009. "First Responders and Crisis Map Symbols: Clarifying Communication." *Cartography and Geographic Information Science* 36 (1): 19–28.

ANSI. 2006. *ANSI INCITS-415 2006 Homeland Security Mapping Standard – Point Symbology for Emergency Management*. Washington, DC: American National Standard for Information Technology.

Buttenfield, B. 1999. "Usability Evaluation of Digital Libraries." *Science & Technology Libraries* 17 (3): 39–59.

Cutter, S. L. 2003. "GI Science, Disasters, and Emergency Management". *Transactions in GIS* 7 (4): 439–446.

Department of Defense, USA. 2008. *Common Warfighting Symbology: MIL-STD-2525C*. Arlington, TX: Department of Defense.

Dymon, U. J. 2003. "An Analysis of Emergency Map Symbology." *International Journal of Emergency Management* 1 (3): 227–237.

Dymon, U. J., and E. K. Mbobi. 2005. "Preparing an ANSI Standard for Emergency and Hazard Mapping Symbology." In *International Cartographic Conference*, edited by Rodolfo Núñez de las Cuevas, July 9–16, A Coruña, Spain.

Harrower, M. and C. A. Brewer. 2003. "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps." *Cartographic Journal* 40: 27–37.

Kostelnick, J. C., J. E. Dobson, S. L. Egbert, and M. D. Dunbar. (2008). "Cartographic symbols for humanitarian demining." *The Cartographic Journal* 45 (1): 18–31.

Miller, G. A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41.

Robinson, A. C., R. E. Roth, J. Blanford, S. Pezanowski, and A. M. MacEachren. 2012. "Developing Map Symbol Standards through an Iterative Collaboration Process." *Environment and Planning B: Planning and Design* 39 (6): 1034–1048.

Robinson, A. C., R. E. Roth, and A. M. MacEachren. 2010. "Challenges for Map Symbol Standardization in Crisis Management." In *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, edited by S. French, B. Tomaszewski, and C. Zobel, May 2–5, Seattle, WA.

Robinson, A. C., R. E. Roth, and A. M. MacEachren. 2011. "Understanding User Needs for Map Symbol Standards in Emergency Management." *Journal of Homeland Security and Emergency Management* 8 (1): 1–16.

Robinson, A. H. 1973. "An International Standard Symbolism for Thematic Maps: Approaches and Problems." *International Yearbook of Cartography* 13: 19–26.

Roth, R. E., B. G. Finch, J. I. Blanford, A. Klippel, A. C. Robinson, and A. M. MacEachren. 2011. "Card Sorting for Cartographic Research and Practice." *Cartography and Geographic Information Science* 38 (2): 89–99.

Saligoe-Simmel, J. 2010. *Ortelius' Community-Based Symbol Collection*. St. Petersburg, FL: North American Cartographic Information Society.

Sheesley, B. 2007. *TypeBrewer: Design and Evaluation of a Help Tool for Selecting Map Typography*. Madison, WI: Department of Geography, University of Wisconsin-Madison.

Sondheim, M., D. Charmley, and G. Leeming. 2010. *Emergency Management Symbology*, Version 1.0, 71. Victoria, BC: Refractions Research.

Tweeting and Tornadoes

Justine I. Blanford
GeoVISTA Center
Pennsylvania State University
jib18@psu.edu

Jase Bernhardt
Geography Dept
Pennsylvania State University
jeb5249@psu.edu

Alexander Savelyev
GeoVISTA Center
Pennsylvania State University
azs5362@psu.edu

Gabrielle Wong-Parodi
Dept of Engineering and Public Policy
Carnegie-Mellon University
gwparodi@gmail.com

Andrew M. Carleton
Geography Dept
Pennsylvania State University
amc7@psu.edu

David W. Titley
Center for Solutions to Weather and Climate Risk, Pennsylvania State University
dwt12@psu.edu

Alan M. MacEachren
GeoVISTA Center
Pennsylvania State University
maceachren@psu.edu

ABSTRACT

Social Media and micro-blogging is being used during crisis events to provide live up-to-date information as events evolve (before, during and after). Messages are posted by citizens or public officials. To understand the effectiveness of these messages, we examined the content of geo-located Twitter messages ("tweets") sent during the Moore, Oklahoma tornado of May 20th, 2013 (+/-1day) to explore the spatial and temporal relationships of real-time reactions of the general public. We found a clear transition of topics during each stage of the tornado event. Twitter was useful for posting and retrieving updates, reconstructing the sequence of events as well as capturing people's reactions leading up to, during and after the tornado. A long-term goal for the research reported here is to provide insights to forecasters and emergency response personnel concerning the impact of warnings and other advisory messages.

Keywords

Twitter, Tornado, Situational Awareness, Emergency Response, Message Warnings and Alerts, Risk Communication, Spatial and Temporal Visualization, GIS

INTRODUCTION

"*The best weather and water forecast can only save lives if it is communicated effectively to at-risk residents and the public officials who are charged with protecting them.*" Pg 37 (Sullivan and Uccellini, 2013). Communicating imminent risk from a severe weather event has the potential to reduce loss of life. Sullivan and Uccellini (2013), in interviews with media, emergency managers and staff at the National Oceanic and Atmospheric Administration (NOAA) and the U.S. National Weather Service (NWS), found that social media is playing an important role in communicating threats from weather events. For example, during Hurricane/Superstorm Sandy the Facebook pages and Twitter feeds of local government offices picked up several thousand followers who were looking for up-to-date information (Sullivan and Uccellini, 2013). In 2011, the NWS embarked on the 'Weather Ready Nation' initiative, to better fulfill its mission of protecting lives and livelihoods. Although much emphasis is being placed on technical capabilities that would enable the NWS to increase its warnings timeliness and accuracy, there is also a component of Weather Ready Nation that emphasizes the need to improve its effectiveness in communicating warnings and providing up-to-date information on high-impact weather events (NWS, 2013).

The power to communicate event information through social media has clearly been demonstrated (e.g., (Starbird and Palen, 2010b; Starbird and Palen, 2010a; Roche, Propeck -Zimmermann and Mericskay, 2011; Bruns, Burgess, Crawford and Shaw, 2012; Sullivan and Uccellini, 2013)). Social media has been useful in the circulation and dissemination of news (Kwak, 2010), as well as in providing live up-to-date information on a variety of environmental hazard events through citizen reporting (e.g., earthquakes in the USA (Crooks, Croitoru, Stefanidis and Radzikowski, 2013) and Haiti (Roche et al., 2011); flooding on the Red River in the USA and Canada (Starbird, Palen, Hughes and Vieweg, 2010; Vieweg, Hughes, Starbird and Palen, 2010); and wildfires in USA (Vieweg et al., 2010)). With the increasing ability to utilize geographic information, either through mining the content of a message or via its geographic properties, it is now much easier to analyze information spatially. For example, SensePlace2 (MacEachren, Jaiswal, Robinson, Pezanowski, Savelyev, Mitra, Zhangi and Blanford, 2011) can integrate geospatial, temporal, and attribute dimensions of Twitter, providing inputs to situational awareness and an understanding of reactions to events (e.g., (Robinson, Savelyev, Pezanowski and MacEachren, 2013)).

For the public to take action, warnings must be understood by the recipient before they can be acted upon. Several factors may affect whether people fully process and understand the information contained in a warning including past experience with natural hazard events, general awareness, and belief and trust (see (Brotzge and Donner, 2013) and references within for details). To be successful, warnings must communicate the necessary information clearly in a timely manner to allow users to react (Brotzge and Donner, 2013). An important question that has had limited attention thus far is how to leverage social media as a lens through which to analyze citizen reactions to natural disasters associated official advisories and warnings. As a step toward addressing this question, we used Twitter to assess the effectiveness of warning messages sent during the Moore, Oklahoma tornado of May 20th, 2013 (+/-1day) by exploring the spatial and temporal relationship of real-time reactions of the general public as the storm system developed into a tornado.

Oklahoma Tornado: During late May, 2013, a series of tornado-producing storm systems swept through the greater Oklahoma City, OK region. A particularly violent tornado touched down at 14:56 hr (all times Central Daylight Time) near Newcastle, 16 minutes after a tornado warning was issued by the NWS Norman, Oklahoma office. The tornado rapidly strengthened, tracking directly over the city of Moore. The tornado travelled a total of 22.5 km in 39 minutes, with a maximum path width of 1.7 km (NOAA, 2013a), and attained a rating of EF-5 on the Enhanced Fujita Tornado Damage Scale, the maximum rating possible. Due to its strength, longevity, and passage over a populated area, the tornado claimed 24 lives and caused extensive damage (Kuligowski, Phan, Levitan and Jorgensen, 2013).

We analyzed 86,100 geo-located tweets collected between May 19th (the day prior to the tornado) and May 21st (the day after the tornado) 2013, for Oklahoma. Tweets were captured using the Twitter Streaming API version 1.1 (https://dev.twitter.com/docs/streaming-apis/streams/public) and saved to a text file in JSON format using a node.js application (http://nodejs.org/).

Tweets containing tornado-relevant information were identified by querying for keywords that included: 'tornado' (+watch, +warning), 'storm', 'weather', 'take/ing cover', 'shelter', 'pray', 'emergency', 'red cross', 'help' and the root of 'devast' (to include devastated and devastation), 'destruct' (to include destructed and destruction), and 'donat' (to include donation(s) and donate). A set of keywords were selected through an iterative process to capture relevant tweets related to the tornado event. We started with words related to tornadoes and then included additional keywords to capture tweets about the tornado before and after the event.

Figure 1 depicts the spatial temporal distribution of tweets with any of the specified keywords. People were clearly interested in the weather, storms and tornadoes. Not surprisingly, on the day of the tornado the number of tweets containing the word tornado increased. During and after the storm the number of tweets including prayers increased as did those relating to destruction, devastation and requests for help and donations.

We next explored the spatial and temporal distribution of tweets on May 20th in more detail using ArcMap 10.2. Figure 1 shows the spatial distribution of tweets by keyword prior to the tornado touching down (i.e. tweets captured before 14hr), during the tornado (14hr – 16hr) and post-tornado (after 16hr). Pre-tornado tweets are primarily composed of those containing keywords that highlight impending threat, such as storm, tornado, and watches / warnings for severe weather and tornadoes. During the tornado itself, the keywords are focused increasingly on the dynamic response to the event, namely deployment of the emergency alert system (sirens), and the resultant precautions taken by individuals in response (shelters). Finally, after the tornado dissipated, the main keywords highlight the destruction and damage that occurred. Figure 2 shows that the frequency of tweets also increases, peaking during the tornado touchdown. After the tornado touched down, the number of tweets containing the word tornado decreased, but much more slowly than it peaked.

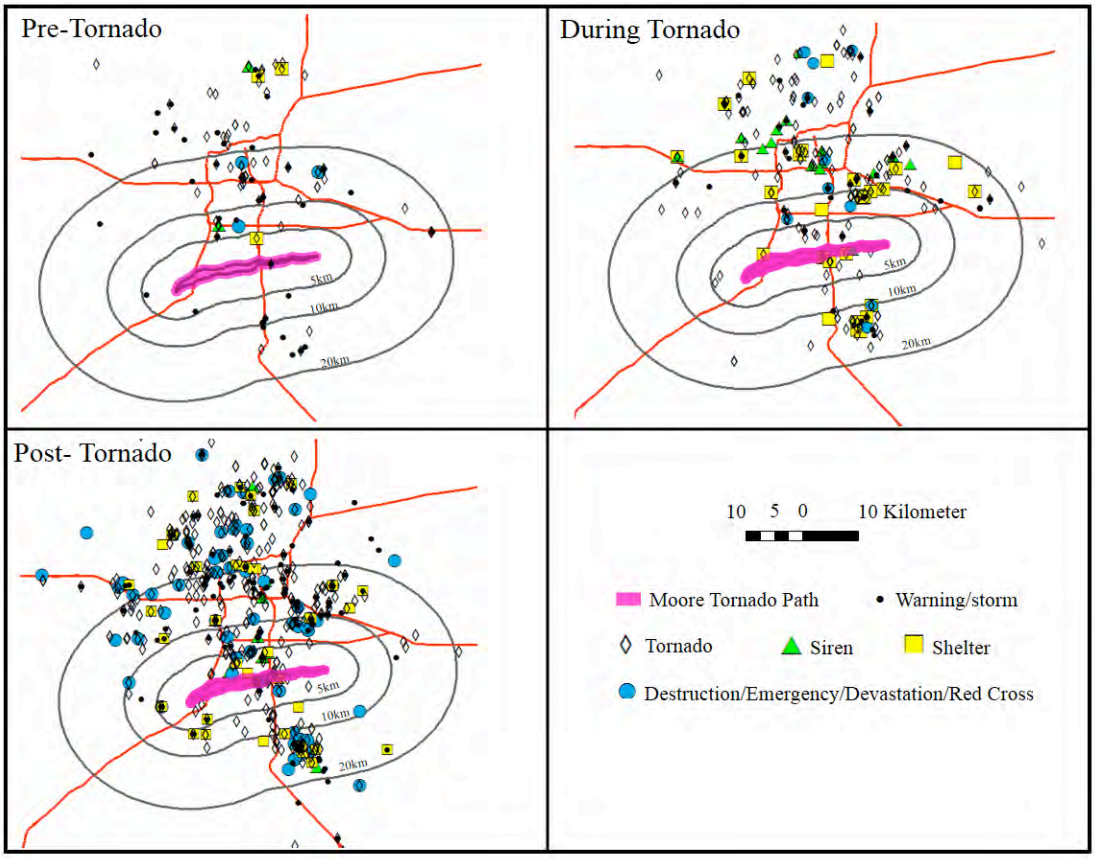
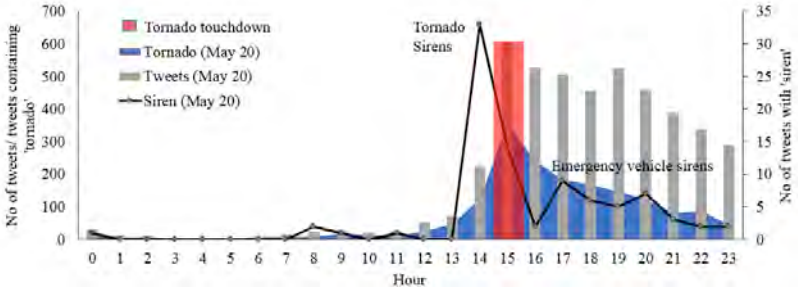


Figure 1: Maps of distribution of tweets with keywords pre, during and post tornado touchdown, May 20th. Tweets within 5, 10 and 20km of the tornado path fall within the concentric buffer zones (grey lines). Tweets outside of this area are greater than 20km from the tornado path.

Between 5:00hr and 15:37hr on May 20th, the NWS office in Norman, OK issued a series of severe weather watches and warnings. These statements were officially disseminated by the National Weather Service through various media, including Twitter (see (NOAA, 2013b)). A total of 36 tweets were posted by NWS, resulting in 2,583 retweets) (see Figure 3). Similar to the tweets illustrated in Figure 2, tweets about the tornado remained low through the day, dramatically increasing in the hour preceding the Moore tornado. Interestingly, of all the retweeted messages, the message requesting users to retweet had the highest retweet (19%). Fifty-eight percent of the retweets occurred during the tornado warning, 33% of these were for messages telling people to take action such as 'take cover'; 18% were for updates on the tornado's movement and location, while the remaining messages highlighted tornado preparedness and what not to do (e.g., be in your car).

Figure 2: Temporal distribution of tweets with at least one keyword (gray bars) (e.g., 'storm', 'weather', 'take/ing cover', 'shelter', 'pray', 'emergency', 'red cross', 'help' and 'devast', 'destruct', and 'donat' in relation to tweets containing the keyword 'tornado' (blue area) or 'siren' (black line) and how these related to the tornado event (red bar). Tweets were summarized for each hour.



Sirens in Moore were sounded six times with the initial siren occurring shortly after the first NWS tornado warning was sent (14:41hr) with the final warning at 15:20hr (Kuligowski et al., 2013). The first mention of sirens also begins at 14:41hr (N=22 tweets in 4 minutes) with tweets such as "*Sirens going off now!! Take cover...be safe!"*, *"Sirens sirens sirens. Becoming so real"*, and *"If u hear a tornado siren, uve got 6-8 minutes…"*

An initial tornado warning was issued for the Moore tornado at 14:40hr, valid through to 15:15hr, based on the projected tornado track. A total of 104 tweets were sent from inside the warning zone. Within 3-7 minutes of the tornado warning being issued, 4 tweets mentioned tornado sirens; 3 of which referenced the sounding of a tornado warning and 27 contained the word 'tornado'. The content of the tweets varied with users providing technical details (re-tweets of storm spotter reports) to the brief ("*Dang this tornado is huge*") and more descriptive "*Saw two, looks like more tornados forming. Taking shelter now."* It was clear that people continued tweeting to provide situation updates, but also expressing their frustrations and sentiments (e.g., *"Why are there no public underground tornado shelters in every OK town?! THIS IS TORNADO ALLEY! @MaryFallin @OKSENATEINFO @OKHouseofReps.")*.

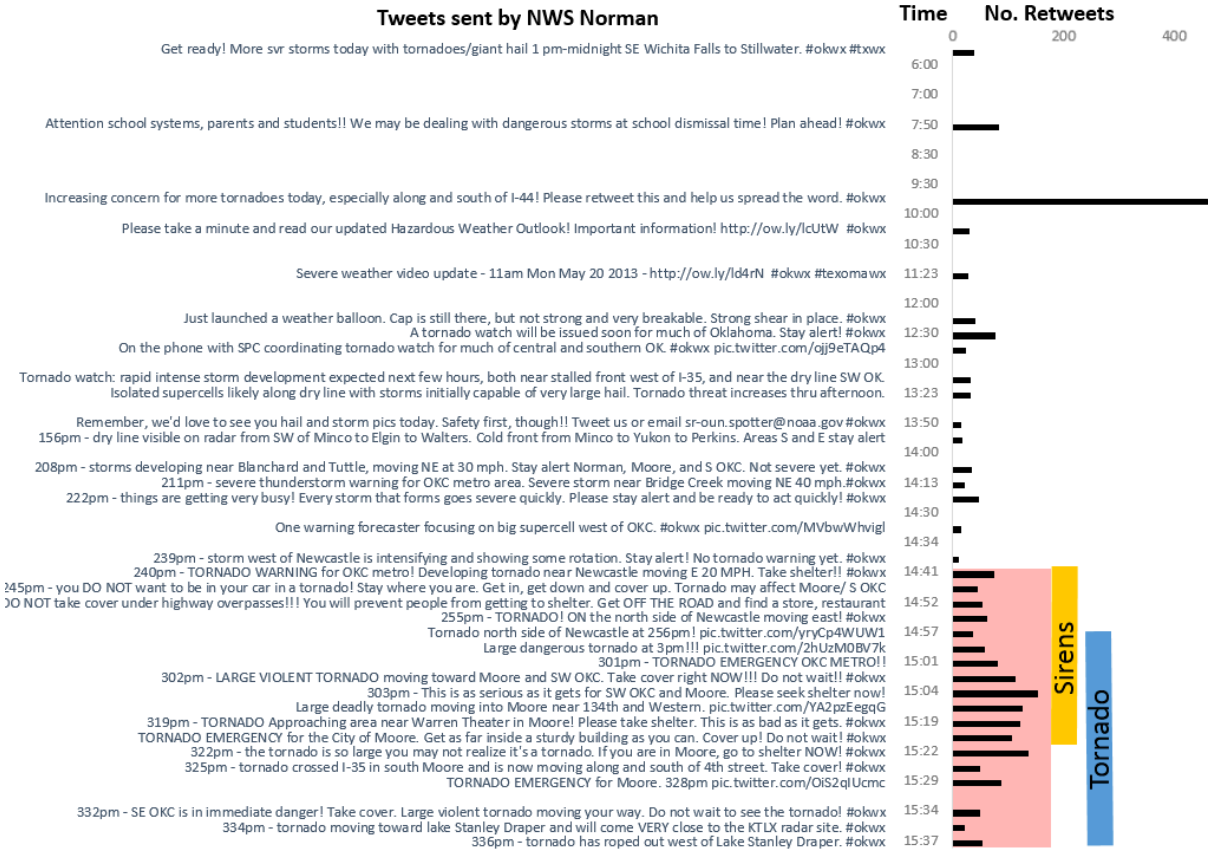


Figure 3: Timeline of warning messages sent by NWS Norman on May 20th with the number of retweets for each message (black bars), sirens (dark yellow), time of tornado touchdown (blue) and the time and duration of the tornado warning (red area).

Content analysis of the tweets found that tweets were used to provide situational updates such as relaying media reports, which included television stations KOCO, KJRH, and KFOR, the University of Oklahoma emergency alert system, and re-tweets of information from the NWS (*e.g.,"KOCO is talking about the south, guys. Wall cloud. No tornado yet."*); providing real-time weather observations (*e.g., "This tornado is about a mile wide. Oh dang."*); providing the location of shelters (*e.g., "If you're on campus, seek shelter in Residence Hall basements, Union basement or Huff. #OU #okwx"*); and communicating personal safety and locating family and friends *(e.g., "We are ok. F4 tornado hit about 2 miles from us. Don't have power right now. Hope this posts."*). After the tornado passed, messages included situational awareness and damage reports (e.g., *"Heavy tornado damage near SE 4th and Bryant. Homes are gone"* and *"Children trapped in #Moore guess ill be #volunteering all night #oklahomatornadoes #oklahoma #okiepride #help #tornado #redcross."*).

CONCLUSION

Twitter is an effective messaging system that enables information to be received and posted in a timely manner. During the Moore, Oklahoma tornado, Twitter was useful for providing updates and relaying of information. By analyzing the text of each tweet and using a list of keywords, we gained insights into what happened on the ground and understood people's interest and reactions, both spatially and temporally. For unpredictable and destructive/hazardous weather events, such as the tornado analyzed here, the time between issuing a warning

and the tornado touchdown can be short, emphasizing the need for clear communication. Including a request to retweet may help facilitate the wider dissemination of critical information via Twitter. The study presented here highlights the need for additional research that should include strategies for prompting retweets, identify messaging that works and does not work and assess the role of volunteer communities in communicating risk obtained from a variety of both formal (e.g., NWS) and informal (e.g., TV stations, social networks, personal observations) sources of information. In this study we analyzed data for a single event. A long-term goal for the research reported here is to provide insights to forecasters and emergency response personnel concerning the impact of warnings and other advisory messages. To broaden the applicability for this kind of data, comparing events of different sizes and duration may provide a deeper understanding of people's responses during catastrophic weather events (Bagrow, Wang and Barabasi, 2011). Future work could also build on previous studies, such as Mendoza, Poblete and Castillo (2010) and analyze warning messages to help maximize the likelihood that the public will take the most appropriate action during an event. In addition, develop a lexicon of the most appropriate words and phrases to use to query the public's response will enable the quick retrieval of relevant tweets before, during and after an event.

REFERENCES

1. Bagrow, J. P., Wang, D. S. & Barabasi, A. L. (2011) Collective response of human populations to large-scale emergencies. *PLoS One,* **6,**
2. Brotzge, J. & Donner, W. (2013) The tornado warning process. A review of current research, challenges, and opportunities. *Bulletin of the American Meteorological Society,* **94,** 1715-1733
3. Bruns, A., Burgess, J., Crawford, K. & Shaw, F. (2012) #qldfloods and @qpsmedia: Crisis communication on Twitter in the 2011 south east Queensland floods pp. 58. ARC Centre of Excellence for Creative Industries and Innovation, Queensland, Australia.
4. Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. (2013) #earthquake: Twitter as a distributed sensor system. *Transactions in GIS,* **17,** 124-147
5. Kuligowski, E. D., Phan, L. T., Levitan, M. L. & Jorgensen, D. P. (2013) NIST special publication 1164: Preliminary reconnaissance of the May 20, 2013, Newcastle-Moore tornado in Oklahoma. pp. 81. National Institute of Standards and Technology Special Publication
6. Kwak, H. (2010) What is Twitter, a social network or a news media? In: *The International World Wide Web Conference*, pp. 591-600. Raleigh, North Carolina.
7. MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhangi, X. & Blanford, J. (2011) Senseplace2: Geotwitter analytics support for situational awareness In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 181-190. Providence, RI
8. Mendoza, M., Poblete, B. & Castillo, C. (2010) Twitter under crisis: Can we trust what we RT? In: *1st workshop on Social Media Analytics (SOMA '10)*, pp. 9. Washington, DC.
9. NOAA (2013a) The tornado outbreak of May 20, 2013. pp. National Weather Service Weather Forecast Office, Norman, OK.
10. NOAA, N. (2013b) Information services chronology for the tornado outbreak of May 20, 2013. pp. NWS NOAA.
11. NWS (2013) Weather-ready nation roadmap v2.0. pp. 81. The US Weather Service, NOAA.
12. Robinson, A. C., Savelyev, A., Pezanowski, S. & MacEachren, A. M. (2013) Understanding the utility of geospatial information in social media. In: *10th International ISCRAM Conference* (Eds, Comes, T., Fiedrich, F., Fortier, S., Geldermann, J. and Muller, T.), pp. 5. ISCRAM, Baden-Baden, Germany.
13. Roche, S., Propeck -Zimmermann, E. & Mericskay, B. (2011) Geoweb and crisis management: Issues and perspectives of volunteered geographic information. *GeoJournal,* **78,** 21-40
14. Starbird, K. & Palen, L. (2010a) Tweak the tweet: Leveraging microblogging proliferation with a prescriptive grammar to support citizen reporting. In: *7th International ISCRAM Conference (Short Paper)*, pp., Seattle, WA.
15. Starbird, K. & Palen, L. (2010b) Pass it on?: Retweeting in mass emergency. In: *7th International ISCRAM Conference*, pp., Seattle, WA.
16. Starbird, K., Palen, L., Hughes, A. L. & Vieweg, S. (2010) Chatter on The Red: What hazards threat reveals about the social life of microblogged information. *2010 ACM Conference on Computer Supported Cooperative Work,* 241-250
17. Sullivan, K. D. & Uccellini, L. W. (2013) Service assessment: Hurricane/post -tropical cyclone Sandy, October 22 – 29, 2012. pp. 66. U.S. Department of Commerce NOAA and NWS, Silver Spring, Maryland.
18. Vieweg, S., Hughes, A. L., Starbird, K. & Palen, L. (2010) Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. *CHI 2010: Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems,* **1-4,** 1079-1088

HAZARDOUS MATERIAL SIGN DETECTION AND RECOGNITION

Albert Parra, Bin Zhao, Andrew Haddad, Mireille Boutin, Edward J. Delp

Video and Image Processing Lab (VIPER)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

In this paper we describe two methods for hazardous material (hazmat) sign recognition. The first method is based on segment detection and grouping using geometric constraints. The second method is based on the use of a saliency map and convex quadrilateral detection. Our experimental results show a detection accuracy of 57.7% on a set of hazmat signs taken in the field under various lightning conditions, distances, and perspectives.

Index Terms— Sign detection, shape detection, saliency map, Hough Transform.

1. INTRODUCTION

Hazardous materials can react differently to environmental stimuli and cause problems in accidents and emergency situations and therefore makes these materials particularly dangerous to civilians and first responders. A federal law in the US requires vehicles transporting hazardous materials be marked with a standard sign (i.e., a "hazmat sign") identifying the type of material the vehicles is carrying [1]. These signs have identifying information described by the sign shape, color, symbols, and numbers. In this paper we describe two methods for hazmat sign detection and recognition. Each method detects (segments) the sign using shape information and then color information is used for sign identification. We first describe a sign recognition method based on segment detection and grouping using geometric constraints. Although this method is fast, it has several disadvantages. For example, low resolution images can cause missed straight edges at $\pm 45^{\circ}$. In order to overcome these issues, we describe a second method that replaces the initial edge detection with a saliency map and combines contour detection and the Hough Transform. The second method is robust to rotation, perspective distortion, sign distance from the camera, distance between multiple signs, and blurred and low resolution images.

2. REVIEW OF EXISTING METHODS

Sign detection can be classified into three main categories: shape-based [2], color-based [3] and saliency-based [4]. Shape-based approaches first generate an edge map and then use shape information to find objects. Color-based approaches overcome the problems of shape variation, partial occlusion, and perspective distortion. However, colors are sensitive to lightning conditions and illumination changes. Saliency-based approaches utilize selective visual attention models. A saliency-based visual attention (SBVA) model was presented in [4] using images features with a Gaussian pyramid. A graph-based visual saliency (GBVS) method was proposed in [5], to highlight conspicuous regions. A histogram-based contract (HC) method and a region-based contract (RC) method were introduced in [6] to construct saliency maps. HC-maps produce better performance over RC-maps but at the expense of increasing the computation time. A saliency map generation method was described in [7] using image signature (IS) to highlight sparse salient regions based on RGB or Lab color spaces.

Sign recognition methods can be classified into: geometric constraint methods, boosted cascades of features, and statistical moments[8, 9, 10]. Methods based on geometric constraints include the use of Hough-like methods [11, 12], contour fitting [13, 14], or radial symmetry detectors [15, 16]. Methods based on the boosted cascades of features commonly use the Viola-Jones framework [17, 18, 19]. These approaches often use object detectors with Haar-like wavelets of different shapes, and produce better results when the feature set is large. Methods based on statistical moments [20, 21, 22] use the central moments of the projections of the object to be detected. These methods are not robust to projective distortions or non-uniform lightning conditions.

3. HAZMAT SIGN DETECTION AND RECOGNITION

We developed two methods for hazmat sign detection and recognition. In the first method we detect and group segments using geometric constraints. In the second method we used saliency map to localize regions that potentially contain hazmat signs and then find the sign in these regions by checking for convex quadrilaterals. Both methods use color information for sign identification.

Segment Detection Using Geometric Constraints: We find edges in the image using the Canny edge detector. Since hazmat signs can be present at various distances, we use median auto-thresholding. To deal with non-uniform illumination changes in the scene, we also grayscale histogram equalize the image. We assume: 1) any sign in the image has to be approximately upright with its major axes aligned with the XY axis (Figure 2 illustrates the difference between an upright sign and a distorted sign); and 2) the projective distortion has to be small. (i.e., edges have to be approximately at $\pm 90^{\circ}$ with respect to each other). Given these assumptions, we use morphological filters to eliminate edges not belonging to a hazmat sign. We create structuring elements at $\pm 45^{\circ}$ and use them separately to erode the Canny edge map. The resulting edge map is the superposition of the two erosions. Signs that do not satisfy the two underlying assumptions will not preserve edges in

This work was partially supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and by NSF grant CCF-0728929. Address all correspondence to Edward J. Delp (ace@ecn.purdue.edu).

the resulting edge map. Once the edge map has been filtered, we find line segments using the probabilistic Hough Transform [23]. We set the minimum gap allowed between points on the same line to 5 pixels and the maximum gap to 5% of the maximum dimension of the original image (width or height). We next proceed to group the segments into candidates. Each candidate consists of a set of segments having one reference segment, at least one parallel segment, and two orthogonal segments (one to the left and one to the right of the reference segment). The reference segment is chosen at random from the list of segments that have not been grouped yet. Parallel segments need to have similar slope and length relative to the reference segment. The thresholds are set so that $|m_p - m_r| < 0.1$ and $|l_p - l_r| < 0.75e$, where m_p and m_r are the slopes of the parallel and reference segments respectively, l_p and l_r are the lengths of the parallel and reference segments respectively, and $e = max(l_p, l_r)$. The distance d between the reference and the parallel segments has to be in the range $0.5e < d < 2.5e$. This distance is defined between the middle points of the parallel and the reference segments. Also, the angle between the reference and the parallel segments has to be less than 20°. This angle is defined by the normal of the parallel segment at its middle point and the vector joining the middle points of the parallel and the reference segments. Orthogonal segments need to have opposite slope and similar length to the reference segment, that is, $|m_p + 1/m_r| < 0.1$ and $|l_p - l_r| < 0.75e$. The distance d between the reference and the orthogonal segments has to be in the range $0.5e < d < 2.5e$. The angle between the reference and the orthogonal segments is defined as positive when the orthogonal segment is to the right of the reference segment, and defined as negative when the orthogonal segment is to the left of the reference segment. For each candidate set satisfying the geometric constraints we compute its minimal bounding box. We then discard any candidate with a bounding box aspect ratio smaller than 1.3. Finally, we check the remaining candidates and remove those that correspond to the same sign. This can be done by first dividing all bounding boxes that overlap more than 50% into groups, and then finding the optimal bounding box for each group. We consider the optimal bounding box to be the one with its nodes closest to its centroid (i.e., closest to a square). Figure 1 illustrates an example of the complete process. Once a hazmat sign is segmented, its color is set to the average hue inside the optimal bounding box and the color is used to identify the sign.

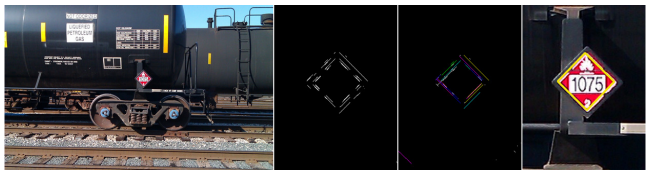


Fig. 1: First method (left to right): original image, segments at $\pm 45^\circ$, grouped segments, optimal bounding box.

Convex Quadrilateral Detection Based on Saliency Map: Our first method described above has some drawbacks. With low resolution images, the resulting edge map will not contain straight edges at $\pm 45^\circ$ and the erosion process will then delete most of them. The same happens with blurry images. Hazmat signs not satisfying the two assumptions of the first method will be removed during the erosion process, as shown in Figure 2. The gap threshold of the probabilistic Hough Transform may cause the segment grouping process to merge two segments from two close signs, as shown in Figure 3. Our second technique replaces the initial edge detection with a saliency map to detect regions potentially containing hazmat signs. The saliency map assigns higher saliency to more visually attractive regions. We used both the Lab and RGB color spaces as a combination of the IS-Lab and the IS-RGB methods in [7] to generate our saliency map. This method, which we call IS-RGB+Lab, gave us the best results in the experiments. We threshold this map to extract the most salient regions in the image. For each salient region found, we detect signs using specific color channels. This allows us to do both sign detection and color recognition at the same time, since we will assume that the color of any hazmat sign found in the region will correspond to the color channel associated to it. Currently the channels used are red, green, blue, and we also use the grayscale version to account for white or black signs. Note that the possible colors for hazmat signs also include yellow and orange, but these can be obtained by transforming the image from RGB to a hue-based color space and then segment the hue image. The grayscale and the color channels are then thresholded to account for highly chromatic areas using an empirically determined threshold. Each of the thresholded images is binarized, and morphologically opened to remove small objects containing less than 0.05% of the total number of pixels. We also use dilation to merge areas that may belong to the same object. We then retrieve the contours from the resulting binary image [24]. For each contour, we use the standard Hough Transform [25] to find straight lines that approximate the contour as a polygon. The intersections of these lines give us the corners of the polygon, which can be used to discard non-quadrilateral shapes. If the contour is approximated by four vertices, we find its convex hull [26]. If the convex hull still has four vertices, we check the angles formed by the intersection of its points. If each of these angles is in the range $90^\circ \pm 1.5^\circ$, and the ratio of the sides formed by the convex hull is in the range 1 ± 0.5, we can assume that we have found a convex quadrilateral. Finally, we use the same technique as in the first method to remove quadrilaterals that correspond to the same hazmat sign. Figure 4 illustrates a successful detection of two signs, one is affected by rotation and perspective distortion. Figure 5 illustrates a successful detection of one sign and also a false positive. In this particular case the issue could be addressed by using an optical character recognition to detect the text inside the sign candidate.

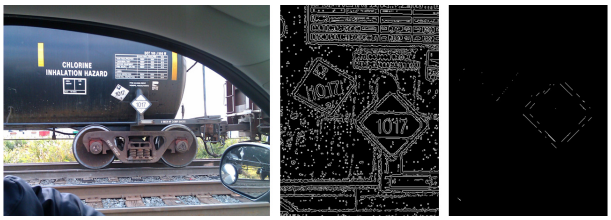


Fig. 2: Issue with first method: sign distortion.

Our second method offers multiple advantages. First, it is robust to rotation, since there is no erosion at $\pm 45^\circ$. Second, it is robust to perspective distortion, since convex quadrilaterals can be skewed. Third, it is able to detect signs close to each other, since there is no overlapping of line segments caused by the probabilistic Hough Transform. Fourth, it is more robust to blurred and low resolution images, since there is no edge detection is performed on the sign recognition step. Lastly, it is more robust to color recognition, since it detects signs already in specific color channels. The only disadvantage is its execution time.

Fig. 3: Issue with first method: segment merging.

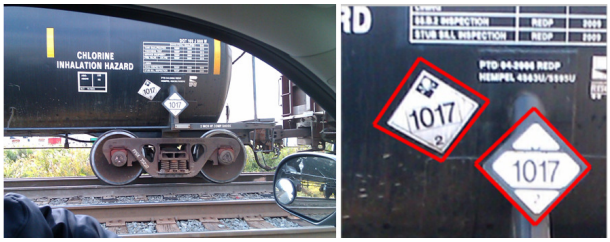


Fig. 4: Second method: true positives.

4. EXPERIMENTAL RESULTS

The first experiment consisted of images from a dataset and manually comparing the results with ground-truth information. The second experiment consisted of evaluating the saliency map methods. The tests were executed on a desktop computer with a 2.8GHz CPU and 2GB RAM. The ground-truth information included the sign distance from the camera, sign color, projective distortion of the sign, image resolution, possible shadow affecting the sign, and sign location on the image. Note that we only used the color and not the text of the sign for sign identification for these experiments. The image dataset consisted of 40 images each containing one or more hazmat signs (52 hazmat signs in total). The images were taken using three different cameras: a 8.2 Mpx Kodak Easyshare C813, a 16 Mpx Nikon Coolpix S800c, and a 5 Mpx camera on an HTC Wildfire mobile telephone. The images were acquired in the field, under various lightning conditions, distances, and perspectives. Among the 40 images: 17 were taken at 10-50 feet, 17 at 50-100 feet, and 6 at 100-200 feet; 3 had motion blur, 8 had geometrical distortion (i.e., perspective or rotation), 7 had shaded signs, and 6 had low resolution. Figure 6 illustrates some samples from the image dataset.

Fig. 5: Second method: True positive/False positive.

Fig. 6: Samples from the image dataset. From left to right: low resolution, blurred sign, shaded sign, geometrical distortion.

Table 1: Image Analysis Results.

Method	Sign	Accuracy	Color	Accuracy	Total
1	16	30.7%	6	11.5%	52
2	30	57.7%	22	42.3%	52

Table 1 shows the results of the first experiment using the two methods proposed in this paper. We determined how many signs were successfully detected (*Sign*) and how many were successfully identified (i.e., sign detected plus correct color (*Color*)). Note that the sign color recognition was done only if a sign was detected. Among the successfully detected signs we had a higher accuracy for color recognition. The first method recognized the correct color in 37.5% of the successfully detected signs, while the second method recognized the correct color in 73.3% of the successfully detected signs. The low accuracy of our first method is caused by multiple factors, including segment merging, edge detection failure on low resolution images, distortion and rotation of the sign, and multi-colored signs. However, note that multi-colored signs may also cause our second method to miss the detection, given that we detect signs at individual color channels. The first method had an average execution time of 2.3 seconds in total. The accuracy of the second method is influenced by the saliency map thresholding, the color recognition method based in specific color channels, and the morphological operations and Hough Transform on low resolution images. The second method had an average execution time of 5.1 seconds in total. Although the first method is faster, the second method doubles the sign detection accuracy, while still being fast enough to be used in real time applications.

Table 2 shows the results of the second experiment, including the average execution time and the score of the saliency map. The saliency map methods evaluated in the second experiment are: SBVA from [4], GBVS from [5], and IS-RGB and IS-Lab from [7]. Figures 7 and 8 illustrate examples of each method. Note that this process is only done as the first step of our second method. It can be

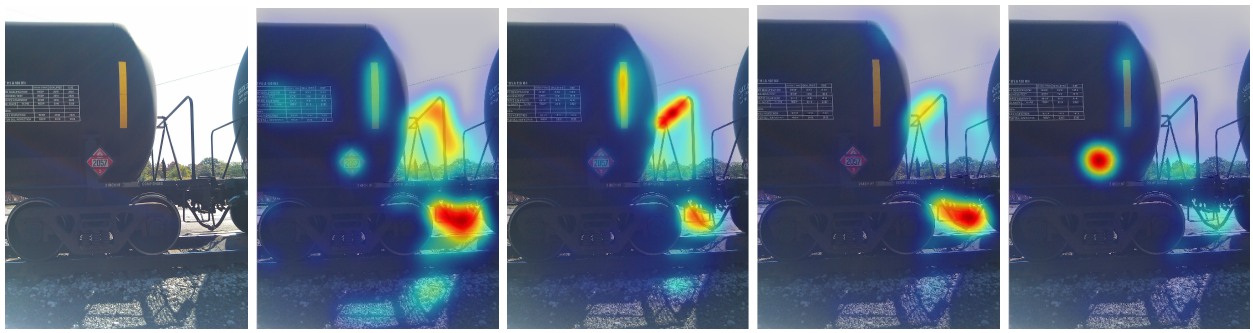


Fig. 7: Examples of saliency maps. From left to right: SBVA, GBVS, IS-RGB, IS-Lab.

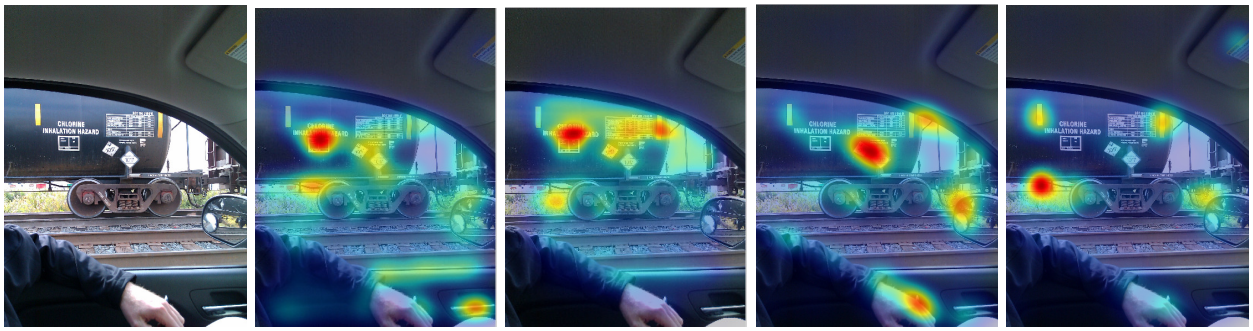


Fig. 8: Examples of saliency maps. From left to right: SBVA, GBVS, IS-RGB, IS-Lab.

Table 2: Average Execution Time, Distribution and Score of Each Saliency Map Method.

Method	Time	Good	Fair	Bad	Lost	Score
SBVA	1.92s	23	15	11	3	110
GBVS	3.10s	19	16	11	6	100
IS-RGB	0.40s	36	5	4	7	122
IS-Lab	0.41s	18	3	22	9	82
IS-RGB+Lab	0.81s	42	3	4	3	136

seen in Figure 8 that although the IS-RGB saliency map covers two close signs in only one region, we can detect them separately (see Figure 4). We manually classified the saliency map results into four categories: good, fair, bad, and lost. Figure 9 illustrates examples of each case. For each of the 52 signs we assigned 3 points to a good map (sign was mostly contained in a high saliency-valued region), 2 points to a fair map (sign was mostly contained in a middle saliency-valued region), 1 point to a bad map (sign was mostly contained in a low saliency-valued region), and 0 point to a lost map (sign was not contained in any saliency-valued region). The score of each saliency map method is the sum of the points assigned to each image in the dataset, which ranges from 0 to 156 for each saliency map method. The IS method using the LAB or RGB color space has higher score and executes faster than the SBVA and the GBVS methods. The IS-RGB+Lab method based on two color spaces had the highest score, at the cost of increasing the execution time, with respect to the two IS methods using a single color space. However, the IS-RGB+Lab method still runs 2.37 times faster than the SBVA method and 3.83 times faster than the GBVS method.

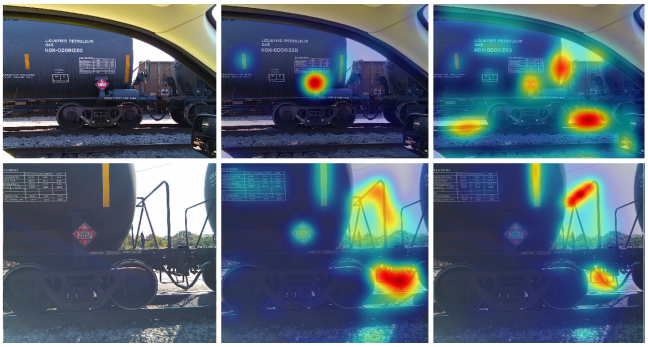


Fig. 9: Saliency map categories (top to bottom, left to right): original image, good, fair; original image, bad, lost.

5. CONCLUSIONS AND FUTURE WORK

We described two methods for hazmat sign detection and recognition. The experimental results showed that our second method is more accurate for both sign detection and color recognition. We can further increase the accuracy by color calibrating the images to enhance the color recognition, developing a blur correction method to reduce the impact of blurred images, and developing an optical character recognition method to interpret the text inside the hazmat signs.

6. REFERENCES

[1] United States Department of Transportation, *Code of Federal Regulations, Title 49, DOT Hazmat*, Labelmaster, 2012 edition, October 2012.

[2] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE Transactions on Image Processing*, vol. 12, no. 10, pp. 1274–1286, October 2003.

[3] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, March 1999.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 545–552, December 2006, Vancouver, B.C., Canada.

[6] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–416, June 2011, Colorado Springs, CO.

[7] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, January 2012.

[8] R. Belaroussi, P. Foucher, J-P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," *Proceedings of the International Conference on Pattern Recognition*, pp. 484–488, August 2010, Istambul, Turkey.

[9] A. Mogelmose, M.M. Trivedi, and T.B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, December 2012.

[10] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 922–934, October 2011.

[11] D.C.W. Pao, H.F. Li, and R. Jayakumar, "Shapes recognition using the straight line hough transform: theory and generalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 11, pp. 1076–1089, November 1992.

[12] S. Houben, "A single target voting scheme for traffic sign detection," *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 124–129, June 2011, Baden-Baden, Germany.

[13] H. Fleyeh and Ping Zhao, "A contour-based separation of vertically attached traffic signs," *Proceedings of the Annual Conference of Industrial Electronics*, pp. 1811–1816, November 2008, Orlando, FL.

[14] L.-W. Tsai, J.-W. Hsieh, C.-H. Chuang, Y.-J. Tseng, K.-C. Fan, and C.-C. Lee, "Road sign detection using eigen colour," *IET Computer Vision*, vol. 2, no. 3, pp. 164–177, September 2008.

[15] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 959–973, August 2003.

[16] N. Barnes, A. Zelinsky, and L.S. Fletcher, "Real-time speed sign detection using the radial symmetry detector," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 2, pp. 322–332, June 2008.

[17] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[18] C.G. Keller, C. Sprunk, C. Bahlmann, J. Giebel, and G. Baratoff, "Real-time recognition of U.S. speed signs," *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 518–523, June 2008, Eindhoven, Netherlands.

[19] X. Baro, S. Escalera, J. Vitria, O. Pujol, and P. Radeva, "Traffic sign recognition using evolutionary Adaboost detection and Forest-ECOC classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 113–126, March 2009.

[20] A.R. Rostampour and P.R. Madhvapathy, "Shape recognition using simple measures of projections," *Proceedings of the Annual International Phoenix Conference on Computers and Communications*, pp. 474–479, March 1988, Scottsdale, AR.

[21] P. Gil-Jimenez, S. Lafuente-Arroyo, H. Gomez-Moreno, F. Lopez-Ferreras, and S. Maldonado-Bascon, "Traffic sign shape classification evaluation. Part II. FFT applied to the signature of blobs," *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 607–612, June 2005, Las Vegas, NV.

[22] A. W. Haddad, S. Huang, M. Boutin, and E. J. Delp, "Detection of symmetric shapes on a mobile device with applications to automatic sign interpretation," *Proceedings of the IS&T/SPIE Electronic Imaging on Multimedia on Mobile Devices*, vol. 8304, January 2012, San Francisco, CA.

[23] N. Kiryati, Y. Eldar, and A. M. Bruckstein, "A probabilistic hough transform," *Pattern Recognition*, vol. 24, no. 4, pp. 303–316, February 1991.

[24] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, April 1985.

[25] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, January 1972.

[26] Jack Sklansky, "Finding the convex hull of a simple polygon," *Pattern Recognition Letters*, vol. 1, no. 2, pp. 79–83, December 1982.

Mobile-Based Hazmat Sign Detection and Recognition

Bin Zhao, Albert Parra, Edward J. Delp
Video and Image Processing Laboratory (VIPER)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

***Abstract*—In this paper we describe a mobile-based hazardous material (hazmat) sign detection and recognition system. Hazmat sign detection is based on visual saliency models. We use saliency maps to denote regions that are likely to contain hazmat signs in complex scenes and then use a convex quadrilateral shape detector to find hazmat sign candidates in these regions. Experimental results show that our proposed hazmat sign detection and recognition method is capable of dealing with projective distorted, blurred, and shaded signs. The test image dataset consists of images taken in the field under various lighting and weather conditions, distances, and perspectives.**

***Index Terms*—sign detection, sign recognition, visual saliency model.**

I. Introduction

The handling and transportation of hazardous materials can cause serious accidents and emergency situations. A federal law in the US requires vehicles transporting hazardous materials to be marked with a standard sign (hazmat sign) identifying the type of hazardous material the vehicle is carrying [1]. Hazmat signs help identify the material and determine what special equipment, procedures and precautions should be taken in the event of an emergency. This information is contained in the Emergency Response Guidebook (ERG) published by the US Department of Transportation (DOT) [2]. There exist several mobile-based applications that provide access to this guidebook for first responders in the field. For example, the official ERG 2012 mobile application lets a user browse the ERG guidebook by sign identifiers, template images, and guide pages [2]. The WISER (Wireless Information System for Emergency Responders) mobile application lets a user browse the ERG guidebook by known substance types and hazard classifications [3]. However, these applications only provide ways of manually searching the guidebook. We have developed an integrated mobile-based system that makes use of location-based services and image analysis methods to automatically interpret the hazmat sign and quickly provide guide information to users. We call this system MERGE (Mobile Emergency Response GuidE) [4]. The MERGE mobile application is capable of detecting hazmat signs from an image and querying an internal database to provide accurate information to first responders in real time. MERGE also provides a complete easily searchable version of the Emergency Response Guidebook (ERG) [2]. The ultimate goal of the MERGE system is to recognize a hazmat sign from a distance of greater than 200 feet using a camera in a mobile device.

Hazmat signs have identifying visual information that can be distinguished from their surroundings by specific colors, shapes, symbols, and numbers. However, there exist challenges for successful automatic detection of hazmat signs in complex scenes. These include various lighting and weather conditions that can deteriorate their shape and color over time. Additionally image distortions may occur, such as blur and change in contrast. In [5] we reported some preliminary in this area where we focused on developing shape detection and color recognition methods. In this paper we propose a hazmat sign detection and recognition system based on visual saliency models. Our system uses saliency maps and convex quadrilateral shape detection to overcome the problems mentioned above.

This work was partially supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. Address all correspondence to Edward J. Delp (ace@ecn.purdue.edu).

II. Review of Existing Methods

Sign detection approaches can be divided into three categories: color-based methods [6], shape-based methods [7] and saliency-based methods [8]. Color-based methods take advantage of the fact that signs often have highly visible contrasting colors. These specific colors are used for sign detection. For example, a color histogram backprojection method is used in [9] to detect interesting regions possibly containing hazmat signs. The luminance homogeneity of blocks is used in [10] to identify homogenous regions as the first step towards detection of information signs containing text. In [11] several color components are used to segment traffic signs in various weather conditions. However, these methods are not robust to lighting conditions and illumination changes. Shape-based approaches first generate an edge map and then use shape characteristics to find signs. For example, in [12] triangular, square and octagonal road signs are detected by exploiting the properties of symmetry and edge orientations exhibited by equiangular polygons. A shape classification method of a road-sign detection system in [13] is based on linear and Gaussian-kernel support vector machines (SVM). Shape-based methods are invariant to translation, rotation and scale. Saliency-based approaches make use of visual saliency models to construct saliency maps that denote areas where signs are likely to be found. For example, in [14] a saliency map of road traffic signs is constructed by a weighted sum of color and edge feature maps. A traffic sign recognition system in [15] uses a visual attention system to denote regions with possible candidates.

Visual saliency models are used to model how the human visual system perceives and processes visual stimuli [16]. The main concept is to generate several types of features from an image and fuse them into a scalar map known as a saliency map. For example, a notable saliency-based visual attention (SBVA) model was proposed in [8] using intensity, color and orientation features with a subsampled Gaussian pyramid. In [17] a graph-based visual saliency (GBVS) method forms the activation map from each feature map based on graph theory. A dynamic visual attention (DVA) model based on the rarity of features is proposed in [18]. A multi-scale dissimilarity aggregation (MSDA) method is used to estimate the saliency of regions in [19]. In [20] an image signature (IS) method for image figure-ground separation is described to generate a saliency map in RGB or Lab color spaces. An saliency detector based on hypercomplex Fourier transform (HFT) is presented in [21] using the convolution of the image amplitude spectrum with a low-pass Gaussian kernel.

III. Mobile-Based Hazmat Sign Detection and Recognition

A. System Overview

Figure 1 shows our mobile-based hazmat sign detection and recognition system (MERGE). It consists of an application running on an Android/iOS mobile device and a backend server where many image processing operations are done. There are two basic operational modes: analysis of new or existing images and internal database searching. The first mode includes capturing or selecting an image from the mobile device and performing image preprocessing and image analysis. When available, the accelerometer is used to detect shaking and avoid the loss of focus before taking an image. The image preprocessing includes blur detection and white balance adjustment. Hazmat sign detection and recognition are done on the backend server and the results are sent back to the mobile device. The second mode includes searching the internal database to obtain guide information about a specific hazmat sign. We designed an internal database based on the contents of 2012 ERG guidebook. Hazmat signs can be seached by UN identifiers, classes, symbols, or template images.

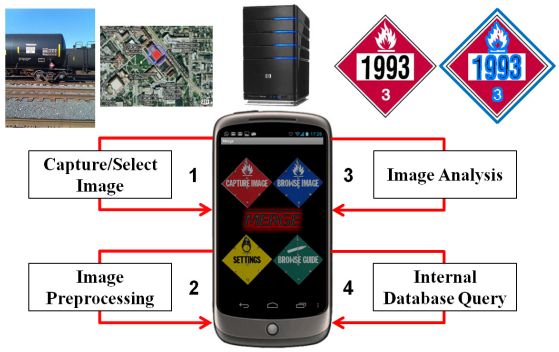


Fig. 1: Mobile-Based Hazmat Sign Detection and Recognition.

Figure 2 shows the user interface and operational workflow. The image analysis results are used for matching related guide pages and querying internal database to retrieve guide information. We display guide information about potential hazards, public safety and emergency response. An evacuation area is also displayed on a map based on the chemical found, the size of the chemical spill and the time of the day. All the information is from the internal database on the mobile application.

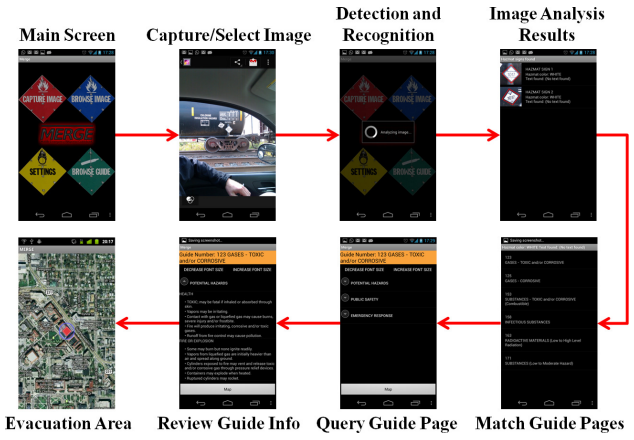


Fig. 2: Mobile Application User Interface and Operational Workflow.

B. Hazmat Sign Detection and Recognition Method

We use the saliency-based method and construct saliency maps to denote regions that are likely to contain hazmat signs in complex scenes. The block diagram in Figure 3 shows the five steps of the proposed hazmat sign detection and recognition method.

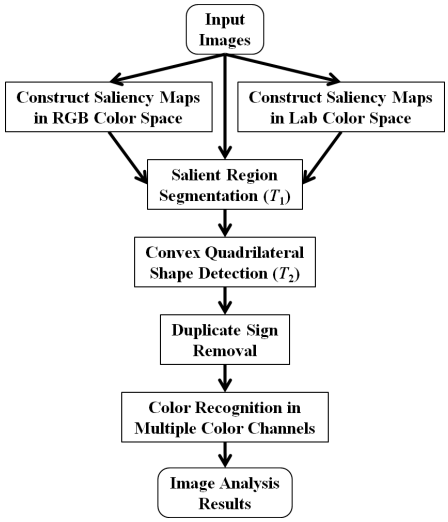


Fig. 3: Proposed Hazmat Sign Detection and Recognition Method.

(1) Construct Saliency Maps in Two Color Spaces: We apply visual saliency models to the input images represented in both RGB and Lab color spaces. In each color space, two saliency maps are constructed using two visual saliency models separately, i.e. IS [20] and HFT [21]. The saliency maps assign higher saliency value to more visually attractive regions. Note that the original HFT method uses the I-RG-BY opponent color space. We modified this method to use RGB and Lab color components with different weights ($[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ for RGB and $[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}]$ for Lab). The combined saliency map method, denoted as IS+HFT(RGB+Lab), generates four saliency maps (two for RGB and two for Lab) and produces the best results in the experiments (see Section IV). **(2) Salient Region Segmentation:** We threshold each saliency map to create a binary mask to segment the salient regions from the original image. The threshold T_1 is determined as k times the average saliency value of a given saliency map. That is, $T_1 = \frac{k}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y)$, where W and H are the width and height of the saliency map, $S(x, y)$ is the saliency value at position (x, y) and k is empirically determined for the combined saliency map method ($k = 4.5$ for IS and $k = 3.5$ for HFT). **(3) Convex Quadrilateral Shape Detection:** For each salient region found, we detect hazmat sign candidates in specific color channels. We used black and white information from grayscale image, and red, green and blue channels from RGB color space. Note that the possible colors for hazmat signs also include yellow and orange, but these can be obtained by transforming the image from RGB to a hue-based color space and then segment the hue channel. The grayscale image and the color channels are thresholded to account for highly chromatic areas using an empirically determined threshold T_2 (85 for black, 170 for white, and 127 for color). Each binarized region is morphologically opened to remove small objects and morphologically dilated to merge areas that may belong to the same object. We then retrieve contours from the resulting binary image using the border following technique proposed in [22]. For each contour, we use the Hough Transform [23] to find straight lines that approximate the

contour as a polygon. The intersections of these lines are the corners of the polygon which can be used to discard non-quadrilateral shapes. If the contour is approximated by four vertices, we find its convex hull [24]. If the convex hull still has four vertices, we check the angles formed by the intersection of its points. If each of these angles is in the range $90^\circ \pm 1.5^\circ$, and the ratio of the sides formed by the convex hull is in the range 1 ± 0.5, we assume that the convex quadrilateral is a hazmat sign candidate. **(4) Duplicate Sign Removal:** For each candidate satisfying the geometric constraints we estimate its minimal bounding box. We then discard any candidate with a bounding box aspect ratio smaller than 1.3. Finally, we check the remaining candidates and remove those that correspond to the same sign. This can be done by first dividing all bounding boxes that overlap more than 50% into groups and then finding the optimal bounding box for each group, which is the box with its nodes closest to its centroid (i.e., closest to a square). Each optimal bounding box is considered to be a detected hazmat sign. **(5) Color Recognition in Multiple Color Channels:** Because signs are detected in specific color channels, the color is recognized directly from the color channel where the sign was identified (black or white for grayscale and red, green or blue for RGB). The recognized color is used for sign identification based on the contents of the 2012 ERG guidebook. Figure 4a illustrates a successful detection of two signs, one of which is affected by perspective and rotation distortion. Figure 4b illustrates a true positive and a false positive.

(a) Two True Positives. (b) One True Positive and One False Positive.

Fig. 4: Examples of Image Analysis Results.

Fig. 5: Samples from the image dataset. From left to right: low resolution, projective distortion, blurred sign, shaded sign.

IV. Experimental Results

We did two experiments to investigate the speed and accuracy of our proposed method. The first experiment consisted of constructing saliency maps using different visual saliency models and evaluating their performance based on ground-truth information. The second experiment consisted of hazmat sign detection and recognition on our image dataset and manually comparing the results with ground-truth information. The tests were executed on a Galaxy Nexus mobile telephone with a dual-core 1.2GHz CPU and 1GB RAM and a backend server with a quad-core 2.4GHz CPU and 4GB RAM. Our image dataset consisted of 50 images, each containing one or more hazmat signs (62 hazmat signs in total). The images were acquired by a third party in the field, under various lighting and weather conditions, distances, and perspectives. Among the 50 images, 23 were taken at 10-50 feet, 23 at 50-100 feet, and 4 at 100-200 feet. Among the 62 hazmat signs, 2 had low resolution, 11 had projective distortion, 8 were blurred, and 6 were shaded. Figure 5 illustrates some samples from the image dataset. The images were taken using three different cameras: a 8.2 MP Kodak Easyshare C813, a 16 MP Nikon Coolpix S800c, and a 5 MP camera on an HTC Wildfire mobile telephone. The ground-truth information included the distance from the camera to the sign, sign color, projective distortion of the sign, image resolution, possible shadow affecting the sign, and sign location on the image. Note that we only used the color and not the text inside the sign for sign identification for these experiments.

TABLE I: Average Execution Time (in Seconds), Distribution and Score of Each Saliency Map Method (Color Spaces).

Saliency Map	Time	Good	Fair	Bad	Lost	Score
SBVA(I-RG-BY)	2.07	34	16	11	1	145
GBVS(I-RG-BY)	3.36	30	15	15	2	135
DVA(RGB)	0.43	19	2	11	30	72
MSDA(RGB)	3.74	22	7	27	6	107
IS(I-RG-BY)	0.43	23	4	17	18	94
IS(RGB)	0.36	45	8	4	5	155
IS(Lab)	0.39	27	5	20	10	111
HFT(I-RG-BY)	0.59	33	8	12	9	127
HFT(RGB)	0.53	38	5	8	11	132
HFT(Lab)	0.55	37	10	8	7	139
IS(RGB+Lab)	0.75	52	6	1	3	169
HFT(RGB+Lab)	1.08	41	6	8	7	143
IS+HFT(RGB+Lab)	1.83	55	4	2	1	175

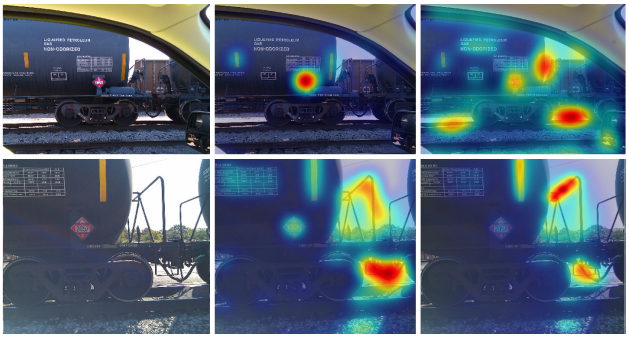


Fig. 6: Saliency map categories (top to bottom, left to right): original image, good, fair; original image, bad, lost.

Table I shows the results of our first experiment, including average execution times and scores. The saliency map methods evaluated in the experiment are: SBVA [8], GBVS [17], DVA [18], MSDA [19], IS [20], HFT [21]. We classified the resulting saliency maps into four categories: good, fair, bad, and lost. For each sign, we assigned 3 points to a good map (sign was mostly contained in a high saliency-valued region), 2 points to a fair map (sign was mostly contained in a middle saliency-valued region), 1 point to a bad map (sign was mostly contained in a low saliency-valued region), and 0 points to a lost map (sign was not contained in any saliency-valued region). Figure 6 illustrates examples of each category. The score of each saliency map method is calculated as the sum of the points assigned to all 62

TABLE II: Image Analysis Results.

Proposed Method	Overall Accuracy	Detected Signs	Total Signs
IS(RGB+Lab)	51.6%	32	62
HFT(RGB+Lab)	38.7%	24	62
IS+HFT(RGB+Lab)	64.5%	40	62

hazmat signs, which ranges from 0 to 186. Compared with the SBVA and the GBVS methods using one color space, the IS and the HFT methods using one color space have comparable scores, while the IS and the HFT methods using two color spaces have higher scores. The IS(RGB+Lab), the HFT(RGB+Lab) and the IS+HFT(RGB+Lab) methods using two color spaces run 2.76, 1.93, and 1.14 times faster than the SBVA method and 4.48, 3.13, and 1.84 times faster than the GBVS method respectively. The results verified that the IS and the HFT methods can be combined to improve the score of IS+HFT method, while still running faster than SBVA and GBVS methods.

Table II shows the image analysis results of our second experiment. The overall sign detection accuracy is closely related to the number of pixels on a hazmat sign, which is mainly influenced by the distance from a camera in a mobile device to a hazmat sign and the resolution of the image captured by the camera. Compared with the proposed IS(RGB+Lab) and the HFT(RGB+Lab) methods using one saliency map method, our proposed IS+HFT(RGB+Lab) method using two saliency map methods has higher accuracy. The proposed IS+HFT(RGB+Lab) method has an overall sign detection accuracy of 64.5% for all 62 hazmat signs. Note that its overall accuracy is 71.9% for the 32 hazmat signs in the 50-100 feet range and 50.0% for the 6 hazmat signs in the 100-200 feet range. We can increase the overall accuracy by improving the adaptive thresholding method used in the saliency region segmentation and the morphological operations used in the convex quadrilateral shape detection. We determined the color recognition accuracy based on how many signs were correctly color recognized after a successful sign detection. The color recognition accuracies of the proposed methods using IS(RGB+Lab), HFT(RGB+Lab) and IS+HFT(RGB+Lab) are 37.1%, 30.6%, and 51.6% respectively. Note that the sign color recognition was done only if a sign was successfully detected, and that multi-colored signs may also cause our method to misidentify the sign color, given that we detect signs at individual color channels. Color recognition accuracy is affected by the absence of color calibration in the step of image preprocessing. The overall average execution times of the proposed methods using IS(RGB+Lab), HFT(RGB+Lab) and IS+HFT(RGB+Lab) are 2.60, 2.49, and 5.09 seconds in total respectively. The proposed IS+HFT(RGB+Lab) method is still suitable for real-time applications.

V. Conclusions

We described a mobile-based hazmat sign detection and recognition system that uses saliency maps to segment salient regions and a convex quadrilateral shape detector to find hazmat sign candidates in these regions. Our experimental results show that our proposed hazmat sign detection and recognition method is capable of dealing with projective distorted, blurred, and shaded signs. Our proposed method has an overall sign detection accuracy of 64.5%, a color recognition accuracy of 51.6%, and an overall average execution time of 5.09 seconds. Further investigation on the influences of distance and image resolution is needed. We will also develop methods for color calibration to improving color recognition accuracy, and methods for sign character recognition to interpret the text inside the detected hazmat signs.

References

[1] United States Department of Transportation, *Code of Federal Regulations, Title 49, DOT Hazmat*, Labelmaster, 2012 edition, October 2012.

[2] ERG, Available: http://www.phmsa.dot.gov/hazmat/library/erg.

[3] WISER, Available: http://wiser.nlm.nih.gov.

[4] MERGE, Available: http://www.hazmat-signs.org.

[5] A. Parra, B. Zhao, A. Haddad, M. Boutin, and E. J. Delp, "Hazardous material sign detection and recognition," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, September 2013, Melbourne, Australia (to appear).

[6] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, March 1999.

[7] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE Transactions on Image Processing*, vol. 12, no. 10, pp. 1274–1286, October 2003.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[9] D. Gossow, J. Pellenz, and D. Paulus, "Danger sign detection using color histograms and SURF matching," *Proceedings of the IEEE International Workshop on Safety, Security and Rescue Robotics*, pp. 13–18, October 2008, Sendai, Japan.

[10] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 922–934, October 2011.

[11] L. Song and Z. Liu, "Color-based traffic sign detection," *Proceedings of the International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pp. 353–357, June 2012, Chengdu, China.

[12] G. Loy and N. Barnes, "Fast shape-based road sign detection for a driver assistance system," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 70–75, September 2004, Stockholm, Sweden.

[13] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. Lopez-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 264–278, June 2007.

[14] W.-J. Won, M. Lee, and J.-W. Son, "Implementation of road traffic signs detection based on saliency map model," *Proceedings of the IEEE Intelligent Vehicles Symposium (IVS)*, pp. 542–547, June 2008, Eindhoven, Netherlands.

[15] R. Kastner, T. Michalke, T. Burbach, J. Fritsch, and C. Goerick, "Attention-based traffic sign recognition with an array of weak classifiers," *Proceedings of the IEEE Intelligent Vehicles Symposium (IVS)*, pp. 333–339, June 2010, San Diego, CA, USA.

[16] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, January 2013.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 545–552, December 2006, Vancouver, BC, Canada.

[18] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 681–688, December 2008, Vancouver, BC, Canada.

[19] Chelhwon Kim and Peyman Milanfar, "Visual saliency in noisy images," *Journal of Vision*, vol. 13, no. 4, pp. 1–14, March 2013.

[20] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, January 2012.

[21] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, April 2013.

[22] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, April 1985.

[23] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, January 1972.

[24] Jack Sklansky, "Finding the convex hull of a simple polygon," *Pattern Recognition Letters*, vol. 1, no. 2, pp. 79–83, December 1982.

Recognition, Segmentation and Retrieval of Gang Graffiti Images on a Mobile Device

Albert Parra, Bin Zhao, Joonsoo Kim, Edward J. Delp
Video and Image Processing Lab (VIPER)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA

***Abstract*—In this paper we describe three methods for recognition, segmentation and retrieval of gang graffiti images. The first method is color recognition based on touchscreen tracing, the second method is color image segmentation based on Gaussian thresholding and the third method is content based image retrieval. Our experimental results show an image retrieval accuracy of 92.8% for gang graffiti scene recognition and an image retrieval accuracy of 50.0% for gang graffiti component classification. The experiments also show an average image retrieval time of 0.56 seconds, from which the scoring process takes on average 984 microseconds.**

***Index Terms*—Color recognition, image segmentation, content based image retrieval, Gaussian thresholding, SIFT.**

I. Introduction

Gangs are a serious threat to public safety throughout the United States. They are responsible for an increasing percentage of crime and violence [1]. Street gang graffiti is their most common way to communicate messages, including challenges, warnings or intimidation to rival gangs. It is an excellent way to track gang affiliation, growth, and membership. The goal of our work is to develop a mobile-based system capable of using location-based-services, combined with image analysis methods, to provide accurate and useful information concerning gangs based on a network connected database of gang graffiti images. In this paper we describe three methods for recognition, segmentation and retrieval of gang graffiti images. Our first method is color recognition of gang graffiti based on touchscreen tracing. Our second method is color image segmentation based on Gaussian thresholding. Our third method is content based gang graffiti image retrieval. Our content based image retrieval method is tested for accuracy and speed in two scenarios: scene recognition and gang graffiti component classification.

Previous work on gang graffiti: In [2] methods for segmenting and retrieving graffiti images are described using global thresholding and template matching. In [3] we described a mobile-based gang graffiti system that uses location information for querying a database of graffiti images. That system does not include image segmentation or matching. Other approaches described in [4], [5], [6], [7] do not use the bag-of-words models for image retrieval of gang graffiti and tattoos and report slower matching and retrieval times than we demonstrate in our experiments. In [8], [9], bag-of-words models for image retrieval of gang or gang-like tattoos are used but are not intended for real-time retrieval in mobile-based environments.

II. Review of Existing Methods

Color Recognition: Gang graffiti are often sprayed in non-uniform surfaces, which makes them difficult to distinguish from the background. Since our system is deployed on a mobile telephone, we take advantage of the touchscreen capabilities of modern mobile devices to aid the recognition of color in gang graffiti images. The touchscreen can be used to detect a path drawn with the finger on the screen for image analysis such as color recognition. Color recognition techniques using tactile feedback use thresholds based on perceptual attributes of specific color spaces. The perceptual thresholds (also known as discrimination thresholds) have been widely studied for human observers [10]. However, some methods do use thresholds based on human perceptibility, but use application based thresholds. For example, some skin detection methods use an adaptive skin color filter to detect color regions, by setting thresholds in both RGB and HSV color spaces [11], [12].

Color Image Segmentation: Image color segmentation techniques can be divided into three categories [13]: physics based, feature-space based, and image-domain based. Methods based on physics include dichromatic reflection models [14] and unichromatic reflection models [15] for single illumination sources, and a more general model of image formation [16] for multiple illuminations. Methods based on feature spaces can be sub-categorized into three groups: clustering of regions given patterns with specific properties, including methods such as k-means clustering [17] or Iterative Self-Organizing Data Analysis Technique (ISODATA) [18]; adaptive k-means clustering, including methods based on maximum a posteriori (MAP) estimation [19] or split-and-merge strategies [20]; and histogram thresholding, including methods based on RGB thresholding and hue information [21], specific skin color domains [22], or entropy thresholding [23]. Methods based on the image-domain can be subcategorized into four groups: split-and-merge, including methods such as region smoothing by Markov Random Fields (MRF) [24] or splitting by either watershed transform [25] or quad-tree image representation for segmentation of skin cancers [26], among others; region growing, including methods such as RGB color distribution growing, HSV morphological open-close growing, or color quantization growing [27]; classification based, including methods such as minimization of Hopfield networks [28], or background extraction using two three-layered neural network [29]; edge based techniques, including methods such as combination of HSI gradients [30], active contours, or the Mumford-Shah variation model [31].

Content Based Image Retrieval: Content Based Image Retrieval (CBIR) consists of four core techniques [32]: visual signature extraction, similarity measures, and classification and clustering. Visual signature extraction usually implies three steps: 1) segmenting images using methods such as k-means clustering [33], normalized cuts [34], or salient region detection [35]; 2) extracting features such as color, texture, or shape [36]; 3) constructing the signatures using distri-

978-1-4799-1535-4/13/$31.00 ©2013 IEEE

butions [37] or adaptivity [38]. Similarity measure methods include manifold embedding [39], and vector quantization [40]. Classification and clustering methods include hierarchical k-means [41], support vector machine [42], or Bayesian classifiers [43].

Our color image segmentation approach falls into the feature-space based techniques. However, our approach differs from the methods mentioned above. Although there are some techniques in the literature that use only hue or luma information, either circular histogram thresholding [44] or one-dimensional histogram thresholding [45], we do not obtain the descriptors of the probability distribution from the color histogram of the image. Instead, the median and the variance obtained from the tracing-bases color recognition process are used for segmentation. Our segmentation approach does not produce binarized images, but grayscale images weighed by a Gaussian distribution, thus creating a probability map for a specific luma or hue. These types of probability maps are used for increased accuracy and robustness in some clustering techniques [46], [47]. Our content based image retrieval approach uses hierarchical k-means to build a vocabulary tree based on the method in [41].

III. Gang Graffiti Color Recognition and Segmentation

One of the goals of our system is to identify the color of graffiti components of an image. We use features and a priori information specific to gang graffiti. First, gang graffiti are mostly monochromatic. This makes it easier to segment graffiti components. Second, gang graffiti are hand-written with each gang member having a different style. This rules out the use of generic Optical Character Recognition methods. Third, gang graffiti are almost always painted on non uniform surfaces with various textures, such as walls, garage doors, or trees. This makes the segmentation of the graffiti contents more challenging. We developed a method for identifying the color of a graffiti component. We call this approach color recognition based on tracing. It can be implemented on a hand-held device without the need of an network connection. We also describe a method for segmenting an image based on the colors determined by our tracing method. We call this color image segmentation based on Gaussian thresholding.

Color Recognition Based on Touchscreen Tracing: In this method the user takes an image of a gang graffiti and traces a path along a colored region using the touchscreen display. Then we recognize the color along the path, and provide a list of gangs related to the color by querying an internal database on the mobile phone. For this method we use an RGB to Y'CH color space conversion. Figure 1 shows an overview of our color recognition method. The path is drawn along a component of the graffiti image assumed to have uniform color. The RGB color components of each pixel on the path are converted to a new luma/chroma/hue color space that we call the Y'CH color space. The Y'CH color space is used because color changes are more intuitive and perceptually relevant to represent in luma or hue than in RGB triplets, in order to obtain the median and the variance of the color along the traced path. Equation 1 shows the mapping between RGB and Y'CH. Note that we use luma (Y') as opposed to luminance (Y) [48]. Third, we compute three medians on the pixel array that forms the path, namely the luma median ($\tilde{Y}$), the chroma median ($\tilde{C}$) and the hue median ($\tilde{H}$). We then define three disjoint regions in our Y'CH color space (labeled 3a, 3b and 3c in Figure 1), delimited by manually set thresholds based on luma ($T_{Y_w} = 0.12$, $T_{Y_b} = 0.85$) and chroma ($T_C = 0.05$). These thresholds were empirically obtained from our database of gang graffiti, consisting of more than 600 gang graffiti images. Depending on the region where the medians are located, we do color recognition based on luma (3a) or hue (3b). Once we have the median, either based on luma or hue, we need to decide which color is associated with it. From all the images in our database, the possible colors for gang graffiti components are black, white, red, blue, green, gold and purple. If the median is based on luma, the color detected is either black ($\tilde{Y} \leq 0.5$) or white ($\tilde{Y} > 0.5$). If the median is based on hue, the color detected is $H_d = \min_i(\theta(\tilde{H}, H_{A_i}))$, where $\theta(\tilde{H}, H_{A_i})$ is the angular distance between the computed hue ($\tilde{H}$) and the i-th component of a set of average hues (H_A), empirically obtained from analyzing 100 color calibrated images taken from our database. These values are (color, H_A)=({Red, 6.10 rad}, {Blue, 4.00 rad}, {Green, 2.20 rad}, {Gold, 0.69 rad}, {Purple, 5.15 rad}). Once the color is detected, we provide a list of gangs related to that color by querying an internal database on the mobile phone. Finally, we also estimate the variance $\sigma^2_{\tilde{X}}$ around the computed median $\tilde{X} = \{\tilde{Y}$ or $\tilde{H}\}$. This variance is used as an input to the color image segmentation method described next. Note that this method can be used with multi-colored graffiti by using it on each trace on the touchscreen.

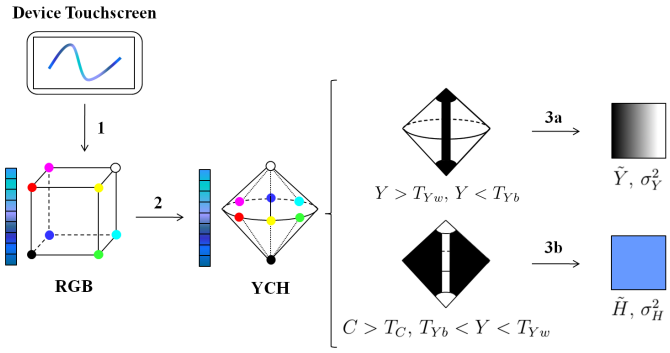


Fig. 1: Color Recognition Using Touch Screen Tracing.

$$
\begin{aligned}
&Y' = 0.299R + 0.587G + 0.114B.\\
&C = max(R,G,B) - min(R,G,B) = M - m\\
&H = \begin{cases} \frac{G-B}{C} & \text{if } M = R \\ \frac{B-R}{C} + 2 & \text{if } M = G \\ \frac{R-G}{C} + 4 & \text{if } M = B \\ 0 & \text{if } C = 0 \end{cases}
\end{aligned} \tag{1}
$$

Color Image Segmentation Based on Gaussian Thresholding: For the segmentation we use a Gaussian threshold near a specific luma or hue value in the Y'CH color space, in order to produce a segmented image where each pixel is given a weight depending on its distance from a median. Figure 2 shows an overview of our color segmentation method divided in 5 steps. We assume that, given a graffiti image, we have the median $\tilde{X}$ and the variance, $\sigma^2_{\tilde{X}}$, of a traced path (step 1a). We then transform the entire RGB image to the our Y'CH color space (steps 1b and 2). We finally segment the image using Gaussian thresholding (steps 3 to 5). The segmentation works as follows. We first ignore all pixels in the image that fall outside our manually set thresholds in the Y'CH color space (step 3). Note that these thresholds are the same as the ones used for the color recognition process. We apply a weight to the rest of the pixels using a normal distribution centered at $\tilde{X}$ and a confidence interval of $2\sigma_{\tilde{X}}$ (step 4). The output

of this process is a grayscale image where each pixel is given a probability based on a normal distribution (step 5). This probability is higher as the pixel value gets closer to $\tilde{X}$. The image is then scaled to $[0, 255]$. Figure 3 shows an example where the color recognition process is performed by tracing a path along the blue numbers "2" and "5" in the graffiti component. Figure 4 shows the effect of the Gaussian thresholding process on the letters "Hill". Note that this method produces a probability map, where the values in a graffiti component decrease as the spray paint fades.

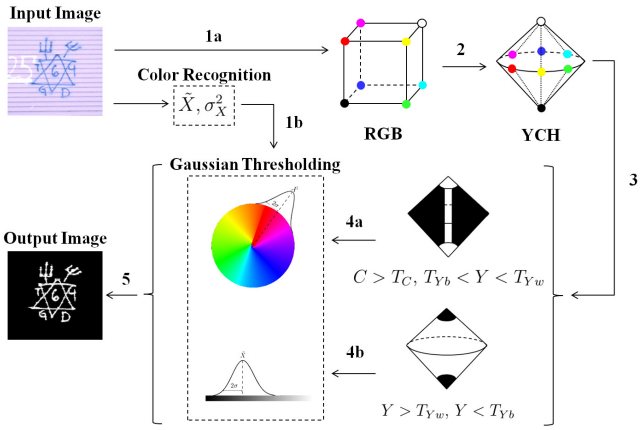


Fig. 2: Color Image Segmentation Using Gaussian Thresholding.

(a) Input Image (b) Segmented Image

Fig. 3: Gaussian Thresholding on blue. $(\tilde{H}, \sigma_{\tilde{H}}^2) = (4.19, 0.05)$.

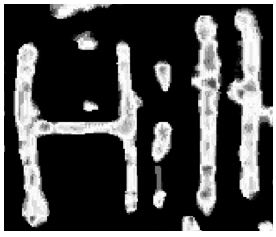


Fig. 4: Probability map created by the Gaussian Thresholding.

IV. Content Based Gang Graffiti Image Retrieval

We describe a method to recognize gang graffiti by matching image features from query images against our database of gang graffiti. The method is currently used in two scenarios: 1) recognize scenes containing graffiti (Figure 5) and 2) classify individual graffiti components (Figure 6). In both cases we use a vocabulary tree [41] to retrieve query images. The vocabulary tree is obtained as follows. First, we extract SIFT features from a set of training images to get sets of 128 dimensional vectors. To recognize scenes containing graffiti, we extract SIFT features from the entire image, similar to the work done in [4], [5], [6], [7], [8], [9] for graffiti and tattoo images. To recognize individual graffiti components, we extract features from segmented graffiti components. By graffiti components we mean the objects and shapes contained in a graffiti image, such as stars, pitchforks, crowns, and arrows. These components are found by first performing color recognition based on touchscreen tracing, then applying our color image segmentation based on Gaussian thresholding, and finally segmenting each graffiti component. Note that at this stage we binarize the probability map returned by the Gaussian thresholding method. That is, we do not take into account the strength of the graffiti trace. Also note that once we segment the gang graffiti from the background we segment the graffiti components manually. Then, we use hierarchical k-means clustering to recursively divide the $\mathbb{R}^{128}$ space into sub-clusters, as in [9]. Each sub-cluster contains the set of descriptors closest to its center. We call each of these sub-clusters a word. This clustering can be interpreted as a vocabulary tree, where k corresponds to the branching factor at each level, and each word corresponds to a leave in the tree. Figure 7 illustrates this equivalence. Each black dot corresponds to a descriptor from a training image. Note that we keep track of the image corresponding to each descriptor. Therefore, each word can be associated to a number representing a path down the vocabulary tree. At the end of the training, each image i can be represented as an n_w dimensional vector d_i, n_w being the total number of words in the tree. At each index j, an entropy weighting [41] is applied so that $d_i[j] = N_j^i \ln \frac{N}{N_j}$, where N_j^i is the the number of descriptors of the i-th training image associated with the j-th word, N is the total number of training images, and N_j is the number of training images with at least one descriptor belonging to the j-th word. Note that d_i is normalized to make it invariant to the total number of descriptors found on the i-th image. Based on the results of [41] we chose $k = 3$ and $10,000$ leaves to create our vocabulary tree. In order to match a query image to an image in our database we first extract SIFT descriptors from the query image. Each of the query descriptors is pushed down the vocabulary tree to find its closest word, and an n_w dimensional vector q is created following the same criteria as in the training process, such that $q[j] = N_j^q \ln \frac{N}{N_j}$, where N_j^q is the the number of descriptors of the query image associated with the j-th word. The closest match to the query image is then $\min_i ||q - d_i||$. Since all the images in our database have location information, we can improve the performance by comparing q to a set of training images in a certain physical radius from the query image. Note that instead of computing norms between the query vector and each training image we could use inverted files in memory to speed up the process [41]. With the use of the location information we reduce the search on average to 10 images instead of the entire training set. The main advantage of using a vocabulary tree for image retrieval is that its leaves define the quantization, thus making the comparison dramatically less expensive than previous methods in the literature. Also, once the vocabulary tree is built, new images can be added by just pushing down its descriptors. Currently, SIFT features are used for both scenes containing graffiti and individual graffiti components. However, note that the k-means clustering accepts any type multi-dimensional vector, hence we can use gang graffiti related features in the future to improve the retrieval performance.

V. Experimental Results

We did two experiments to determine the accuracy and the speed of our image retrieval approach. The tests were executed in a desktop computer with a 2.8GHz CPU and 2GB RAM. The goal of the first

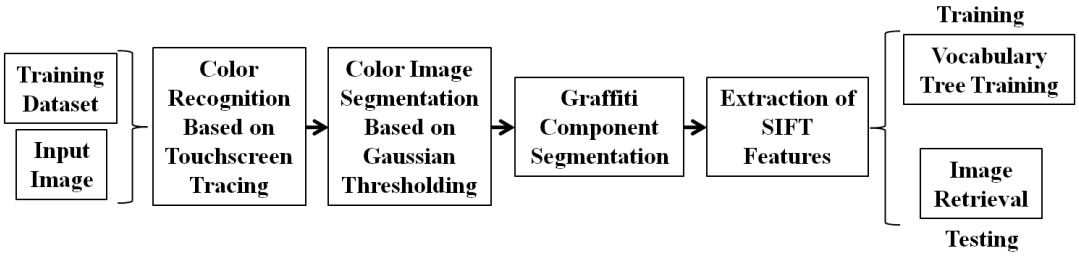


Fig. 6: Gang Graffiti Component Classification.

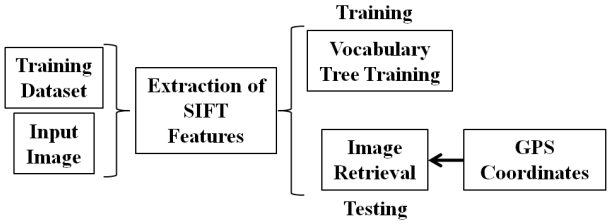


Fig. 5: Gang Graffiti Scene Recognition.

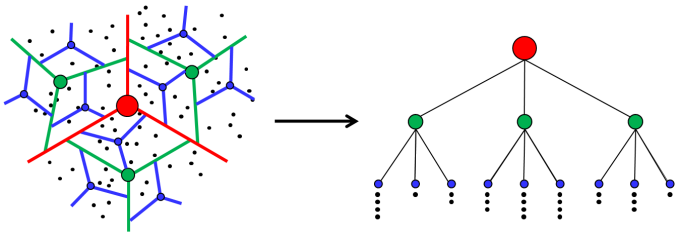


Fig. 7: Vocabulary Tree From Hierarchical k-Means.

experiment was to match query images to images in our database based on the scene. That is, by finding features not only from the graffiti in the image, but also of the background. We trained 708 images from our database and used hierarchical k-means to create a vocabulary tree. Figure 8 shows some sample images. Based on the results of [41] and the size of our dataset we set $k = 3$ and the number of leaves in the vocabulary tree to 300. A separate set of 100 images was used for testing. Each of the test images corresponded to one of the scenes in our database, but with different viewpoint, rotation, and illumination. For each test image we retrieved its 5 closest matches from the training set and we gave it a score from 5 to 0, 5 meaning that the matching scene was in first position, and 0 meaning that there was no matching scene in the top 5 results. Table I summarizes the results of the first experiments. It is worth noting that although this experiment only accounted for scene recognition, we found that sometimes the results returned included scenes close to the query. Figure 9 illustrates an example. These results can be used to recognize nearby graffiti or even graffiti that have been removed. The average execution time of the first experiment was 0.55 seconds. The goal of the second experiment was to classify query images into categories based on a set of gang graffiti symbols. We trained 162 images, each one consisting of one graffiti component in black with white background. A separate set of 40 images was used for testing. Each of the test images also consisted of one graffiti component in black with white background. We considered 33 distinct graffiti components (i.e., classes), including 5-point star, 6-point star, pitchfork, or arrow. Figure 10 illustrates an example of each class. For each test image we retrieved its 5 closest matches on the training set and we associated a class to it based on a scoring scheme. Given the scores of the 5 for closest matches $s = \{s_1, s_2, s_3, s_4, s_5\}$ in ascending order, we normalize them and invert them so that the new scores become $p = \{p_1, p_2, p_3, p_4, p_5\}$, where $p_i = 1 - \frac{s_i}{\sum s_i}$. Then, we manually group the 5 closest matches into N classes, $N \in \{1, ..., 33\}$. We add up the new scores associated to each class, and we assign the class C with the highest score to the query image, such that $C = \underset{n}{\text{argmax}}\{\sum_k p_k^{(n)}\}$, where k is the set of indeces of s belonging to the n-th class, $n \in \{1, ..., N\}$. Given the 40 testing images, containing multiple examples from the 33 classes, the classification accuracy is 50.0%. The classification accuracy for the class *6-point star* is 66.7%, while the classification accuracy for the class *pitchfork* is 33.6%. Thus, using SIFT features for graffiti component classification is good for some classes but poor for others. The fundamental reason why SIFT is not good enough for component classification is that the some classes overlap. For example, *5-point star* includes *X*, and *pitchfork* includes *3*. The average execution time of the second experiment was 0.57 seconds. Note that although the training set for the second experiment is smaller than the training set for the first experiment, the running times for the two experiments are similar. This is because the process that takes longer (an average of 0.45 seconds in both experiments) is finding the paths of the query image features down the vocabulary tree. The scoring process takes an average of 984 microseconds in both experiments.

TABLE I: Experimental Results For Scene Recognition.

	Score						
Images	5	4	3	2	1	0	Accuracy
100	91	1	1	1	0	6	92.8%

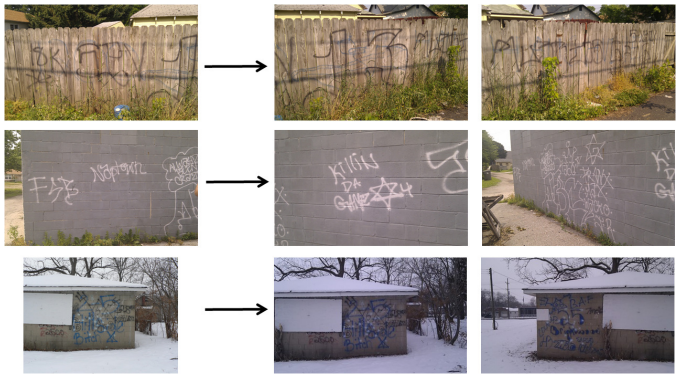


Fig. 9: Query images (left) and similar retrieved scenes (right).

Fig. 8: Sample images from our training image dataset.

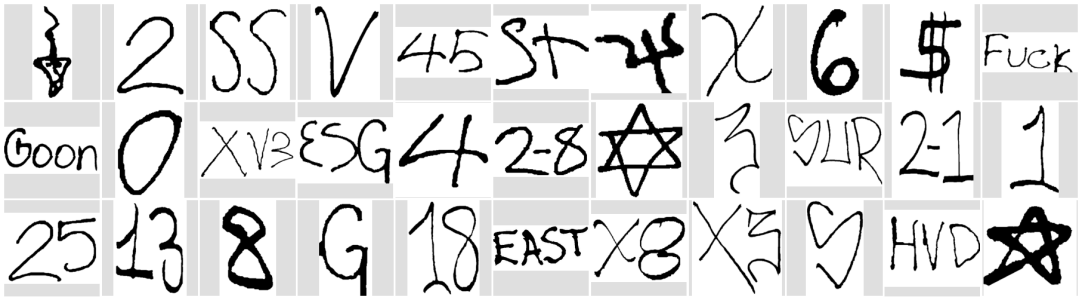


Fig. 10: Gang Graffiti Component Classes.

VI. Conclusions and Future Work

We described two methods for color recognition and segmentation of gang graffiti and one method for content based gang graffiti image retrieval. We evaluated the accuracy and speed of the content based gang graffiti image retrieval in two scenarios: scene recognition and gang graffiti component classification. The experimental results showed that the use of SIFT features for scene recognition produce very accurate outcomes, but the use of SIFT features for component classification does not produce accurate results. The experiments also showed that the image retrieval is fast in both scenarios. We can increase the accuracy of the component classification by creating our own set of gang graffiti component features. The set would include information such as the component color, aspect ratio, or position and alignment with respect to other components. We can also use the probability map created by the Gaussian thresholding method to estimate the direction of the graffiti trace, and use that information as a component feature. These new features can still be used to train a vocabulary tree using the current approach. We are also investigating methods for automatic graffiti component segmentation based on elastic shape recognition [2]. Finally, we are working on implementing a mobile-based version of our classification system by optimizing the feature extraction (i.e. using Dense SIFT features instead of SIFT) and the image retrieval method (i.e. using inverted file lookup methods).

References

[1] National Drug Intelligence Center (NDIC), *Attorney General's Report to Congress on the Growth of Violent Street Gangs in Suburban Areas*, United States Deptartment of Justice, April 2008.

[2] C. Yang, P. C. Wong, W. Ribarsky, and J. Fan, "Efficient graffiti image retrieval," *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pp. 36:1–36:8, June 2012, Hong Kong, China.

[3] A. Parra, M. Boutin, and E. J. Delp, "Location-aware gang graffiti acquisition and browsing on a mobile device," *Proceedings of the IS&T/SPIE Electronic Imaging on Multimedia on Mobile Devices*, pp. 830402–1–13, January 2012, San Francisco, CA, USA.

[4] A. K. Jain, J.-E. Lee, R. Jin, and N. Gregg, "Content-based image retrieval: An application to tattoo images," *Proceedings of the International Conference on Image Processing*, pp. 2745–2748, 2009.

[5] J.-E. Lee, A.K. Jain, and R. Jin, "Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification," *Proceedings of the Biometrics Symposium*, pp. 1–8, September 2008, Tampa, Florida, USA.

[6] W. Tong, J.-E. Lee, R. Jin, and A. K. Jain, "Gang and Moniker Identification by Graffiti Matching," *Proceedings of the 3rd ACM Workshop on Multimedia in Forensics and Intelligence*, November 2011, Scottsdale, AZ.

[7] A. K. Jain, J.-E. Lee, and R. Jin, "Graffiti-ID: Matching and Retrieval of Graffiti Images," *Proceedings of the 1st ACM Workshop on Multimedia in Forensics*, pp. 1–6, October 2009, Beijing, China.

[8] D. Manger, "Large-scale tattoo image retrieval," *Proceedings of the Conference on Computer and Robot Vision*, pp. 454–459, May 2012, Toronto, Ontario, Canada.

[9] J.-E. Lee, R. Jin, A. K. Jain, and W. Tong, "Image retrieval in forensics: Tattoo image database application," *IEEE MultiMedia*, vol. 19, no. 1, pp. 40–49, 2012.

[10] J. Krauskopf and G. Karl, "Color Discrimination and Adaptation," *Vision Research*, vol. 32, no. 11, pp. 2165–2175, January 1992.

[11] K.-M. Cho, J.-H. Jang, and K.-S. Hong, "Adaptive Skin-Color Filter," *Pattern Recognition*, vol. 34, no. 5, pp. 1067–1073, May 2001.

[12] R.M. Jusoh, N. Hamzah, M.H. Marhaban, and N.M.A. Alias, "Skin Detection Based on Thresholding in RGB and Hue Component," *Proceedings of the 2010 IEEE Symposium on Industrial Electronics Applications*, pp. 515–517, October 2010, Penang, Malaysia.

[13] S. R. Vantaram and E. Saber, "Survey of contemporary trends in color image segmentation," *Journal of Electronic Imaging*, vol. 21, no. 4, pp. 040901–1–040901–28, October 2012.

[14] R.T. Tan and K. Ikeuchi, "Separating reflection components of textured surfaces using a single image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 178–193, February 2005.

[15] G. Healey, "Segmenting Images Using Normalized Color," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 64–73, January 1992.

[16] B. A. Maxwell and S. A. Shafer, "Physics-Based Segmentation of Complex Objects Using Multiple Hypotheses of Image Formation," *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 269–295, November 1997.

[17] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 604–610, October 2005, Montbonnot, France.

[18] Y. Tarabalka, J.A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 8, pp. 2973–2987, August 2009.

[19] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 313–320, June 2003, Urbana, IL, USA.

[20] A.L.N. Fred and A.K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, June 2005.

[21] H. Gomez-Moreno, S. Maldonado-Bascon, P. Gil-Jimenez, and S. Lafuente-Arroyo, "Goal evaluation of segmentation algorithms for traffic sign recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 4, pp. 917–930, December 2010.

[22] S.L. Phung, A. Bouzerdoum, and Sr. Chai, D., "Skin segmentation using color pixel classification: analysis and comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 148–154, January 2005.

[23] C-I Chang, Y. Du, J. Wang, S-M Guo, and P.D. Thouin, "Survey and comparative analysis of entropy and relative entropy thresholding techniques," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 153, no. 6, pp. 837–850, December 2006.

[24] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, July 2003.

[25] V. Grau, A. U J Mewes, M. Alcaniz, R. Kikinis, and S.K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, April 2004.

[26] A.J. Round, A.W.G. Duller, and P.J. Fish, "Colour Segmentation for Lesion Classification," *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, pp. 582–585, November 1997, Chicago, IL, USA.

[27] Y. Deng and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, August 2001.

[28] M.E. Plissiti, D.I. Fotiadis, L.K. Michalis, and G.E. Bozios, "An automated method for lumen and media-adventitia border detection in a sequence of ivus frames," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 2, pp. 131–141, June 2004.

[29] N. Funakubo, "Feature Extraction of Color Texture Using Neural Networks for Region Segmentation," *Proceedings of the 20th Annual Conference of IEEE Industrial Electronics*, vol. 2, pp. 852–856, September 1994, Bologna, Italy.

[30] T. Carron and P. Lambert, "Color Edge Detector Using Jointly Hue, Saturation and Intensity," *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, pp. 977–981, November 1994, Austin, TX, USA.

[31] T.F. Chan and L.A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, February 2001.

[32] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, May 2008, New York, NY, USA.

[33] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, July 2002.

[34] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, August 2000.

[35] F. Zhu, M. Bosch, N. Khanna, C.J. Boushey, and E.J. Delp, "Multilevel segmentation for food classification in dietary assessment," *Proceedings of 7th International Symposium on Image and Signal Processing and Analysis*, pp. 337–342, September 2008, Dubrovnik, Croatia.

[36] D.E. Ilea and P.F. Whelan, "CTex - an adaptive unsupervised segmentation algorithm based on color-texture coherence," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1926–1939, October 2008.

[37] J. Li and J.Z. W., "Studying digital imagery of ancient paintings by mixtures of stochastic models," *IEEE Transactions on Image Processing*, vol. 13, no. 3, pp. 340–353, March 2004.

[38] H. Muller, T. Pun, and D. Squire, "Learning from user behavior in image retrieval: Application of market basket analysis," *International Journal of Computer Vision*, vol. 56, pp. 65–77, January 2004.

[39] J. He, H. Tong, M. Li, H-J. Zhang, and C. Zhang, "Mean version space: a new active learning method for content-based image retrieval," *Proceedings of the ACM SIGMM international workshop on Multimedia information retrieval*, pp. 15–22, October 2004, New York, NY, USA.

[40] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," *IEEE Transactions on Image Processing*, vol. 13, no. 5, pp. 699–709, May 2004.

[41] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2161–2168, June 2006, Washington, DC, USA.

[42] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," *Proceedings of the ACM international conference on Multimedia*, pp. 107–118, October 2001, Ottawa, Canada.

[43] Z. Su, H. Zhang, S. Li, and S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 924–937, August 2003.

[44] D.-C. Tseng, Y.-F. Li, and C.-T. Tung, "Circular Histogram Thresholding for Color Image Segmentation," *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, vol. 2, pp. 673–676, August 1995, Montreal, Canada.

[45] D.-C. Tseng and C.-H. Chang, "Color Segmentation Using Perceptual Attributes," *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, vol. 3, pp. 228–231, September 1992, La Haye, Holland.

[46] J.D. Brand and J.S.D. Mason, "Skin Probability Map and Its Use in Face Detection," *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp. 1034–1037, October 2001, Thessaloniki, Greece.

[47] Z. Xue, D. Shen, and S. Wong, "Tissue Probability Map Constrained CLASSIC for Increased Accuracy and Robustness in Serial Image Segmentation," *Proceedings of the 2009 SPIE Symposium on Medical Imaging*, vol. 7258, pp. 725904–1–9, February 2009, Lake Buena Vista, FL, USA.

[48] C. Poynton, *Digital Video and HDTV Algorithms and Interfaces*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2003.

EuroVis Workshop on Visual Analytics (2014)
M. Pohl and J. Roberts (Editors)

Visual Analytics for Risk-based Decision Making, Long-Term Planning, and Assessment Process

Silvia Oliveros-Torres[†1], Yang Yang[†1], Yun Jang[‡2], David Ebert[†1]

[1]Purdue University, USA, [2]Sejong University, South Korea,

Abstract

Risk-based decision making is a data-driven process used to gather data about outcomes, analyze different scenarios, and deliver informed decisions to mitigate risk. We describe the design and application of integrated visual analytics techniques and components to support risk-based decision making following a structured risk management process in the US Coast Guard domain. The components proposed perform the following interactive tasks: the identification of risk priority areas, the distribution of pre-computed risk values, and the analysis of coverage versus risk, all of which equip analysts with the tools to examine the different decision factors and assist course of action development in the long-term planning and assessment process.

1. Introduction

Risk-based decision making is a growing operational and business trend that currently lacks interactive tools to aid the decision makers. The term risk is defined as the "potential for an unwanted outcome resulting from an incident, event, or occurrence, as determined by its likelihood and the associated consequences" [Com10]. Therefore, risk-based decision making can be defined as a process that collects and organizes information about different possible outcomes in an ordered structure that helps analysts make informed choices [MMGW04]. Risk-based decision making provides a framework for making decisions and helps identify the greatest risk so the decision maker can prioritize efforts in order to minimize risk and support long-term planning.

However, performing risk analysis and long-term planning is a complex and challenging analytical task, in which the decision maker must set up the problem and determine inputs, outputs, and other factors that might influence the decisions. Research in other areas has shown that individuals often make sub-optimal decisions due to cognitive limitations [SLFE11] and information overload [EM08]. Moreover, the analyst could base his/her decisions on subjective, rather than objective, perception of the risk at hand.

Therefore, we have developed several visual analytics components that can facilitate and improve the process of risk-based decision making. These components, developed through a collaborative user-centered process with the U.S. Coast Guard, use graphical depictions to assist the cognitive process of quantifying and comparing lines of evidence [LCG*12]. Our interactive components facilitate thinking, thereby improving the analyst's understanding of the data and speeding the overall decision making process. The components include feedback and exploratory abilities to examine, filter, and modify certain parameters.

During development, we followed a procedure similar to Sedlmair et al.'s [SMM12] nine-stage framework for conducting design studies. The new components were added to the framework described by Malik et al. [MMME11] because the end users have an understanding and working knowledge of the system.

The new risk-based visual analytics components being applied to visualize and compare risk include the following:

- The use of interactive graphics and choropleth maps to visualize operational risk profiles.
- A method to visualize and identify areas of high risk and compare the changes in risk priority areas over time.
- A method to spatially evaluate and distribute precomputed risk values based on the underlying distribution of cases over time.

† e-mail: {solivero|yang260|ebertd}@purdue.edu
‡ e-mail: jangy@sejong.edu

2. Related Work

In this section, we review previous works that describe the use of visual analytics in communicating risk, some existing models for risk analysis, and different tools to address risk in the maritime security domain.

In risk communication, Lipkus and Hollands [LH99] demonstrated that static images displaying risk characteristics such as risk magnitude and cumulative risk communicate the risk values more effectively than a display of numbers. Savikhin et al. [SME08] demonstrate the benefits of applying visual analytics techniques to aid users in their economic decision making. In contrast, our components provide not only visualizations, but also integrated techniques to analyze the changes of risk values both spatially and temporally.

For risk analysis and modeling, Bonafede and Marmo [BM08] demonstrate that the use of graphs can reduce search times for solutions and for identification of data. They propose four sub-plots with bar graphs and parallel coordinates to compare clients. Feather et al. [FCKM06] describe a risk based decision process with a model that takes into account requirements, risks, and mitigation strategies using bar charts and treemaps. Both papers emphasize that no single visualization technique serves all purposes and instead it is better to use a mix of several. One limitation in their systems is the lack of support of spatiotemporal data. Migut and Worring [MW10] developed a framework that integrates interactive visual exploration with machine learning techniques to support the risk assessment and decision making process. Their visualizations include scatterplots and mosaic plots as tools to build classification models.

Willems et al. [WvdWvW09] presented a geographical visualization using density estimated heatmaps to display vessel movements and support coastal surveillance systems. Pelot et al. [PP08] created a grid colored map representing vessel traffic where they model and identify vulnerable areas. Marven et al. [MCK07] analyzed Search and Rescue operations for the Canadian Coast Guard, exploring the clustering of incident areas with two different models: a Spatial and Temporal Analysis of Crime (STAC) and kernel density estimation (KDE). Abi-Zeid et al. [AZF05] developed SARPlan, a geographic decision support system for planning search and rescue missions, originally developed for aeronautical incidents. Orosz et al. [OSB*10] developed PortSec for decision-making and planning of port resources to address security needs to outside threats and hypothetical scenarios.

3. Visual Analytics in the Risk Management Process

We used the risk management process originally specified in ISO 31000:2009 [ISO09] to provide the initial principles and generic guidelines for risk management. Based on this process, we developed specific goals that our new visual analytics components should achieve:

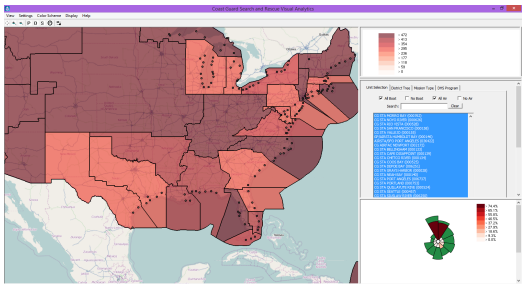


Figure 1: *View of the overall Visual Analytics System*

- Understand areas and missions driving the risk values.
- Identify risk priority areas and how they evolve over time.
- Visualize the geographical distribution of operations.
- Visualize the spatial distribution of the risk.
- Obtain details on demand about the operations.
- Provide a feedback loop if certain parameters change.

Malik et al. [MMME11] focused on the consequences of station closures, but the new additions to the system focus on Risk at the operational level. Such risk is assessed by the USCG Operational Risk Assessment Model (ORAM) [USC12]. Analysts at the Coast Guard Atlantic Area's Operations Analysis Division created this model to support mission planning and analysis of the Coast Guard's mission set. The model combines quantitative and qualitative theoretical frameworks to calculate and compare risk between the eleven Coast Guard statutory missions and geographical areas by providing the Risk Index Numbers (RIN) [USC12]. The RIN is a numerical value that characterizes and quantifies the qualities of risk. RIN values provided include both total risk and residual risk values as shown in Equation 1 [Com10].

$$\text{Total RIN} = \text{Residual RIN} + \text{Mitigated RIN} \qquad (1)$$

3.1. Operational Risk Profiles

The first step is to acquire an understanding on how the risk numbers behave for each district as well as how much risk was mitigated. Therefore, there are two main goals in visualizing the Operational Risk Profiles:

- Compare the RIN values between the districts for any given mission or combination of missions.
- Compare the RIN values between missions for any given district.

When performing total versus residual risk analysis, the ratio between the RIN values is more critical than the raw numbers; therefore, we choose a radial layout to focus on ratios and relative values since such layouts inhibit the analysts innate tendency to focus on these numerical details.

We went through several design iterations and presented different alternatives to our end users to gain feedback in terms of which design was the most effective in conveying the information and comparing the distribution of risk. A risk

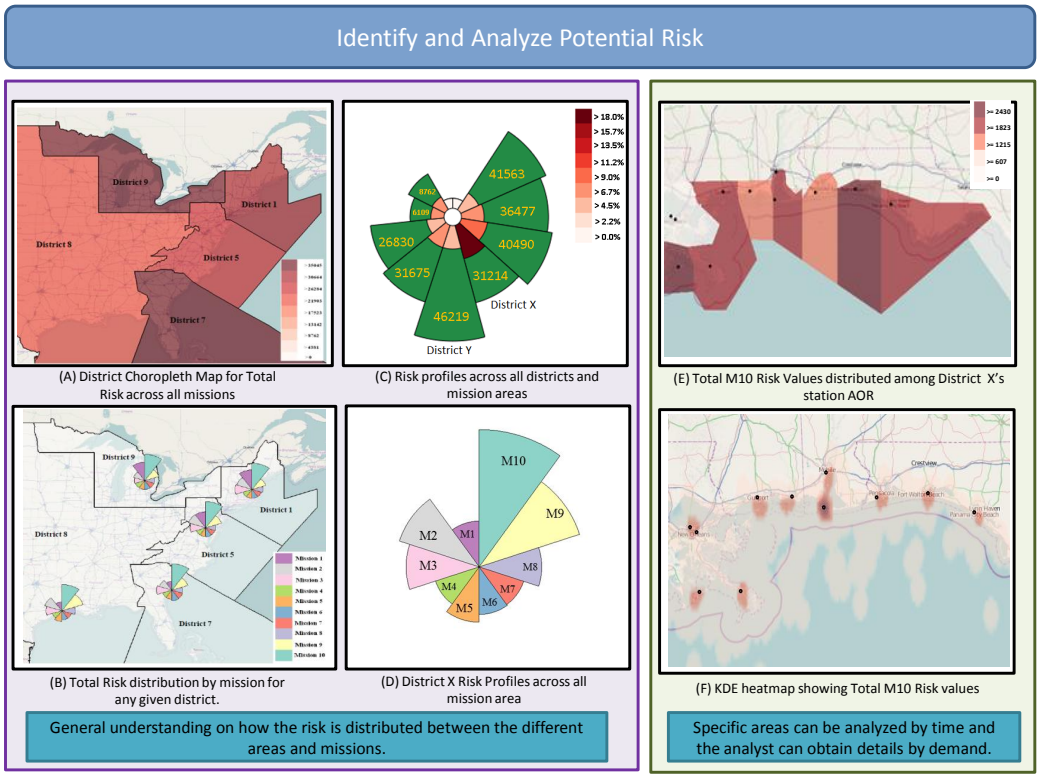


Figure 2: *General process for identifying and analyzing potential risk.*

pie graph was created with eleven fixed pie slices each representing a Coast Guard district as shown in Figure 2-C. The area of each outer pie slice is used to encode the comparison of total risk across districts, with larger pie slice corresponding to higher total risk. The area of inner pie slices represent the comparison of residual risk across districts. Each inner pie slice is also colored on a sequential red scale indicating the ratio of residual versus total risk for a given district. The choice of color (green indicates mitigated risk and red indicates residual risk) is consistent with the Coast Guard's Green-Amber-Red model. We allow interactive filtering by missions to analyze and compare the spatial distribution of risk across districts for any given mission or combination.

3.2. Risk Visualization Using Heatmaps

Next, we need to analyze risk priority areas and how they evolve over time. To quickly identify hotspots, a modified variable kernel density estimation technique (KDE) is employed on the map. Risk at the strategic level is not assigned to a specific unit or station, instead the analyst is able to observe areas with a high density of incidents independent of station location. The heatmap can display the RIN values for total, residual, and mitigated risk. The analyst can switch between the total risk and the residual risk to find hotspots where the risk has not been mitigated and examine the incident details in these zones. Analyzing the incident helps the analyst develop new strategies and courses of action to mitigate the risk.

3.3. Risk Distribution using Choropleth Maps

We utilize choropleth maps in two different ways to help visualize risk. The first option is to visualize any of the Risk values for any given mission or combination of missions by district (Figure 2-A), providing an effective way to present and share the information about risk levels within the U.S.

The second use of choropleth maps (Figure 2-E) highlights the risk distribution of the RIN values per district. During the process it is useful to visualize risk at the station level by using each individual station's Area of Responsibility (AOR). Certain mission's RIN numbers are computed at a district level rather than the station level. Therefore, in order to distribute the RIN values across the stations' AORs, we analyze the underlying incident distributions for a given time period. We use the incidents distribution as a basis to assign risk values across stations given the pre-computed total RIN values by district. The mathematical formula used to compute distributed RIN value for a particular station X that belongs to district Y is:

$$\text{station X RIN} = \frac{\text{Incidents in X}}{\text{Incidents } \forall \text{ stations in Y}} \times \text{district Y RIN} \quad (2)$$

The risk distribution choropleth map provides an easy way to visualize the variations in risk values for individual station's AOR and help identify stations that will potentially require allocation of more resources.

3.4. Visual Analytics System

The overall system provides multiple linked windows and advanced filtering techniques to perform spatio-temporal analysis on the risk data as shown in Figure 1. The system allows the user to visualize historical Coast Guard Data, such as the number and location of incidents that occurred during a certain period of time. It can analyze incidents occurring on specific date ranges to explore seasonal trends and it can filter incidents relevant to the analyst's hypothesis. The addition of the new components enables the Coast Guard analyst to perform risk-based analysis of the operation as well as long term planning by providing new visualization along with feedback loops that control resource allocation.

4. Case Study: Identify and Analyze Potential Risks

To illustrate the use of our system, we present an example use scenario using notional data. In decision making, several questions will drive the analyst in developing the planning strategy: What risks exist in the region and where they are distributed? Where are our resources allocated? What constraints exist in the system that will require a prioritization of resource use?

In a resource constrained environment, we want to use resources in the mission area that provides the greatest return on investment (large amount of total risk but very little residual risk). The first step in the risk management process is to identify potential risks; therefore the analyst begins by looking at the operational risk profile and the district risk choropleth map to observe the risk values at the district level across all mission areas.

Figure 2-C displays the total and residual risk and the ratio between them for all the districts across all mission areas. In this case, we can observe that although District Y has the largest total risk values, it mitigated most of the risk effectively. On the other hand, District X shows less total risk, but the amount of residual risk as well as residual to total risk ratio is the highest as encoded by the darkest red shade. District X can be seen as more problematic than District Y; thus, the analyst will focus more attention on analyzing this particular district. This visualization provides a starting point in understanding how risk is distributed among the different districts and focusing on districts with high risk concentration.

After identifying that District X has the greatest residual risk and the highest risk concentration, the next step is determining the key drivers of risk within a district. This leads the analyst to leverage other components of the risk visual analytics tool to specifically evaluate District X. For instance, the analyst can examine the distribution of risk across different missions in District X as shown in Figure 2-D to identify which mission type has the greatest risk in this district. The analyst can observe that most of the operational risk emerges from one of the missions, in this case M10.

New questions emerge at this stage: Are there several big events that drive the risk, or are there many small events with smaller consequences accumulated to affect the operation? So now we examine the spatial distribution of M10 risk within District X to analyze specific areas of high residual risk. Depending on the data quality regarding spatial location, the analyst has two options for drilling down into specific areas within District X. The first option is to use the risk heatmap described in Section 3.2 to locate risk priority areas, as seen in Figure 2-F. If the spatial location is not available, then we re-distribute the risk to station AORs as described in Section 3.3 and as seen in Figure 2-E.

5. Domain Expert Feedback

The prototype components went through an iterative design refinement process with the collaboration of four Coast Guard personnel: an operation research analyst, a former Coast Guard officer, one in-field officer, and a high level officer. Informal feedback is given below:

"These components aid the analyst in answering the questions that come from developing the planning strategy, often with a speed that was previously unattainable with the Coast Guard's usual brute force processing of thousands of lines of data to calculate summary statistics."

"This system provides a risk informed process for building a defensible planning baseline for the long-term planning process. Understanding the risk profiles provides analytic justification for resource use, and can aid in demonstrating effective application of resource use based on risk."

6. Conclusions

We have demonstrated how our interactive visual analytics components can facilitate the risk management process and evaluate courses of action. Within the maritime context, our interactive visual analytics environment utilizes KDE heatmaps to help identify risk priority areas, multiple designs to visualize risk profiles, a risk distribution choropleth map to visualize the spatial distribution of pre-computed risk values, and the coverage map overlaid with risk distribution for analysis of coverage capability/efficiency as well as potential need for resource reallocation or assets upgrade. Finally, we included a case study that examines the efficiency of Coast Guard operations and provides useful visual reference that can communicate recommendations based on risk management. The described risk-based decision making process serves as a blueprint for future systems dealing with risk values and resource planning.

Acknowledgment

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003. Jang's work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170).

References

[AZF05] ABI-ZEID I., FROST J. R.: Sarplan: A decision support system for canadian search and rescue operations. *European Journal of Operational Research 162*, 3 (2005), 630 – 653. Decision-Aid to Improve Organizational Performance. 2

[BM08] BONAFEDE C., MARMO R.: Operational Risk Visualization. *Science* (2008), 100–103. 2

[Com10] COMMITTEE R. S.: *DHS Risk Lexicon*. Homeland Security, 2010. 1, 2

[EM08] EPPLER M., MENGIS J.: The concept of information overload - a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004). In *Kommunikationsmanagement im Wandel*, Meckel M., Schmid B., (Eds.). Gabler, 2008, pp. 271–305. 1

[FCKM06] FEATHER M., CORNFORD S., KIPER J., MENZIES T.: Experiences using Visualization Techniques to Present Requirements, Risks to Them, and Options for Risk Mitigation. *2006 First International Workshop on Requirements Engineering Visualization (REV'06 - RE'06 Workshop)* (Aug. 2006), 10–10. 2

[ISO09] ISO 31000, Risk Management Principles and Guidelines, Geneva : International Standards Organisation, 2009. 2

[LCG*12] LINKOV I., CORMIER S., GOLD J., SATTERSTROM F. K., BRIDGES T.: Using our brains to develop better policy. *Risk Analysis 32*, 3 (2012), 374–380. 1

[LH99] LIPKUS I. M., HOLLANDS J. G.: The visual communication of risk. *Journal of the National Cancer Institute. Monographs 27701*, 25 (Jan. 1999), 149–63. 2

[MCK07] MARVEN C., CANESSA R., KELLER P.: Exploratory spatial data analysis to support maritime search and rescue planning. *Geomatics Solutions for Disaster Management* (2007), 271–288. 2

[MMGW04] MACESKER B., MYERS J., GUTHRIE V., WALKER D.: *Quick Reference Guide to Risk Based Decision Making (RBDM): A Step by Step Example of the RBDM Process in the Field*. EQE International, Inc., an ABS Group Company Knoxville, Tennessee, 2004. 1

[MMME11] MALIK A., MACIEJEWSKI R., MAULE B., EBERT D.: A visual analytics process for maritime resource allocation and risk assessment. In *Visual Analytics Science and Technology (VAST), IEEE Conference on* (oct. 2011), pp. 221 –230. 1, 2

[MW10] MIGUT M., WORRING M.: Visual exploration of classification models for risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (October 2010), pp. 11–18. 2

[OSB*10] OROSZ M., SOUTHWELL C., BARRETT A., CHEN J., IOANNOU P., ABADI A., MAYA I.: Portsec: A port security risk analysis and resource allocation system. In *Technologies for Homeland Security (HST), 2010 IEEE International Conference on* (Nov. 2010), pp. 264 –269. 2

[PP08] PELOT R., PLUMMER L.: Spatial analysis of traffic and risks in the coastal zone. *Journal of Coastal Conservation 11* (2008), 201–207. 2

[SLFE11] SAVIKHIN A., LAM H., FISHER B., EBERT D.: An experimental study of financial portfolio selection with visual analytics for decision support. In *System Sciences (HICSS), 44th Hawaii International Conference on* (2011), IEEE, pp. 1–10. 1

[SME08] SAVIKHIN A., MACIEJEWSKI R., EBERT D.: Applied visual analytics for economic decision-making. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (October 2008), pp. 107–114. 2

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 18*, 12 (2012), 2431–2440. 1

[USC12] USCG: *Joint, CG Atlantic Area and CG Pacific Area Operational Risk Assessment Model (ORAM), Executive Summary*. Unpublished Technical Document, 2012. 2

[WvdWvW09] WILLEMS N., VAN DE WETERING H., VAN WIJK J. J.: Visualization of vessel movements. *Comput. Graph. Forum 28*, 3 (2009), 959–966. 2

Analyzing High-dimensional Multivariate Network Links with Integrated Anomaly Detection, Highlighting and Exploration

Sungahnn Ko, Shehzad Afzal, Simon Walton, Yang Yang, Junghoon Chae, Abish Malik,
Yun Jang, Min Chen and David Ebert, *Fellow, IEEE*

Abstract—This paper focuses on the integration of a family of visual analytics techniques for analyzing high-dimensional, multivariate network data that features spatial and temporal information, network connections, and a variety of other categorical and numerical data types. Such data types are commonly encountered in transportation, shipping, and logistics industries. Due to the scale and complexity of the data, it is essential to integrate techniques for data analysis, visualization, and exploration. We present new visual representations, *Petal* and *Thread*, to effectively present many-to-many network data including multi-attribute vectors. In addition, we deploy an information-theoretic model for anomaly detection across varying dimensions, displaying highlighted anomalies in a visually consistent manner, as well as supporting a managed process of exploration. Lastly, we evaluate the proposed methodology through data exploration and an empirical study.

1 Introduction

The recent trend of increasing size, complexity, and variety in datasets (e.g., spatial, temporal, quantitative, qualitative, network data) makes analysis and decisions from these data more challenging, often called the *big data* problem [21, 32, 37]. One very challenging type of big data is multivariate network data, especially when there are multivariate values for both nodes and links. For example, transportation, shipping, logistics, commerce, trading, electricity and communication industries [7, 41] have many connected operational locations where multiple variables describe each location's operations. With flight delay network data, various multivariate operational aspects are considered simultaneously: types of delay, patterns based on airport location, trends in time, and relationships among the airports. To reduce the analysts' information overload and to enable effective planning, analysis and decision making, an interactive visual exploration and analysis environment is needed as traditional machine learning and big data analytics alone are insufficient [10].

While, various systems and techniques for network visualization have been proposed [19], few support analyzing both multivariate network data (e.g., [39] and [25]) and map-based spatial network data (e.g., [17] and [7]). There still remains a gap in effective multivariate spatial network data exploration and analysis to efficiently answer challenging questions such as the following: What are the patterns in multivariate variables on a node or among node-node pairs? Are the patterns relevant to specific regions and times? Is there any seasonality in the patterns? Can we verify the patterns on a map? Which network nodes and links could be anomalous?

In this work, we fill this gap by integrating a family of visual analytics techniques for exploring and analyzing such complex data. We employ multiple linked views [30] (see Fig. 1), two new multivariate visualization techniques, *petals* and *threads*, and an information-theoretic analytical backend engine for aggregate-level and detail-level network analysis.

Petals and *threads* efficiently present a simplified representation of many-to-many networks where multi-attribute vectors represent the size of attributes in different directions. Specifically, *petals* represent an aggregated summary view of directional data (Fig. 3) and *threads* encode multiple variables of links (Fig. 2). An information-theoretic model provides our analytical engine the ability to highlight anomalies in the data. The anomaly detection can be dynamically configured based on new contextual requirements that usually result from user-generated hypotheses stimulated from visualization and exploration of data. The analytical method provides visualization with additional warning signals and enables users to prioritize their exploration strategy.

The contributions of our work in the multivariate spatiotemporal network visualization and analysis domain are 1) designing *petals* and *threads* for high-dimensional multivariate network link analysis, 2) evaluating *petals* and *threads* with a user study, 3) designing and implementing a visual analytics system using multiple coordinated views, 4) integrating an information-theoretic anomaly detection method in the interactive visualization analysis process, and 5) exploring complex data (e.g., flight delay network) to illustrate the use and potential of our designs in the multiple-coordinated views.

Our system can be applied to exploration of any multivariate spatiotemporal, network link data such as transportation, shipping, logistics, commerce, trading, communication industries (e.g., AT&T communication network data [7] and electric power grid data [41]).

2 Related Work

While the research topics in network visualization are as numerous as the visualizations themselves [19], in this work, we consider network visualization techniques and tools that are pertinent to multivariate geospatial network data. For multivariate network visualization research, Wattenberg [39] has designed PivotGraph, a software tool focusing on the relationships between node attributes and connections of multivariate graphs on a grid layout. Ploceus [25] enables multi-dimensional and multi-level network-based visual analysis on tabular data while Honeycomb [38] focuses on scalability (e.g., millions of connections) using a matrix representation that is also incorporated in our pixel matrix view. For geospatial network visualization, Guo [17] has developed an integrated, interactive visualization framework that visualizes major flow structures and multivariate relations at the same time. SeeNet [7] visualizes geospatial network data in a communication industry; however, its visualization focuses on univariate data. In contrast to the previous work, our system allows users to analyze all combinations of spatial, temporal, multivariate, and network characteristics simultaneously. Herman et al. [19] surveyed other network visualization techniques beyond our paper's scope.

In order to visualize multivariate data, and to display the maximum amount of data relative to the available screen space, a pixel-based visualization was developed by Keim et al. [20]. In the pixel-based visualization, each data element is assigned to a pixel, and a predefined color map is used to shade the pixel to represent the range of

- *Sungahn Ko, Shehzad Afzal, Yang Yang, Junghoon Chae, Abish Malik and David S. Ebert are with Purdue University in West Lafayette, IN, USA. E-mail: {ko, safzal, yang260, jchae, amalik, ebertd}@purdue.edu.*
- *Yun Jang is with Sejong University in Seoul, Korea. E-mail: {jangy}@sejong.edu.*
- *Simon Walton and Min Chen are with Oxford University in Oxford, UK. E-mail: {simon.walton, min.chen}@oerc.ox.ac.uk.*

Submitted to IEEE VAST 2014. Do not redistribute.

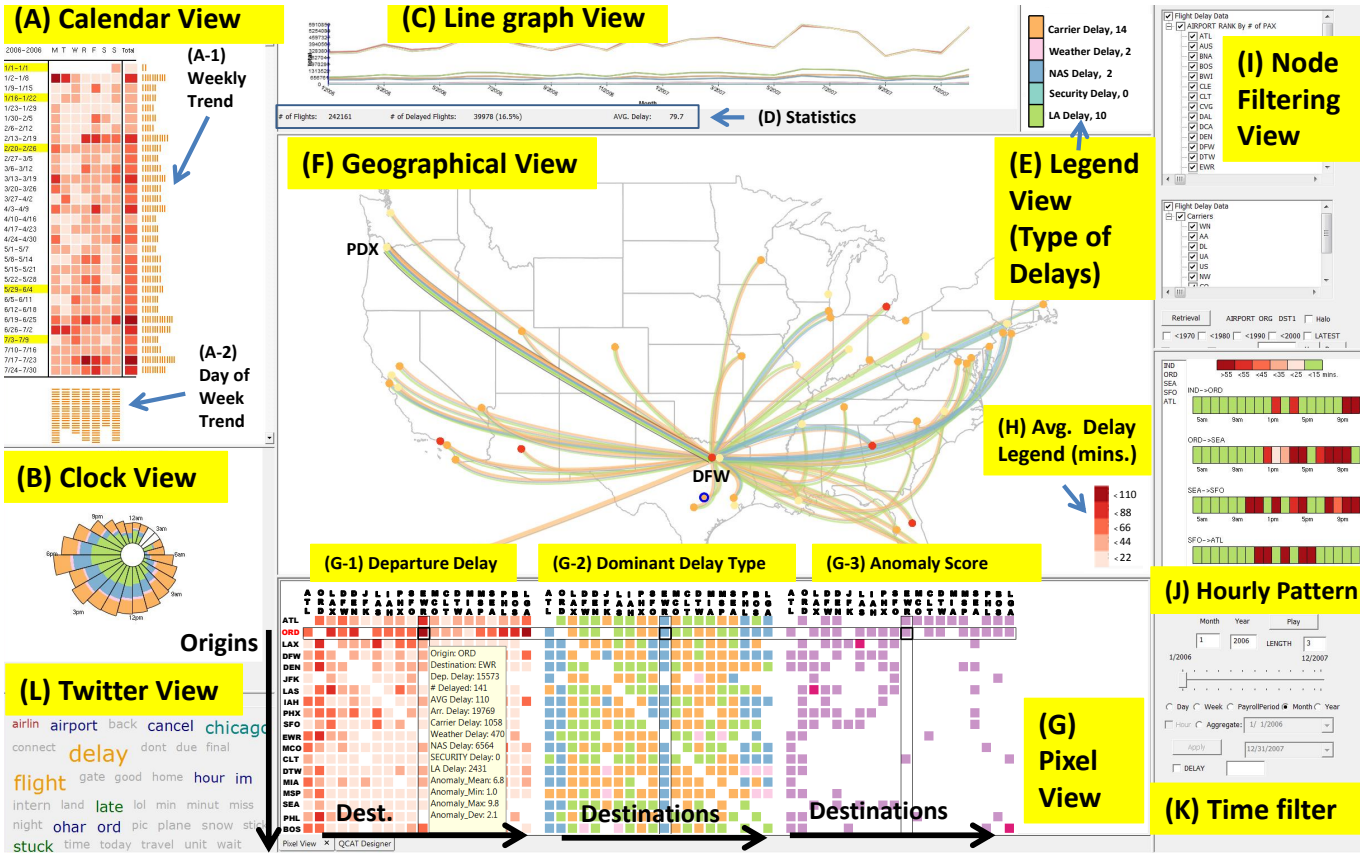


Fig. 1. Our system consists of multiple coordinated and linked views: (A) Calendar view, (B) Clock view, (C) Line graph view, (D) Statistics, (E) Legend view for displaying types of delays, (F) Geographical view, (G) Pixel view, (H) Legend view for delay type and time, (I) Node filter, (J) Pattern on itinerary view, (K) Time and aggregation filter, and (L) Twitter tag cloud view. In the (H) legend, the darker the red, the longer the average delay is. A route from Dallas (DFW) to Portland (PDX) is specified in (F), and the top 20 airports in terms of delays are visualized in (G) for explanation.

the data attribute. Thus, the amount of information in the visualization is theoretically limited only by the resolution of the screen. Borgo et al. [8] present how the usability of the pixel-based visualization varies across different tasks and block resolutions. Oelke et al. [29] study visual boosting techniques for pixel-based visualization such as halos and distortion. Ziegler et al. [43] present how pixel-based visualizations help analysts gain insight for long-term investments while Ko et al. [22] demonstrate the effectiveness of pixel-based visualization in analyzing corporate competitive advantages. Our system incorporates this pixel-based visualization not only to visualize as much data as possible, but also to describe origin-destination network status because the conventional layout of placing pixels side-by-side naturally builds a node-link network.

To help users visually explore multivariate data, many systems have been developed in research and commercial areas [42] (e.g., Spotfire [5], QlikView [3], and Tableau [4]). Common among these systems is that they make extensive use of interactive techniques for brushing, linking, zooming, and filtering to refine the user's queries. Of the systems, Tableau [4], which has become popular due to its flexible operation, allows analysts to easily access and effectively analyze their data [42]. Although multivariate and time-series data analysis is possible in the tool, comparison among multivariate, spatial-temporal, and network-based attributes with geographical components is not well supported by Tableau. In our system, all attributes and characteristics in the data are incorporated and visualized using multiple linked views for simultaneous comparisons. For visualizing multivariate data, Duffy et al. [13] use a glyph encoding some 20 variables while Scheepens et al. [33] focus on a method for reducing visual clutter and occlusion among glyphs.

Analysis of spatiotemporal social media data can have a significant impact to increase situational awareness and provide insights for investigations. Sakaki et al. [31] introduce a natural disaster alert system for earthquake epicenter estimation by analyzing Twitter messages as virtual sensors. Scatterblogs [9,36] is a scalable system enabling analysts to find quantitative information and to detect spatiotemporal anomalies within a large volume of geo-located microblog messages. Chae et al. [11] propose an interactive social media data analysis and visualization system for anomalous circumstance detections as well as examinations of abnormal topics and events from various social media data sources. In our user study, we use Twitter tags to help analysts find any correlation between data and public reactions.

Lee and Ziang [24] provide an overview of using information-theoretical measures for anomaly detection, including entropy, conditional entropy, information gain, and information cost. A number of case studies are also provided in the domain of network security. Chandola et al. provided a comprehensive survey on methods for anomaly detection [12]. Arackaparambil et al. [6] use information theory to monitor network streams for anomalies in network traffic, and to explore the challenges of providing a scalable implementation using a distributed approach to computing entropy and conditional entropy. Kopylova et al. [23] investigate the use of mutual information in network traffic anomaly detection using Rényi entropy rather than the traditional Shannon entropy measure.

3 Multivariate Network Visualization

To effectively reveal as many aspects of the data characteristics as possible, we explore the data in a series of linked visualizations. Fig. 1illustrates how our system provides a comprehensive multivariate network information in multiple linked views. For illustration, we use a flight delay network dataset [1] as an example of multivariate geospatial network data, but any multivariate network data can be populated into our system. Multivariate network information is provided in the

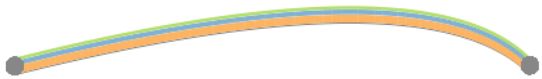


Fig. 2. *Thread* example, showing a single link with multiple *threads*. Width of each *thread* within a link is adjusted based on the contribution of each variable. Contribution of each variable in this example is as follows: Variable 1 (Orange) = 0.5, Variable 2 (Blue) = 0.3, Variable 3 (Green) = 0.2.

geographical view (F) where any operational variable can be used for coloring the node (e.g., anomaly score). The user can explore the data in either a pixel view or a parallel coordinate view (G). Note that (G) has two tab views at the bottom, and a parallel coordinate view example in (G) is shown in Fig. 8. Similarly, time-varying variables (e.g., delays) are presented in different linked visualizations for efficient exploration in the calendar view (A), clock view (B), and line graph view (C). The Twitter view (L) presents important tags with different font sizes and colors. Hourly (delay) patterns can be found in (J) on a series of linked nodes. For example, when a user plans to travel in a certain order (e.g., IND, ORD, SEA, SFO, and ATL), the user can easily find when severe delays are caused in each origin–destination pair. In the system, the line graph view presents temporally aggregated data (e.g., weekly, monthly, yearly). The parallel coordinate view (discussed later in Section 5.2) can be be used to explore the attributes and their value distributions, as well as designing and selecting Query Conditional ATtributes(QCATs, discussed in Section 4) for anomaly detection. Based on characteristics of the data, perceptually appropriate color maps are chosen from both sequential and qualitative color maps from ColorBrewer [18].

3.1 Spatial Multivariate Network Visualization

Unfortunately, a barrier exists in analyzing multivariate network data because visual clutter and complexity often occur in visualizing multiple variables for a node with multiple links between nodes in the map. To reduce such clutter and complexity in the analysis, we design *threads* (see Fig. 2) and *petals* (see Fig. 3) for exploring multivariate link network data. *Threads* connect an origin to each destination and visualize multiple link variables. Because visual clutter around the origin is often generated by link visualization and our *threads*, we also design *petals* to present aggregated and simplified many-to-many network link data. *Threads* and *petals* are designed based on the following requirements for the visualization:

R1 A visualization should present multiple variables describing the relationships between an origin and multiple destination nodes on the map. Here, users should be able to see an overview of the multivariate relationships and discern at least the largest variable in the visualization for both one-to-one and one-to-many relationships.

R2 The visualization should provide simplified one-to-many multivariate spatial networks with minimum visual clutter. Use of node rearrangement techniques (e.g., force-based model algorithm [28]) is not allowed to maintain geospatial semantic meanings.

R3 Users should be able to discern in the visualization for R2 which one-to-many network has the largest aggregate value and which variable has the largest contribution for the largest aggregate value of the one-to-many network.

R4 Multiple variables describing the statistics for a node should be visually presented.

For goal R1, we design *threads*, and for goals R2–R4 we design *petals*. In the following sections, we explain their visual representations in detail.

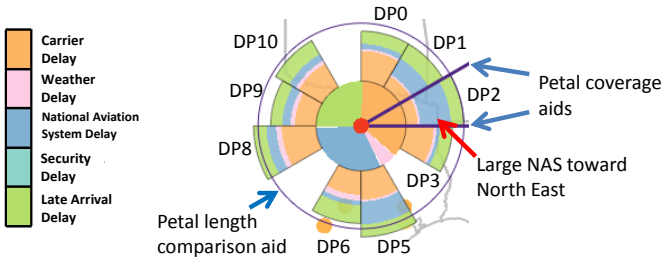


Fig. 3. To show the *petal* coverage including destinations, *petal* coverage guide lines are provided. For comparison of *petal* lengths, equal-radius circles are drawn on all *petals* as shown. The radius of the circles is the length of the *petal* where a user's mouse is hovering.

3.1.1 Thread Visual Representation

We design the *thread* visualization for representing multiple link variables with a focus on the relationship in an origin-destination pair (R1). Each network link consists of multiple *threads*, and each *thread*'s width is scaled based on a link variable's value. Therefore, each link has the same number of *threads* as the number of link variables, but with varying *thread* widths. While GreenGrid [41] utilizes the force-directed layout [28] and presents a (combined) variable on its links, the *threads* are placed on physical locations and present multiple variables. Users can choose the node variables to be encoded in the *thread* link width. Fig. 2 illustrates an example presenting how link variables can be mapped to *threads*. In this work, we use the departure delay times for each cause of delay as the link variables. This visual representation helps users easily identify which link has the largest delay and which delay type contributes most to the delay. In addition, when a link is specified as an anomalous link, it is located on the top in the stack of *threads* and other links become transparent so that the anomalous link can be highlighted as shown in Fig. 1 (F). Note that Bezier curves are utilized for the link visualizations, and *threads* can be sorted (e.g., departure delays or anomaly scores in our implementation). To help user perception, our system provides zooming (with a mouse wheel) and allows users to select the *thread* base width.

3.1.2 Petal Visualization

We introduce *petals*, a new directionally-aggregated radial visual representation as shown in Fig. 3 (Dallas, TX). In this representation, we can provide aggregated directional multivariate network link visualization with minimal visual clutter because we avoid link crossings [7]. Moreover, the spatial and multivariate characteristics are preserved and emphasized. Each directional *petal* (DP) encodes various information between one origin and multiple destinations in a given aggregate direction. Many transportation and logistics problems do have variable variation that is directionally dependent due to transportation paths, weather, routing, etc. By radiating from the origin location to multiple directions (one- to-many), a *petal* presents the geospatial relationships (R2). The *petal* length encodes a selected variable value (R2). Additional variable information is then encoded as radial sections within each *petal* (R3). For example, with the flight delay network data, the average departure delay for the flights heading for airports in a certain radial direction is mapped to the length of the *petal*. Then, the five types of delays are encoded inside the *petal* presenting the contributions of each delay type. Thus, we interpret that DP2 in Fig. 3, has a large NAS (National Aviation System, pointed by an red arrow) delay from Dallas. This indicates a large air traffic delay for the destinations, especially toward the airports in New York. Within a *petal*, we insert a pie chart visualization to show comprehensive overviews and comparisons among multiple variables in a node (R4). In the system, users can turn the *petal* display on and off. By default, we assign 12 *petals* for each origin but the users can merge two adjacent *petals* or split one *petal* into many *petals*. To help users easily recognize the destinations included in a *petal*, our system provides *petal* coverage guide lines as shown in Fig. 3. In addition, when the mouse hovers on a *petal*, the destinations included in the *petal* turn red for better recog-

nition. Lastly to ease comparison of *petal* lengths, equal-radius circles are drawn on all *petals*. The radius of the circle is the length of the *petal* where a user's mouse is hovering (e.g., the radius of the current circle in Fig. 3 is the length of DP2). Note that the data for the destinations within a *petal*'s coverage are aggregated and visualized together in the corresponding *petal*.

3.2 Pixel-oriented Network Matrix

Pixel-based visualization [20] is a technique for visualizing multivariate data but it can also be used for presenting cross-variable data (e.g., origin-destination pair data) by placing pixels side-by-side. We utilize this flexibility in our system to provide more complete multivariate network information, as shown in Fig. 1 *(G)*. Our system allows up to three square pixel matrices, where the *y*-axis of all matrices are the origins while the *x*-axis represent destinations. For example, for a flight delay network data for 20 airports, we place the departure delay matrix in *(G-1)*, the dominant delay type matrix (e.g., weather, security) in *(G-2)*, and the anomaly Z-score matrix *(G-3)* from our information-theoretic model as discussed in Section 4. Note that a Z-score filter is applied so that red pixels have Z-scores larger than 2 (97.7%) and purple pixels have Z-scores between 1 and 2 (84.1%). In our implementation, users can optionally make *G-3* present additional delay information (e.g., delay by airplane ages and by airlines as shown in Fig. 7 (c) and (d)). When a mouse hovers on a pixel, a tooltip pops up to display detailed information including delays of different types, the number of flights, and the anomaly scores, as shown in Fig. 1 (G-1). This interaction method is useful when a user wants to find out whether a delay type presented as a dominant type in (G-2) is indeed dominant among all delay types.

3.3 Time Series Displays

In order to present temporal trends, our system provides various time-series views: a calendar view *(A)*, clock view *(B)*, and line graph view *(C)* in Fig. 1. With the calendar representation [40] that applies a calendar metaphor to effectively reveal seasonality and cyclic trends, our system presents the delays by using different shading levels. For instance, the longer delays are presented with darker red. In addition, to help users identify any holiday effect, the week including a holiday has a yellow background. In order to supplement the functionality of the monthly trend line graph, our calendar representation provides additional weekly information on the right side of the calendar *(A-1)* and day of weekly patterns at the bottom of the calendar *(A-2)* in Fig. 1. The clock representation *(B)* is an efficient tool to detect daily trends [15], and we encode variables using areas to enhance visual perception according to Stevens' power law [35]. The line graph view *(C)* presents the types of aggregated delays as well as statistics such as the number of total flights, delayed flights, and average delay time.

3.4 Twitter Information Display

Analysis of Twitter messages (Tweets) generated within operational locations increases situational awareness and provides further insights for investigations by understanding responses from people and analyzing relationships between the responses and the data [11, 31, 36]. In this case, Twitter messages can be a good reference. In order to enable such analysis, we have set up a tweet collection system where our framework can retrieve tweets (posted after November 2011). In the collecting process, the text content of the Tweets is tokenized into words that are stemmed before the queries. Based on this infrastructure, our system provides Twitter view as shown in Fig. 1 (L) that can assist a user in examining the responses from people which were triggered by their delayed flights, and in finding additional information and correlation from the extracted related key tags. Once the user clicks one of the tags in the view, the dates of Tweets containing the tag are highlighted with blue outlines in the calendar view as shown in Fig. 6. We use opacity to encode the word frequency. Moreover, if the user selects a day in the calendar view, the actual Tweets of the day are displayed in the Tweet tab for further analysis.

4 Anomaly Detection and Highlighting

The visualizations in our system are able to draw upon an information-theoretic model for anomaly detection in a context-sensitive manner, utilizing the anomaly data for a consistent highlighting strategy shown throughout the visualization pipeline. For example, while Fig. 1 *(G-3)* explicitly encodes the anomaly score as the primary visual attribute, Fig. 1 *(F)* focuses on highly anomalous routes with thin outlines. In this case, attribute $a_{origin} = DFW$ (Dallas) is set as the condition in the model. What defines an 'anomalous' record depends upon the user's design and definition of individual anomaly detectors, *QCATs*, discussed in detail in this section. From a visual analytical perspective, these QCATs provide an overview of records where important attributes deviate from usual for specific conditions.

4.1 Overview of Anomaly Detection Method

Chandola et al. provided a comprehensive survey on methods for anomaly detection [12], categorizing them based on the nature of inputs, instance types, algorithmic mechanisms, and forms of outputs. For multivariate network data, we are interested in methods that can:

- Handle multi-dimensional records – because the main flight data concerned is a structured data stream consisting of 29 attribute dimensions (e.g., ≥ 10);
- Address the need for detecting contextual anomalies – which can provide a high-degree of flexibility and accommodating dynamic data and task variations in different detection scenarios;
- Facilitate an unsupervised algorithmic mechanism – alleviating the lack of training data in many situations;
- Generate anomaly scores as outputs that can be effectively conveyed by most visualization techniques.

In general, the family of statistical and information-theoretic methods can address the above-mentioned requirements better than the families of classification-, nearest neighbor- and clustering-based methods. As information theory is fundamentally built on probabilistic and statistical measures, information-theoretic methods may also be considered as a subset of the family of statistical methods. In this work, we use an information-theoretic method because of advantages as highlighted in [12]. "(1) They can operate in an unsupervised setting. (2) They do not make any assumptions about the underlying statistical distribution for the data."

Let $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ be a set of n variables. Each data record, $R = \{v_1, v_2, \ldots, v_n\}$ be a n-tuple, where v_i represents a valid value of attribute $\mathbf{a}_i$. In a practical scenario, an attribute, $\mathbf{a}_i$, may have a very large or infinite number of valid values. Binning is normally used to facilitate more accurate estimation of the probability of each valid value. In the following discussion, the probability distribution of an attribute, $p(\mathbf{a}_i)$, is assumed to be estimated in conjunction with an appropriate binning scheme.

The attribute set, $\mathbf{A}$, is divided into three mutually-exclusive subsets, $\mathbf{A}_{cnd}$, $\mathbf{A}_{von}$, and $\mathbf{A}_{ins}$. As anomalies are context-sensitive, $\mathbf{A}_{cnd}$ defines the context of a type of anomaly as a particular condition, such that all attributes in $\mathbf{A}_{cnd}$ are associated with specific values. For example, we may have $\mathbf{a}_4 = 1(Monday)$, $\mathbf{a}_{17} = JFK$, $\mathbf{a}_{18} = LHR$. The attributes in $\mathbf{A}_{cnd}$ are referred to as *conditional attributes*. In some situations, a conditional attribute may also take a range of values, e.g., $\mathbf{a}_4 = 1, 2, 3, 4$ or 5 (Monday–Friday).

The attributes in $\mathbf{A}_{von}$ play the primary role in determining an anomaly score for each record that has met the condition defined by $\mathbf{A}_{cnd}$. These attributes are referred to as *Variants of Normality* (VON). The remaining attributes, which are grouped into $\mathbf{A}_{ins}$, are considered to have "insignificant" influence on the type of anomaly concerned and are therefore excluded in the computation. Such a decision is usually made based on some known factors or logical reasoning by the user.

A combined configuration of $\mathbf{A}_{cnd}$ and $\mathbf{A}_{von}$ in relation to the overall attribute set $\mathbf{A}$, subsequently, determines how anomaly scores are estimated for each record. Given a record R, we first retrieve all records

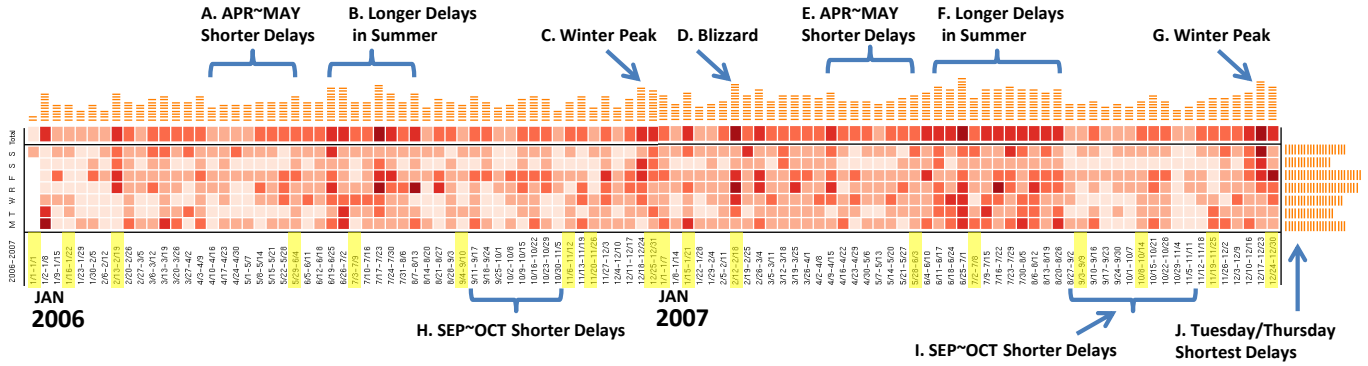


Fig. 4. Calendar view showing delay patterns for 2006–2007. In general, there were long delays in the summer and winter seasons, while APR–MAY and SEP–OCT did not have as many delays. Some delays increased around the holidays, but not all holidays had much impact on the delays.

that have the same conditional attribute values as R. Let this collection of records be $R_1, R_2, \ldots, R_W$, where W is usually a very large number. We now consider only the variants of normality defined by $\mathbf{A}_{von} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_s\}$. In conjunction with a binning scheme, each attribute, $\mathbf{x}_j$, may take valid values that are mapped to a set of t_j bins $B_j = \{b_{j,1}, b_{j,2}, \ldots, b_{j,t_j}\}$. For the s attributes in $\mathbf{A}_{von}$, there are a total of: $t_1 \times t_2 \times \ldots \times t_s$ different combinations of bins across different attributes. These combinations collectively define an alphabet $\mathscr{Z}$, and each unique combination is a letter $z \in \mathscr{Z}$.

The selection of an appropriate binning scheme for each attribute $\mathbf{x}_j$ is essential for ensuring that the total number of letters $|\mathscr{Z}|$ is smaller than the total number of records W. Ideally, we have $|\mathscr{Z}| << W$. We can, then, estimate the probability of each letter $z \in \mathscr{Z}$ based on the collection of records $R_1, R_2, \ldots, R_W$, resulting in a probability distribution function $p(z)$. For the given record R, we obtain its probability $p(R)$ by mapping it to its corresponding letter in $\mathscr{Z}$. The level of self-information is $I(R) = -\log_2(p(R))$, which is also called *surprisal*. We use this surprisal value as the anomaly score for the given record R. The level of uncertainty of this score can be defined as $H(\mathscr{Z})/\log_2(|\mathscr{Z}|)$, where $H(\mathscr{Z})$ is the entropy of the alphabet $\mathscr{Z}$.

It is necessary to emphasize that the anomaly score obtained for R reflects only the type of anomalies encoded by the specific configuration of $\mathbf{A}_{cnd}$ and $\mathbf{A}_{von}$. Hence, each configuration is only for queries of a specific type of anomaly in a particular context. We call each configuration a QCAT (**Q**uery **C**onditional **AT**tributes). It is not difficult to see that a visual analytics system can be equipped with one or more QCATs. For a given record, scores obtained using different QCATs can be aggregated, though it is necessary to understand the semantic implication of combining different QCATs and the difference between different aggregation methods (e.g., mean or max). Section 5.2 discusses the workflow for working with QCATs in a visual analytical system.

4.2 Implementation & Scalability

We have conducted a series of tests on the scalability of QCATs. Two implementations, client- and server-based, have been developed using Postgres. The former performs the grouping and aggregation on the client (i.e, in native code), and the latter uses a stored procedure hosted by the database server. Both server- and client-based implementations show that QCATs are linearly scalable in relation to the number of records used in the computation; the server-based implementation is about 2.5 times faster than the client-based implementation. Additionally, the client implementation is more sensitive to the network bandwidth and latency to the database server.

In our scalability tests, we have found that the performance of the server-based solution can be seriously affected by the number of VONs in $\mathbf{A}_{von}$, while the client-based implementation shows steady linear scalability in relation to the increasing number of VONs. The largest factor is the amount of shared buffers provided to Postgres. The scalability of entropy computation is linear but does rely on recomputing past data due to updated probability masses. However, Arackaparambil et al. [6] show that a distributed method for conditional entropy computation is feasible, while Guba et al. [16] demonstrate entropy estimation in streaming insert-only datasets. In the following sections, we describe how our system presents multivariate network data and visualizes the detected anomalies.

5 Geospatial Multivariate Network Data Exploration

As an example, we will use flight delay data from the Bureau of Transportation Statistics (BTS) [1] where each data row provides information for an individual flight including origin, destination, day of week, day of month, scheduled (departure/arrival) time, and real (departure/arrival) time and type of delay. There are five types of delays. Carrier delay is a problem within the airlines' control including mechanical problems of aircrafts, while NAS delay is caused by the control of the National Aviation System (NAS) including heavy traffic volume. Late Arrival Delay (LAD) is caused by the late arrival of the same aircraft at a previous airport. Security delay includes re-boarding time due to security breach and waiting time at the screening equipment. Weather delay means delay caused by extreme weather conditions at point of departure or arrival. Note that NAS delay and Security delay might be caused by the government organizations, while Carrier delay and LAD are caused by the airlines. We use the top 50 airports according to the number of passenger boardings that encompasses FAA's OEP-35 (Operational Evolution Partnership 35) airports accounting for more than 70% of the entire number of passengers [2].

5.1 Flight Delay Network Exploration

In this section, we explore the flight delay network data from 2006-2007 and summarize delay patterns in terms of temporal (e.g., summer, winter, holidays, weekly, hourly, and day of week) and spatial effects including special conditions such as severe weather (e.g., blizzards). First, we use the calendar view to investigate data patterns. In Fig. 4, we can see long delays as prominent seasonal patterns in the summer (B, F) and winter (C, G), while shorter delays were recorded during April–May and September–October. Another visible pattern is that there were fewer delays on Tuesday and Saturday in (J). We find that the patterns are related to holidays that are concentrated in summer and winter (e.g., Independence day in July, Christmas in December, personal vacations) but long delay patterns are not indicated for Martin Luther King day in January and Labor day in September. Moreover, long delay patterns tend to increase in 2007, especially in the summer (B and F). Also, there is a sudden spike (D) shown with the darkest red that might be another point for investigation.

Next, we can explore the aggregated delays for two years in the pixel view as shown in Fig. 5 (a, b), where we see some interesting patterns. The most prominent pattern is the series of horizontal and vertical dark red pixels (long delays) generated at the Chicago O'Hare

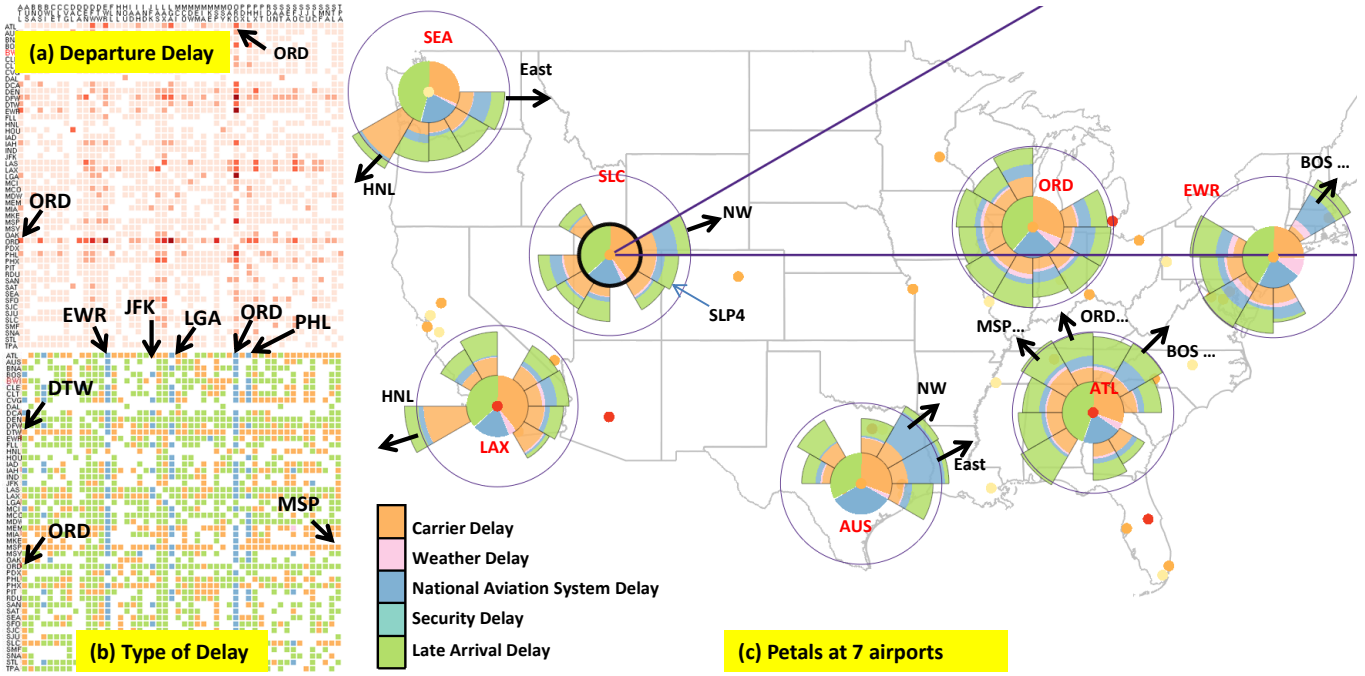


Fig. 5. (a) ORD is the most congested airport for both in-bound (vertical) and out-bound (horizontal). It is notable that carrier delay is the prevalent (out-bound) delay for DTW and MSP while NAS delay is the prominent delay for the incoming flights (vertical) in at EWR, JFK and LGA (b). (c) Flights heading to Hawaii from west cost airports in winter had long delays. Flights heading for ORD, ATL and airports from mid-east and east usually suffer from NAS delays.

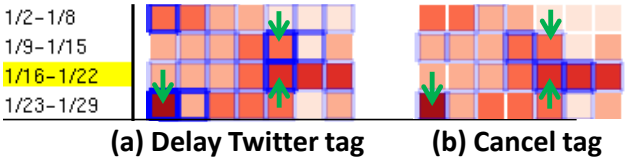


Fig. 6. Possible correlation between Twitter messages and delays. The opaque blue lines represent the day when many users posted messages related to the delay. The tag "delay" is selected in (a) and the tag "cancel" is selected in (b). Those days with arrows had severe weather reports on January 2012.

Airport (ORD) in (a), which indicates that both in-bound (horizontal) and out-bound (vertical) flights were severely congested. We also observe that such delays in ORD were caused mainly by late arrivals of aircraft (green) shown in (b). In addition, we notice that there are five distinguishable vertical blue lines in the matrix (b) and four of them (EWR, JFK, LGA, and ORD) were regulated by the High Density Rule (HDR) enacted in 1969 by the FAA due to severe congestion. This may indicate that the rule might not be strong enough to prevent such long delays. The delays in DTW (Detroit) and MSP (Minneapolis), which are two of the biggest hubs of Delta Airlines, are not very long compared to those in other top congested airports. However, it is interesting that the major type of delay is carrier delay (orange) caused by the airline itself.

Since one of the highest delays is observed in winter as shown in Fig. 4, we use our *petal* visualization with winter seasonal data for finding patterns and types of delays in the network as shown in Fig. 5 (c). We can select as many *petals* as designed for the exploration as long as minimal visual clutter is maintained. One interesting finding is that the flights heading for HNL (Hawaii) from the west coast airports (SEA and LAX) have relatively long delays (e.g., 120 minutes on average) and the prevalent cause for the delay is the carrier delay. Moreover, those airports also have relatively long NAS delays for flights heading for north-east destinations (ORD, and airports around New York).

The next interesting aspect is the delay distribution by time as shown in Fig. 1 (B) in the proportional mode with area encoding for each delay type. Here, we see a trend showing that delays increased from 6 am and had a peak around 6 pm. It is noted that this is the same pattern shown in the late aircraft delay while other types retained their proportion. This suggests that delays propagate during the day, a problem that Mazzeo termed "cascading delays" [26]. Such trends may imply that delays might be effectively reduced because these delays can be controlled either by the airlines (carrier delay/late arrival delay) with enough of an interval or layover time between two consecutive flight schedules, or by a government agency (e.g., Federal Aviation Administration) with advanced systems for air traffic control.

During exploration, we found that ORD (Chicago) generally shows longer delays than others in winter. To find any correlation between delay and Tweets, we use the twitter view with Tweet data generated within 3 miles of ORD in January 2012, as shown in Fig. 1 (L). We can see that several related Twitter tags occur, including as Chicago, airport, flight, delay, cancel, late, and hour. When delay and cancel tags are clicked, we find a possible correlation between the tags and the delays on the calendar (January, 2012) as shown in Fig. 6. In the figure, the opaque blue outlines presenting frequencies of Tweet messages are placed on the dates when large delays were recorded.

Of primary interest are the patterns in the length and types of delays that can be can be better explored by sorting airports. We see that the ranks change with little variation based on seasons, but most delays are caused by some major airports including ORD (Chicago), ATL (Atlanta), LGA (New York City), EWR (New York City), DTW (Detroit), LAX (Los Angeles), LAS (Las Vegas), and DFW (Dallas) as shown in Fig. 7 (a). From the type matrix Fig. 7 (b), we notice that in many highly-ranked airports, the main type of delay is the late arrival delay in busy travel seasons while the NAS delay is dominant at other times. This implies that the NAS might not be properly adapting to the current increasing traffic in terms of delays. On the other hand, we notice that the two distinguishable airlines causing delays are AA (American Airline) and UA (United Airline) in the two most delayed airports as shown in Fig. 7 (c). The dominant delay type matrix in Fig. 7 (b) indicates that the airlines are responsible for solving

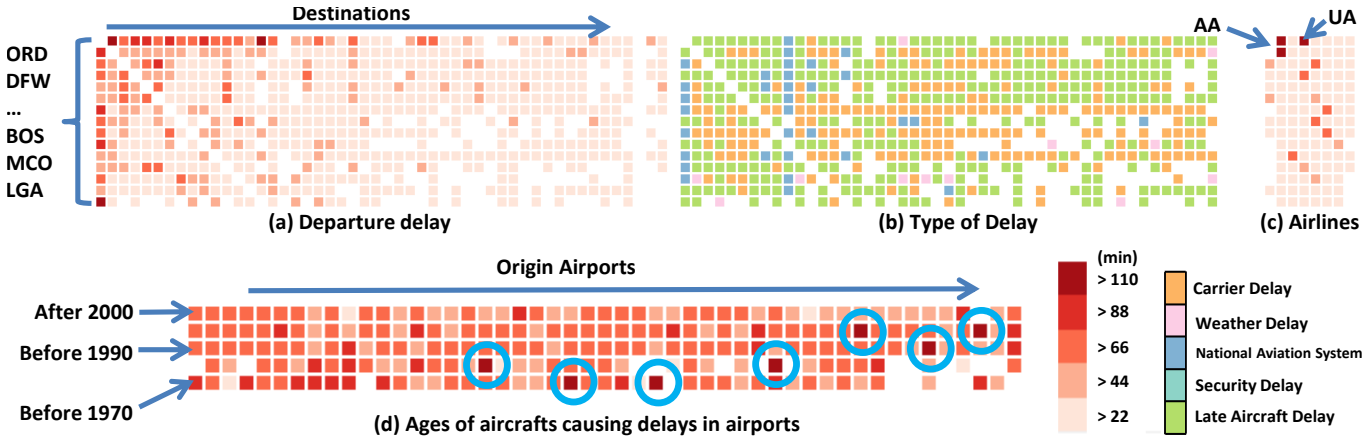


Fig. 7. Airports are sorted by delays. ORD shows the longest delays in many out-bound flights in (a). The dominant type of delay was carrier delay and LAD in (b). UA and AA had the longest delays in ORD when ORD was top by delays in (c). Old airplanes generally caused long delays in (d).

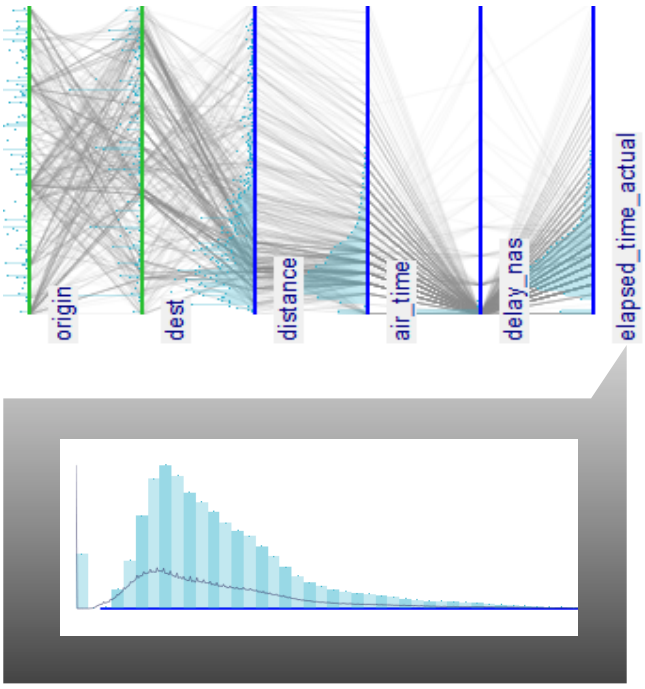
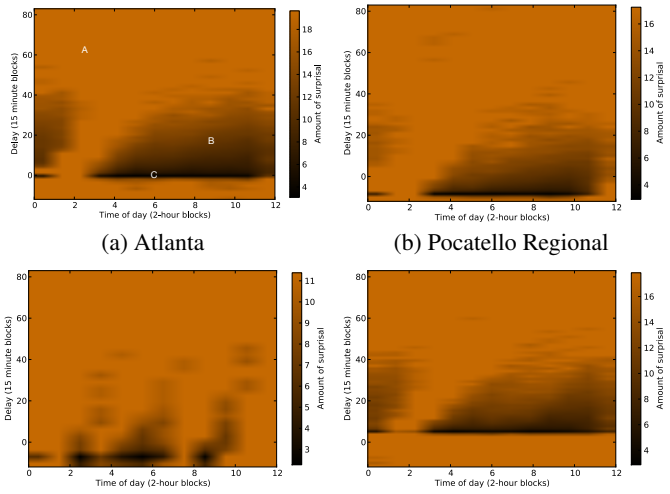


Fig. 8. Using Parallel Coordinates to Design QCATs: *(top)* Exploring the attributes as a parallel coordinate plot; *(bottom)* Specifying an individual attribute's bin specification



(a) Atlanta (b) Pocatello Regional

(c) Sitka (d) John F Kennedy Intl

Fig. 9. Heatmaps representing the surprisal spaces of flights leaving four different airports, with *(x)* Time of day *(bin size: 2 hrs)*, and *(b)* Departure delay *(bin size: 15 mins)*

the delay problem because the dominant types of delays were carrier delay and late aircraft arrivals. Finally, it is also noted that there are old airplanes operating as shown in (d, rotated) where we see that 7 airports have the darkest red colors and only the older airplanes cause the severe delays. It is also notable that airplanes manufactured before 1970 have higher average delays across airports.

5.2 QCAT Workflow

As discussed in Section 4, our system features an information-theoretic anomaly detection system that is comprised of a set of user-defined QCATs. The design of a QCAT can be based on a specific hypothesis, or as a more general monitoring system for one or more attributes. Ideally, in a deployed system, the roles of QCAT *designer* and overall *analyst* would be disparate, with the analyst analyzing the data for anomalies and reporting back to the designer to refine the QCATs based on new trends.

To assist the user in defining the QCATs in the system, we provide a designer tool based on parallel coordinates (see Fig. 8 *(top)*) while the user is able to explore the attribute space by adding/removing attribute dimensions, observing their value distributions (e.g., probability mass functions), as well as viewing the record relationship between attributes afforded by a standard parallel coordinates representation. The role of an attribute can be toggled between conditional (green) and VON (black) using the right mouse button. The user is also able to explore an individual attribute in more detail by clicking the left mouse button that expands the attribute to the full view to show its distribution in more detail (Fig. 8 *(bottom)*). The detail view also shows the attribute's bin width specification, which can be modified per QCAT. The user's choice of bin width has an effect on the anomaly results and reflects the user's knowledge of the attribute's semantic meaning. The system maps data types to suitable bin width granularities automatically. For example, timestamp datatypes are divided into bins of n minutes; categorical data such as strings are unbinned. Since integer types may represent categorical, interval or ratio measurements, we assume a default bin width of 1 and let the user decide upon a more suitable width.

Once the user has defined a QCAT, it can be saved to the QCAT library and selected as the active QCAT. Anomaly-supporting visualiza-

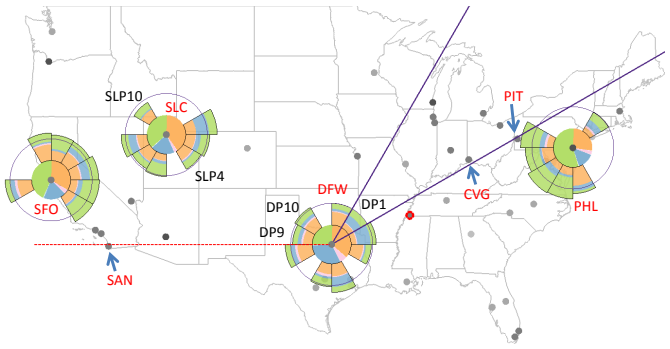


Fig. 10. An example for a *petal* experiment. With the visual aid, users could better tell that CVG is included in DP1 while PIT is included in DP2.

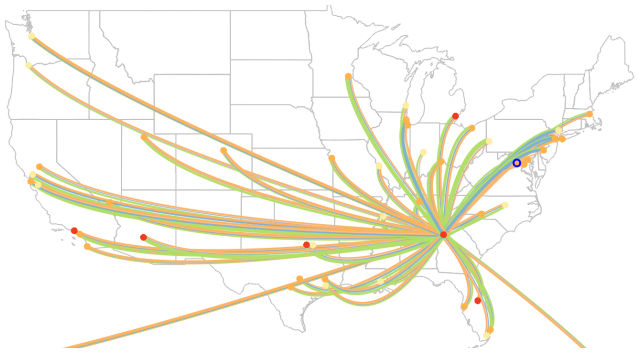


Fig. 11. An example of a *thread* experiment. 40% of the participants answered incorrectly that green was the prevalent delay due to the severe color concentration around the origin.

Petal Index	Difference (%)	Accuracy (%)	Time (s)
DP1	4 (Small)	76.7	8.2
DP10	21 (Large)	100	3.7

Table 1. Participants found the longest delay inside a *petal* more accurately in less time as the difference became larger (HP1).

Petal Index	Difference (%)	Accuracy (%)	Time (s)
DP9	3 (Small)	46.7 (83.3)	6.6 (5.4)
SLP10	12 (Large)	96.7 (100)	3.9 (2.6)

Table 2. As the difference became larger, the participants better detected the shortest delay (HP2). Visual aids improved both the accuracy and efficiency (HP5).

tions in our system such as the pixel-oriented network matrix update to reflect the anomaly scores by completing the relevant conditionals in the QCAT (i.e., origin and destination pairs) and executing the QCAT on the data to obtain statistics (i.e., mean, max, variance) on the surprisal values for records matching the conditionals. Our system by default displays the maximum surprisal value as the anomaly value mapped to a visual attribute (i.e, halo) in the visualization. The anomaly values in the visualizations guide the user to identify abnormal flights based on their own criteria specified in the design of each QCAT. Anomalous results can then be explored further using the available visual analytical tools to understand why the anomaly value was high and report these findings to the QCAT's designer.

For a QCAT consisting of two VONs, we can illustrate the anomaly distribution using a heatmap. Fig. 9 shows the anomaly space for flights leaving four different airports for the years 2006 and 2007. The x-axis shows the time of day divided into two-hour blocks, and the y-axis shows the amount of delay in 15-minute blocks (notice that flights can leave early). Areas of low surprisal value are black and become amber with higher surprisal values. It is clear that for this airport, flights around 4AM are uncommon, and the amount of delay seems to increase steadily throughout the day until late afternoon before leveling out. For the Atlanta airport, three example records, A, B and C are shown of high ($\approx$19.68), slightly above average ($\approx$14.36), and low ($\approx$3.478) surprisal values, respectively. Investigating these flights using *threads* shows that late aircraft were largely to blame for both A and B; however, in the case of A the high surprisal value indicates that such a large delay is unusual at this time of the morning. At C, we find ourselves in the 'usual' low-anomaly area for this airport, where delays are close to zero for most of the day.

6 User Study

In order to evaluate the *petal* and *thread* designs, we performed a user study with 30 participants recruited from various majors at our university. In the study, the participants were given computer-based tasks for verifying hypotheses. Various difference levels in the flight delay network data were used in the tasks. Note that the *difference level* in this section means the difference between the longest (shortest) and second longest (shortest) delays. Note that the numbers in parentheses in the summary tables are the results with visual aids. We use a paired t-test to check if our experimental result obtained is significant (p-value $<$ 0.05) within a 95% confidence interval.

6.1 Petal User Study Results

We first set up the following hypotheses for the *petals* visualization as follows:

HP1 As the difference becomes larger, users will show high accuracy and speed in detecting the longest delay inside a *petal*,

HP2 As the difference becomes larger, users will show high accuracy and speed in finding the shortest (or longest) delay among *petals* for one operational place,

HP3 Users will show lower accuracy in finding the shortest (or longest) *petal* among the *petals* at multiple operational places,

HP4 Users will show low accuracy and speed in finding whether an airport is included in *a petal* as the distance between the *petal* and the airport becomes longer and as an airport is close to the boundary of the *petal*, and

HP5 Visual aids will improve accuracy and speed.

TASK1 for verifying HP1 asked the participants to choose the longest delay inside a *petal* in 2 locations: DP1 (delay difference: 4%)and DP10 (21%), as shown in Fig. 10. The participants showed higher accuracy and speed as the difference increased (Table 1, p-value $<$ 0.05). In TASK2, for verifying HP2, the participants were asked to select the shortest *petal* in 2 locations: DFW (3%) and SLC (12%). For a small difference (3%), 46.7% of the participants answered correctly. As the difference became larger and the visual aid (circle) was provided (HP5), both accuracy and speed were improved (Table 2, p-value $<$ 0.05). TASK3 was the same as TASK2 but multiple *petals* at Salt Lake City (SLC), San Francisco (SFO), and Philadelphia (PHL) were presented concurrently. Here, the participants showed lower accuracy (from 46.7% to 36.7%) and slower speed (from 6.6s to 10.9s) compared to the results in TASK2. The visual aid (HP5) improved both accuracy and speed in the lower difference (Table 3, p-value $<$ 0.05). In order to evaluate if users accurately recognized the coverage of each *petal* (HP4), TASK4 asked the participants to select airports that were included in DP1 and DP9, as shown in Fig. 10. As summarized in Table 4, the participants showed low accuracy (23.3% and 60%). The main reason for such low accuracy was that it was hard for them to find whether CVG (Cincinnati) and PIT (Pittsburgh) were included in DP1. In the same context, only 60% of the participants

Petals	Diff. (%)	Accuracy (%)	Time (s)
SLC+SFO+PHL	2 (Small)	36.7 (73.3)	10.9 (6.8)
SLC+SFO+PHL	12 (Large)	96.7 (96.7)	4.7 (5.1)

Table 3. Users had difficulty finding the longest delay among distant *petals* with a small (2%) difference (HP3). The visual aid helped the users better answer with a small difference (HP5).

Petal Index	# of Airports	Accuracy (%)	Time (avg.)
DP1	8	23.3 (83.3)	1.96 (1.18)
DP9	8	60.0 (93.3)	2.3 (1.3)

Table 4. The participants had difficulty finding whether CVG and PIT were included in DP1 (HP4). The visual aid helped the users better recognize if an airport was included in a *petal* or not (HP5).

correctly found that SAN was not included in DP9. However, with the visual coverage line (HP5), both the accuracy and speed improved.

6.2 Thread User Study Results

Next, we set up hypotheses for the *threads*. As the difference between the longest and the second longest delays becomes larger, the users will produce better results in HT1) detecting the longest delay inside the *threads*, and in HT2) choosing the most prevalent delay among all the *threads*. TASK5 for verifying HT1 asked the participants to select the color of the thickest *thread* for small (3.1%) and large (28%) difference levels. As summarized in Table 5, when the difference was small (3.1%), it was hard for the participants to tell the longest delay (66.7% accuracy). On the other hand, when the difference became larger, they answered very accurately and spent less time (p-value $<$ 0.05). TASK6 for verifying HT2 asked the participants to tell the color of the longest delay when all the *threads* were considered. Here we see a similar result as in TASK5: the larger the difference, the higher the accuracy and the slower the speed (Table 6). In TASK6, we had an interesting result showing that special concentration of a color may interfere with accurate visual perception. For example, we can see LAD (green) is concentrated on short-haul routes as shown in Fig. 11. In this case, 40% of the participants thought that LAD was the longest delay for flights leaving from Atlanta, but in fact the carrier delay was 23% larger than LAD. This error rate is unexpected compared to the result in TASK5 where the participants showed higher accuracy and speed with a similar difference (28.6%). Conversely, we think it is possible that users could assume that the color on long-haul routes has the largest value if the color is concentrated in long-haul *threads*. To prevent this, our system provides numeric information in the legend view that users can refer to, as shown in Fig. 1 (E).

7 LIMITATIONS AND DISCUSSION

Petals have a similar appearance to the rose diagram (or sunburst) visual representation that has been adapted in various contexts [14, 27, 34]. The contribution of *petals* lies in extending the usability of the family of the rose diagram by allowing geographically-directional, multi-variate, and aggregated network analysis simultaneously. Discerning widths of *thread* can be hard when each variable has similar values or when a unit *thread* within a route is not thick enough for visual perception. In addition, when a color is concentrated on long-haul or short-haul routes, it could be hard to select the largest value among all *threads*. To help users with those issues with threads, our system provides the numeric information of the variables in the legend view when a user specifies an area of *threads* (aggregated) and in a tooltip when the user's mouse hovers over an airport (origin to destination). The tooltip in the pixel view can be used for verifying that the presented dominant delay (Fig. 1 (G-2)) is indeed dominant compared to others.

Difference (%)	Accuracy (%)	Time (s)
3.1 (Small)	66.7	5.2
28 (Large)	100	2.9

Table 5. The participants made errors and spent more time finding the longest delay when the difference in *threads* was small (HT1).

Difference (%)	Accuracy (%)	Time (s)
11 (Small)	90	5.3
28.6 (Large)	100	2.9
23 (Large)	60	5.1

Table 6. 40% of the participants answered incorrectly with a large difference (23%) in finding the prevalent delay among all *threads*. This may indicate that color concentration on long-haul or short-haul *threads* interferes with visual perception.

8 CONCLUSION AND FUTURE WORK

We have explored complex multivariate network links with multiple tightly-integrated interactive visualizations. We have introduced two new visual representations, *petals* and *threads*, for spatial multivariate link visualization. Our sortable matrix displays have the ability to represent multiple origin and destination pairs with enhanced pixel-based visualization, while the linked line graph, calendar, and clock views give opportunities to find temporal characteristics. An information-theoretic anomaly detection model was introduced based on conditional attributes, with the visualizations in the system utilizing the surprisal values for visual highlighting of anomalies in multiple visualization components in a unified manner.

Our system has several benefits compared to previous systems. Our system allows users to investigate the data status of a large number of operational locations by simultaneously observing various data characteristics at both aggregate (entire network) and detailed levels (e.g., origin-destination pairs) using our multiple linked view. Our new visual representations, *petals* and *threads*, help users find features of multiple spatial network variables with minimum visual clutter; the pixel-based network matrices aid in analyzing the entire network in terms of multiple origin-destination pairs as well as origin-attribute pairs. Seasonal and cyclical trends can be efficiently detected in the calendar, line graph, and clock visualizations from our system. Lastly, our system provides an information-theoretic model for detecting anomalies based on conditions. For the evaluation of our system, we presented an example using flight delay network data from the top 50 airports to illustrate the use and potential of our designs and the user study results.

Our system can be easily applied to analysis with any other multivariate spatiotemporal, network-based data such as transportation and logistics, trading, and communication industries [7]. As a future work, we plan to incorporate the ability to help users find correlations using *petals* and *threads*. The capability for visualizing cascading effects and clusters of operational places that have the same characteristics will also be investigated. In addition, we would like to explore our anomaly detection more by investigating methods of combining the anomaly values for groups of QCATs.

REFERENCES

[1] Bureau of Transportation Statistics (Accessed 20 Mar 14. `http://www.rita.dot.gov/`.

[2] Operational Evolution Partnership 35. `http://aspmhelp.faa.gov/index.php/OEP_35`.

[3] Qlikview. `http://www.qlikview.com/`.

[4] Tableau. `http://www.tableausoftware.com`.

[5] C. Ahlberg. Spotfire: An information exploration environment. *ACM Special Interest Group on Management of Data Record*, 25(4):25–29, 1996.

[6] C. Arackaparambil, S. Bratus, J. Brody, and A. Shubina. Distributed monitoring of conditional entropy for anomaly detection in streams. In

Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, pages 1–8, 2010.

[7] R. A. Becker, S. G. Eick, and A. R. Wilks. Visualizing network data. *IEEE Transaction on Visualization and Computer Graphics*, 1(1):16–21, Mar. 1995.

[8] R. Borgo, K. Proctor, M. Chen, H. Janicke, T. Murray, and I. Thornton. Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):963–972, 2010.

[9] H. Bosch, D. Thom, M. Worner, S. Koch, E. Puttmann, D. Jackle, and T. Ertl. Scatterblogs: Geo-spatial document analysis. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 309 –310, 2011.

[10] D. Brooks. What data can't do. *The New York Times*, Feb. 2013.

[11] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 143–152, 2012.

[12] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, July 2009.

[13] B. Duffy, J. Thiyagalingam, S. Walton, D. J. Smith, A. Trefethen, J. C. Kirkman-Brown, E. A. Gaffney, and M. Chen. Glyph-based video visualization for semen analysis. *IEEE Transactions on Visualization and Computer Graphics*, 99:1, 2013.

[14] N. Elmqvist, J. Stasko, and P. Tsigas. Datameadow: A visual canvas for analysis of large-scale multivariate data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 187–194, 2007.

[15] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2013.

[16] S. Guha, A. Mcgregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.

[17] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048, 2009.

[18] M. A. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting color schemes for maps. *Cartographic Journal*, 40(1):27–37, 2003.

[19] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. In *IEEE Transactions on Visualization and Computer Graphics*, volume 6 (1), pages 24–43. 2000.

[20] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[21] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann, 2010.

[22] S. Ko, R. Maciejewski, Y. Jang, and D. S. Ebert. Marketanalyzer: An interactive visual analytics system for analyzing competitive advantage using point of sale data. *Computer Graphics Forum*, 31(3):1245–1254, 2012.

[23] Y. Kopylova, D. Buell, C.-T. Huang, and J. Janies. Mutual information applied to anomaly detection. *Communications and Networks, Journal of*, 10(1):89–97, 2008.

[24] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143, 2001.

[25] Z. Liu, S. B. Navathe, and J. T. Stasko. Network-based visual analysis of tabular data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 41–50, 2011.

[26] M. Mazzeo. Competition and service quality in the u.s. airline industry. *Review of Industrial Organization*, 22(4):275–296, June 2003.

[27] F. Nightingale. *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison and Sons, 1958.

[28] A. Noack. An energy model for visual graph clustering. In *Graph Drawing*, volume 2912 of *Lecture Notes in Computer Science*, pages 425–436. Springer, 2003.

[29] D. Oelke, H. Janetzko, S. Simon, K. Neuhaus, and D. A. Keim. Visual boosting in pixel-based visualizations. *Computer Graphics Forum*, 30(3):871–880, 2011.

[30] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[32] A. Z. Santovena. Big data : evolution, components, challenges and opportunities. Master's thesis, Massachusetts Institute of Technology, Sloan School of Management, 2013.

[33] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *IEEE Symposium on Pacific Visualization*, pages 17–24, 2014.

[34] Z. Shen and K.-L. Ma. Mobivis: A visualization system for exploring mobile data. In *IEEE Symposium on Pacific Visualization*, pages 175–182, 2008.

[35] S. S. Stevens. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley, 1975.

[36] D. Thom, H. Bosch, S. Koch, M. Woerner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *IEEE Pacific Visualization Symposium (PacificVis)*, 2012.

[37] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[38] F. van Ham, H.-J. Schulz, and J. M. DiMicco. Honeycomb: Visual analysis of large scale social networks. In *Proceedings of International Conference on Human-Computer Interaction*, volume 5727 of *Lecture Notes in Computer Science*, pages 429–442. Springer, 2009.

[39] Wattenberg, Martin. Visual exploration of multivariate graphs. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, volume 1 of *Visualization 1*, pages 811–819, 2006.

[40] J. V. Wijk and E. V. Selow. Cluster and calendar based visualization of time series data. In *1999 IEEE Symposium on Information Visualization (INFOVIS '99)*, pages 4–9, Oct. 1999.

[41] P. C. Wong, K. Schneider, P. Mackey, H. Foote, G. Chin, Jr., R. Guttromson, and J. Thomas. A novel visualization technique for electric power grid analytics. *IEEE Transactions on Visualization and Computer Graphics*, 15(3):410–423, May/June 2009.

[42] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstädt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. A. Keim. Visual analytics for the big data era - A comparative review of state-of-the-art commercial systems. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 173–182. IEEE Computer Society, 2012.

[43] H. Ziegler, T. Nietzschmann, and D. A. Keim. Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. In *Proceedings of International Conference on Information Visualisation*, pages 287–295. IEEE Computer Society, 2008.

Bristle Maps: A Multivariate Abstraction Technique for Geovisualization

SungYe Kim, Ross Maciejewski, *Member*, *IEEE*, Abish Malik, Yun Jang, *Member*, *IEEE*, David S. Ebert, *Fellow*, *IEEE*, and Tobias Isenberg, *Member*, *IEEE*

Abstract—We present Bristle Maps, a novel method for the aggregation, abstraction, and stylization of spatiotemporal data that enables multiattribute visualization, exploration, and analysis. This visualization technique supports the display of multidimensional data by providing users with a multiparameter encoding scheme within a single visual encoding paradigm. Given a set of geographically located spatiotemporal events, we approximate the data as a continuous function using kernel density estimation. The density estimation encodes the probability that an event will occur within the space over a given temporal aggregation. These probability values, for one or more set of events, are then encoded into a bristle map. A bristle map consists of a series of straight lines that extend from, and are connected to, linear map elements such as roads, train, subway lines, and so on. These lines vary in length, density, color, orientation, and transparency—creating the multivariate attribute encoding scheme where event magnitude, change, and uncertainty can be mapped as various bristle parameters. This approach increases the amount of information displayed in a single plot and allows for unique designs for various information schemes. We show the application of our bristle map encoding scheme using categorical spatiotemporal police reports. Our examples demonstrate the use of our technique for visualizing data magnitude, variable comparisons, and a variety of multivariate attribute combinations. To evaluate the effectiveness of our bristle map, we have conducted quantitative and qualitative evaluations in which we compare our bristle map to conventional geovisualization techniques. Our results show that bristle maps are competitive in completion time and accuracy of tasks with various levels of complexity.

Index Terms—Data transformation and representation, data abstraction, illustrative visualization, geovisualization

1 Introduction

As data dimensionality increases, the encoding of variables and their relationships is often abstracted down to a representative subset for analysis in a single display, or dispersed across a series of coordinated multiple views [1], [2], [3]. Moreover, many techniques have been developed to visually encode multiple data attributes/variables for each data sample to enable interactive analysis, ranging from discrete glyph attribute encoding [4] to more spatially continuous color, transparency, and shading encodings [5], [6], [7]. As the number of visualized variables increases, the amount of information that can be effectively displayed becomes limited due to overplotting and cluttering [8]. This is especially a problem in geographical visualization as a key attribute of the data is the location within the two-dimensional map space.

In geographical visualization, data can be described at any given location on a map. The data being described can come from an aggregated measurement, a direct event occurrence, or various other means. In dense data sets, plotting events as symbols on the map (e.g., Fig. 1a) leads to cluttering and is often unable to convey a meaningful sense of event magnitude within the data. Aggregation of the data by defined boundaries, such as county or census tract boundaries (e.g., Fig. 1b), leads to a loss of specificity in data location and runs afoul of the Modifiable Areal Unit Problem [9]. Furthermore, it is known that the level of data aggregation can affect aspects of task complexity such as information load and the user's ability to recognize patterns within the data [10]. To combat problems associated with areal aggregation, dasymetric mapping focuses on using zonal boundaries that are based on sharp changes in the statistical surface being mapped [11]. However, even when grouping data into small spatial quadrats, data can either be overaggregated or underaggregated. A third option is to estimate the discrete event points as a continuous function (e.g., Fig. 1c); such a mapping, however, only allows for the use of color as a means of representing data variables. As an encoding based on underlying network data, Fig. 1d shows a traditional line map. However, its representation is still restrained by the color and thickness of the lines.

To increase the amount of information that can be visualized within the constraints of a thematic map, this paper explores a novel method of multivariate encoding. Inspired by ideas of symbolic encoding from Spence [12] and choices of visual encodings by Wilkinson [13], we have developed the bristle map (Fig. 1e), a novel method for the aggregation, abstraction, and stylization of geographically located spatiotemporal data. The bristle map consists of a series of straight lines extended from and connected to linear map elements (roads, train lines, subway lines, etc.) that have some contextual relationship with the data being

- *S. Kim, A. Malik, and D.S. Ebert are with the School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Avenue, West Lafayette, IN 47907. E-mail: {inside, amalik, ebertd}@purdue.edu.*
- *R. Maciejewski is with School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Mail Code 8809, Tempe, AZ 85287-8809. E-mail: rmacieje@asu.edu.*
- *Y. Jang is with Department of Computer Engineering, Sejong University, 98 Gunja-dong Gwangjin-gu, Seoul, 143-747, South Korea. E-mail: jangy@sejong.edu.*
- *T. Isenberg is with Team Aviz, INRIA-Saclay, Bât 650, Université Paris-Sud, 91405 Orsay Cedex, France. E-mail: tobias.isenberg@inria.fr.*

Manuscript received 14 Feb. 2012; revised 7 Nov. 2012; accepted 2 Mar. 2013; published online 20 Mar. 2013.
Recommended for acceptance by M. Agrawala.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2012-02-0030.
Digital Object Identifier no. 10.1109/TVCG.2013.66.

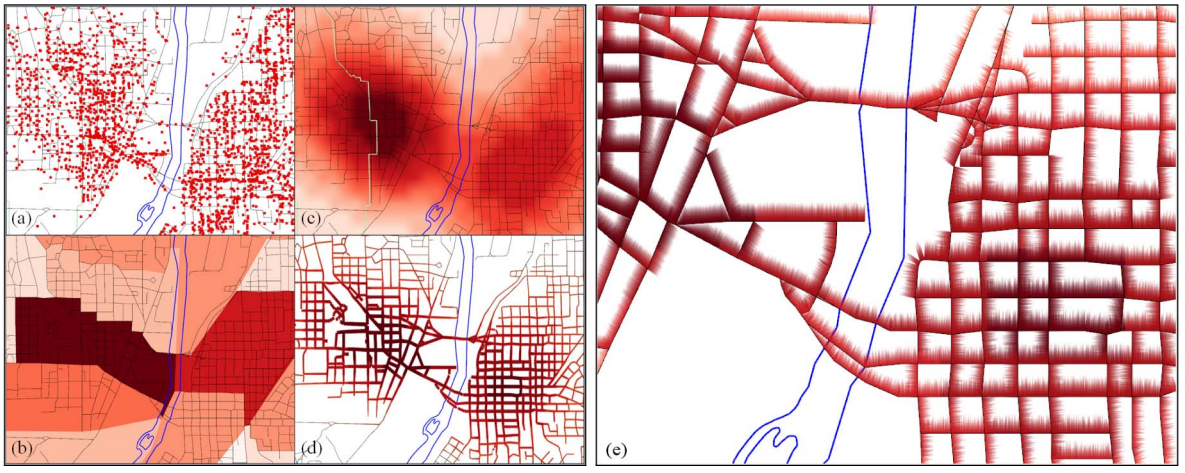


Fig. 1. Data abstraction in geovisualization. In this image, we show crimes in West Lafayette and Lafayette, Indiana, where the blue line represents the Wabash River. (a) Plotting events as points. (b) Aggregation of points by areal units. (c) Approximation of a continuous domain from point sampling. (d) Approximation of a continuous domain using solid lines applied to roads. (e) Our abstraction using a series of bristle lines applied to roads.

visualized. We vary these lines with respect to their color, length, density, and orientation to allow for a unique encoding scheme that can be used to create informative maps. With respect to the other representations shown in Fig. 1, our technique utilizes the underlying geographical context as a part of its symbology, thereby directly incorporating geographical elements within its encoding scheme. One of the major advantages of the bristle map technique is that the basis domain of the data (e.g., street network) remains highly visible regardless of the color scale being used. If one compares Figs. 1c and 1e, the street network in Fig. 1e is clearly visible because the lines only "bristle off" to one side, whereas in Fig. 1c some streets are hardly discernible due to the dark colors.

To demonstrate our technique, we focus on categorical spatiotemporal event data (e.g., emergency department logs, crime reports). In such data, events consist of locations in time and space where each event fits into a hierarchical categorization structure. These categories are typically processed as time series and snapshots of time are aggregated and typically visualized on a choropleth map [14]. Past work [6], [15] has shown that the use of kernel density estimation (KDE) [16] is highly suitable in the spatial analysis of such data. Thus, our approach incorporates kernel density estimation as a means of estimating the underlying distribution of spatiotemporal events. Using the estimated distribution in an area for a given category (or categories) and temporal unit, we incorporate the underlying geographical network structure into the visual encoding. Bristles are extended from this underlying structure, and the color, length, density, transparency, and orientation of each bristle are mapped to a particular variable (or set of variables). Schemes presented in this paper include combinations of the following mappings:

- length, density, and color as data magnitude,
- orientation and coloring for bivariate mapping,
- color and length for bivariate mapping,
- color and density for bivariate mapping, and
- length and transparency for temporal variance.

Given the available parameters for visual encoding within the bristle map, other encodings also exist, which illustrate the flexibility and power of our technique. Our work focuses on showing how bristle maps can be used to show spatial and temporal correlations between variables, encode uncertainty in a unique and aesthetically informative way, and maintain geographical context through linking our visual encoding directly to geographical components. As such, the bristle map is a powerful multivariate encoding scheme that is adaptable to various attribute encodings to create richly informative visualizations.

2 Related Work

Many techniques in multivariate data visualization focus on a means of reducing clutter and highlighting information through a variety of approaches including filtering (e.g., [17]), clustering (e.g., [18]), and sampling (e.g., [19]). In this section, we focus particularly on techniques within geographical visualization for improving the understanding of thematic/statistical maps, as Wilkinson [13] noted that the problem of multivariate thematic symbology for maps is that they are not only challenging to make, but also challenging to read.

In geographical visualization, the most common means of data representation is the choropleth map in which areas are shaded or patterned in proportion to a measured variable. Such maps are typically used to display only one variable, which is mapped to a given color scale. Other research has focused on encoding multivariate information into choropleth maps (such as uncertainty) with textures and patterns [20], creating bivariate color schemes for visualizing interactions between two variables [21], [22], or animating choropleth maps to enhance the exploration of temporal patterns and changes [23]. We present bristle maps as a robust alternative to these schemes in which multivariate attributes are instead mapped to a variety of graphical properties of a line (length, density, color, and orientation), as opposed to utilizing a bivariate color scheme, texture overlays, or animation.

More recent geographical visualization techniques have included extensions to choropleth mapping ideas. Hagh-Shenas et al. [24] compared the effectiveness of visualizing geographically referenced data through the use color blending (in which a single composite color conveys the values of multiple color encoded quantities) and color weaving methods (in which colors of multiple variables are separately woven to form a fine grained texture pattern). The results from their study indicate color weaving to be more effective than color blending for conveying individual distributions in a multivariate setting. Saito et al. [25] proposed a two-tone pseudo coloring method for visualizing precise details in an overview display. Under this scheme, each scalar value is represented by two discrete colors. Sips et al. [26] focused on revealing clusters and other relationships between geospatial data points by their statistical values through the overplotting of points. This work was later extended [27] to combine a cartogram-based layout to provide users with insight to the relative geospatial positioning of the data set while preserving cluster information and avoiding overplotting. Other cartogram techniques include the WorldMapper Project [28] which is used to represent social and economic data of the countries of the world. In each of these, novel data visualization techniques are created; however, the distortion of spatial features (country boundaries, roads) is often undesirable. While these techniques focus on displaying large amounts of aggregate data on small screens, our technique focuses on enhancing details of geographical context within the data. A similar concept of preserving data context is found in Wong et al.'s [29] GreenGrid in which they visualize both the physics of the power grids in conjunction with the geographical relationships using graph-based techniques.

Along with the previously described map schemes and cartogram distortions, there has been work in the use of heatmaps based on spatial data. Fisher [30] applied heatmaps to visualize the trends of the interactions of users with interactive maps that are based on their view of the geographic areas. Maciejewski et al. [6] used heatmaps as one of the tools to find aberrations or hotspots that facilitate the exploration of geo-spatial temporal data sets. Work by Chainey et al. [15] illustrated a number of different mapping techniques for identifying hotspots of crime and demonstrated that kernel density estimation provides analysts with an excellent means of predicting future criminal activities.

In conjunction with previous visualizations, other research has focused on expanding the dimensionality of the data being displayed by utilizing three-dimensional visuals. Van Wijk and Telea [7] utilized color and heightfields to visualize scalar functions of two variables. Tominski et al. [31] explored embedding 3D icons into a map display as a means of representing spatiotemporal data. In contrast, our work focuses on a two-dimensional encoding scheme that incorporates a variety of the visual variables described by Bertin [32] and Wilkinson [13] as a means of representing multivariate data.

Finally, it is important to note that our technique is akin to traditional traffic flow maps (e.g., Fig. 1d) seen in a variety of atlases; however, provides more generalized schemes. In traffic flow maps, the amount of data that can be displayed is restrained by the color and the width of the line representing linear elements (i.e., roads) on the map. Our work is similar to that of the traffic flow maps in that we utilize width (specifically, matched to the length in our bristle maps) and color as underlying visual variables of our encoding. However, our work also incorporates bristle density as a means of further encoding parameters. In the following sections, we compare our encodings to a variety of methods including the point, color, and flow line maps.

3 Bristle Map Generation

In Fig. 1, we developed our motivation for the need to directly incorporate geographic features to the underlying data to better preserve contextual information. It is clear that the aggregation of data into arbitrary geographical areas obscures data, while the continuous approximation of an underlying data source can lead to incongruent mappings with respect to geographic features. Furthermore, both these mappings are limited in the fact that only color and texture are available for variable encoding, limiting the amount of data that can be displayed to either a single variable or possibly two variables in the case of a bivariate color map. The goal of this work is to create visual encodings for higher order structures.

The bristle map was inspired by the Substrate simulation of Tarbell [33] and abstract renderings of map scenes in work by Isenberg [34]. Given these images, our work focuses on using the underlying visual properties to intelligently encode information for display. In *The Grammar of Graphics* [13], Wilkinson discusses the combination of several perceptual scales into a single display. Here, he notes the idea of separable dimensions of the data is a key issue, where discriminations between stimuli are of key importance in the visualization. The Substrate aesthetic directly lends itself to this approach as color, line length, and orientation are distinct classes within Wilkinson's table of aesthetic attributes and each of these visual parameters directly contributes to the substrate aesthetic.

Fig. 2 illustrates the bristle map generation pipeline. Given underlying data events, we compute a continuous distribution. We also create a topology graph from given geographically relevant linear content for clutter reduction described in Section 5. As an example of geographical content, if the underlying data was water pollution we could use a city sewage map for the geographic components, for our crime data examples we use roadways. Each linear geographic component consists of a series of line segments, and we extend *bristle lines* from these line segments. These bristle lines emerge perpendicularly from the underlying geographical line segment and are allowed to vary in length, density, color, transparency, and orientation, to facilitate multivariate data encoding. The third stage of the bristle map generation pipeline (Fig. 2) illustrates the bristle line concept for each geographical line segment, $\overline{SE}$, and $\overline{P_1P_2}$ defines our generated bristle line. Each bristle line is created using the vector equation of a line as shown in

$$P_2 = P_1 + \vec{V_l}L_l = P_1 + \vec{V_l}(t \times L_{lmax}). \quad (1)$$

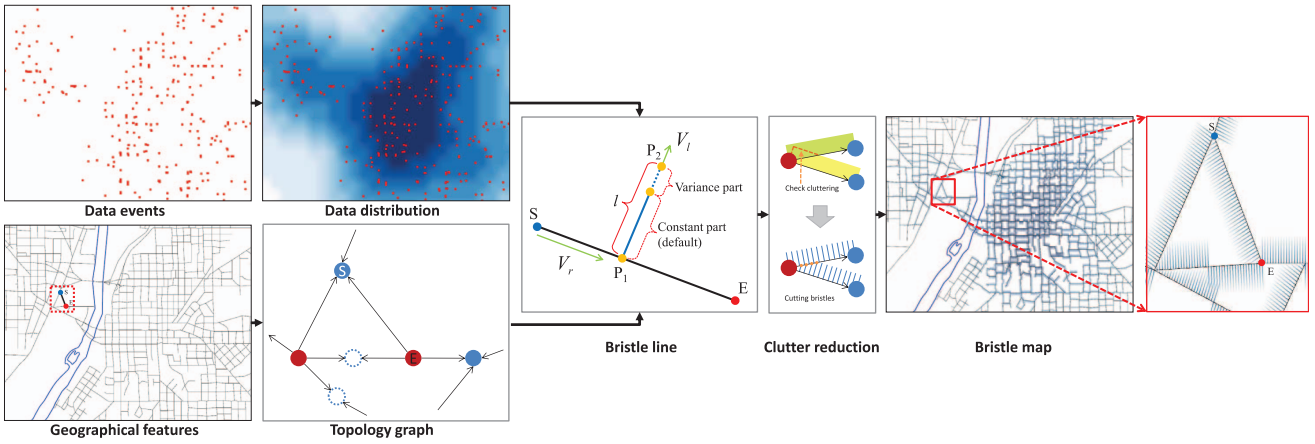


Fig. 2. The bristle map generation pipeline. Beginning with data events, a continuous abstraction is created. We also create a topology graph from contextually important linear features (in this case roads). Next, bristles are extended from these features based on the continuous abstraction and the topology. Clutter reduction is performed when generating each bristle, and finally the resultant bristle map is generated.

Here, P_1 is a point on the contextually relevant geographic line segment, $\overline{SE}$, $\overrightarrow{V_l}$ is a unit vector perpendicular to the line $\overline{SE}$, and L_{lmax} is the maximum length of the $\overline{P_1P_2}$. L_l is the length of the $\overline{P_1P_2}$ determined by a parameter t.

Each line from P_1 to P_2 is drawn in such a manner that it will either encode different properties of a multivariate data set, or use a data reinforcement technique where properties are encoded to the same variable to provide redundant cues. We utilize three encoding properties for each bristle: length, color, and orientation. The length of a line $\overline{P_1P_2}$ is separated into two portions: a constant component, which is proportional to the magnitude of the variable being encoded, and a variance component. It can also capture other properties such as level of certainty. The color of a bristle $\overline{P_1P_2}$ is proportional to the underlying variable distribution to be encoded at point P_1. When the variance component is used, its transparency is adjusted as a means of visually distinguishing it from the constant component. Orientation of the bristle line is always perpendicular to $\overline{SE}$ and is utilized for bivariate comparison (i.e., day/night, two data types) and/or clutter reduction. To summarize, length and color represent a local data magnitude property at point P_1. We also choose to encode redundant information into the density of the number of bristles placed on a given line segment, where the density of the bristles along $\overline{SE}$ is decided by an average data value on a line segment $\overline{SE}$.

For each visual encoding, the underlying data is assumed to be continuous over a given geographical segment, such that for all points between any two nodes on the underlying contextual geographic structure, a data distribution value is associated with the point. In the case of a discrete data set (e.g., crime locations), the choice of an appropriate means of data interpolation with regards to the underlying geographic information is dependent on the data analysis being performed. Based on the recommendations of Chainey et al. [15], we apply a kernel density estimation [16] to approximate the underlying distribution of crimes over the geographic features. The kernel density estimation procedure used is defined by the following equation:

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{\mathbf{x}-X_i}{h}\right). \tag{2}$$

Here, the window width of the kernel placed on point x is proportional to a window bandwidth, h, and the total number of samples, N. We utilize the Epanechnikov kernel [16]:

$$K(\mathbf{u}) = \frac{3}{4}(1-\mathbf{u}^2)1_{(||\mathbf{u}||\leq 1)}, \tag{3}$$

where the function $1_{(||\mathbf{u}||\leq 1)}$ evaluates to 1 if the inequality is true and zero for all other cases.

Thus, given a multivariate data set where locations in space and time correspond to a series of categorized events, we can create bristle maps that encode various properties of the data. Note that this technique relies on the data being contextually relevant to an underlying geographical network. For example, crime event data with its 2D geographical coordinates is recorded and hence defined by addresses on streets; thus, it is contextually relevant to a street network. Data sets in which this contextual relationship does not exist should utilize other visual encoding schemes. Table 1 shows the parameters in our bristle map and their corresponding potential variables being encoded to each parameter. In the following section, we present a series of potential parameter combinations for various bristle map encodings and discuss the various results.

TABLE 1
Parameters, Corresponding Variables, and Ranges

Parameters	*Potential variables*	*Range*
Base position (P_1)	Geographic location	(Double, Double)
Length 1 (constant portion)	Data magnitude	Double
Length 2 (variance portion)	Temporal variance, accuracy	e.g., monthly/yearly
Color	Data magnitude	Discrete, Continous
Transparency	Temporal variance, accuracy	Double [0.0, 1.0]
Orientation ($\overrightarrow{V_l}$)	Temporal difference, data type	Clock-wise, Counter clock-wise
Density	Average data magnitude on an area ($\overline{SE}$)	Double

4 Encoding Schemes

The bristle map is a powerful visual encoding scheme that lends itself to a variety of data encodings, examples of which we present next. For demonstration purposes, we employ categorical spatiotemporal police reports collected in Tippecanoe County (specifically West Lafayette and Lafayette, IN, USA), from 1999 to 2010. The data set contains the date, time, crime type (e.g., armed/unarmed aggravated assault, armed robbery, burglary, homicide, noise, other assaults, rape, rape attempted, residential entry, robbery, theft, vandalism, and vehicle theft), and the address of each recorded criminal event. Note that other data sets can be easily encoded with bristle maps, and our choice of data was only made to illustrate the technique.

Utilizing this multivariate crime data set, we discuss potential encoding schemes for multivariate spatiotemporal data. We then provide illustrations of each described encoding scheme with respect to our crime data set. Encoding schemes presented in this section include the use of bristle color, length, and density to encode data magnitude, the use of bristle orientation to inform temporal comparison, and the encoding of temporal variance in the bristle lengths.

4.1 Color, Length, and Density as Data Magnitude

Here, we discuss our technique for encoding the color, length, and density of the bristles into two separate variable groups. As both color and length (size) fall into two distinct categories of aesthetics according to Wilkinson [13], the use of separate variables for both categories allows for a distinguishable visual data encoding. In both cases, we assign data magnitude to both a color scale and a length scale. We note that such an encoding scheme has the potential to portray data more effectively than visualizations that map each data variable to a single display parameter. As noted in the arguments for the use of redundant color scales by Rheingans [35], the use of different display parameters is able to convey different types of information. Furthermore, by combining encodings in a redundant manner, it is possible to reinforce the encoding scheme. The utility of redundant color scales was confirmed by Ware [36].

In our encoding scheme, each bristle line's length, L_l, is calculated using (4) based on a parameter, t, and the maximum length, L_{lmax}:

$$L_l = t \times L_{lmax} = (\alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1}) \times L_{lmax}. \quad (4)$$

For this visual encoding of the bristles, the parameter t is defined by the ratio of the data value at P_1, which we call κ_{P_1}, the ratio of the temporal variance at P_1, υ_{P_1}, and a set of tuning parameters (α and β) that provide weights to the constant and variance components as shown in Fig. 2. In this work, we use $\alpha = 1.0$ and $\beta = 0.3$. Note that the choice of encoding the variance at a 30 percent value was chosen through trial and error by generating visualizations that the authors found to be the most useful and aesthetically pleasing. For problems where determining exact data values from the visual encoding is required (as opposed to approximating high and low rates), the variance portion is removed from the equation entirely by using $\beta = 0.0$. As such, by creating the encoding scheme with diverse parameters, we are able to generate more aesthetic choices and visualizations. It is important to note that not all encodings will be appropriate and are most likely task dependent.

The L_{lmax} portion of (4) is defined in

$$L_{lmax} = \rho \times log_b \left(\frac{1}{N_r} \sum_{i=0}^{N_r - 1} L_{\overline{SE}} \right). \quad (5)$$

In this equation, we take the average length of all line segments (where N_r is the total number of line segments in the map) and calculate L_{lmax} using a nonlinear function such that the length of bristle lines does not grow in an unbounded manner when zooming in. Moreover, L_{lmax} is modified by the parameter ρ, where ρ is the ratio of the current zoom level to the initial zoom level, to decouple our technique with the zoom level. In this work, we use $b = 15$ for the base of a log function.

Next, we determine the number (or density) of bristles, N_l, to be drawn on each line segment $\overline{SE}$ using

$$N_l = \rho \left(\frac{\zeta}{\lambda} L_{\overline{SE}} \right) \kappa_{\overline{SE}}. \quad (6)$$

Here, N_l is calculated using two user-defined constants λ and ζ, where λ is the unit geographical length (distance) and ζ is the number of bristle lines per unit geographical length. We use $\lambda = 0.0009$ and $\zeta =$ 3-15 in our current visualization. As the bristle density may also be used to encode data magnitude parameters in bristle map generation, N_l should be proportional to the ratio of average data value on $\overline{SE}$, $\kappa_{\overline{SE}}$. Moreover, we also apply ρ such that N_l will be independent of the zoom level to preserve the extent of density.

For color, we allow users to choose either a continuous or a sequential color scheme from Color Brewer [37]. Then, data are linearly mapped to a probability that a crime of type A will occur at geographic point B, where the probability is estimated from the underlying data distribution using kernel density estimation as described in Section 3.

Fig. 3 illustrates our length, density, and color encoding using the previously described crime data set. Burglary is encoded with the red color scheme, and color is proportional to the probability (calculated from the underlying point distribution using kernel density estimation) that a burglary occurred at a given location. Fig. 3(left) shows our bristle map encoding for burglary rates with a color scheme and bristle density, and Fig. 3(right) shows a line map encoding the same information with a color scheme and line thickness for comparison to our bristle map. Compared to this line map, our bristle map provides the advantages of additional dimensionality through the density of bristle lines. In this scheme, one is able to easily encode two variables in different combinations of bristle map parameters (i.e., color and density with a constant length, color and length with a constant density), and provide users with distinguishable visual parameters that seem to focus attention to various details.

4.2 Multivariate Encoding: Separating Length, Density and Color, and Using Orientation

In the previous section, we illustrated how our method can be utilized for univariate encoding by using a redundant encoding scheme. However, a major benefit of bristle maps

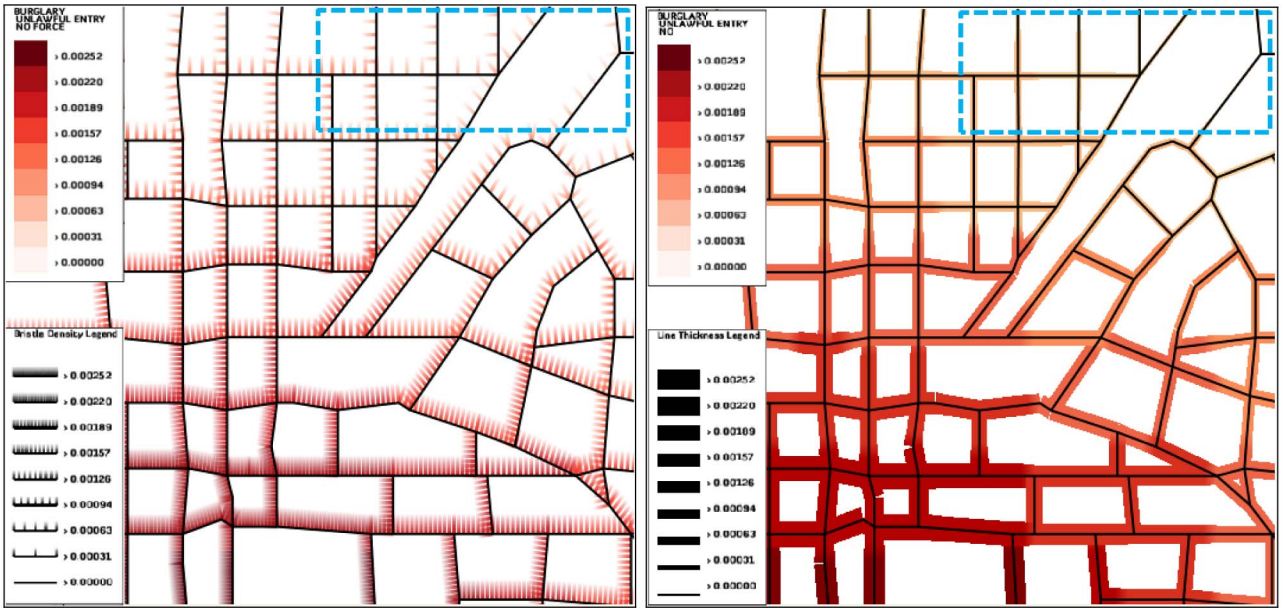


Fig. 3. (Left) Our bristle map encodes burglary rates with both bristle color and bristle density. (Right) A line map encoding burglary rates as both line color and line thickness. Compared to the line map, our bristle map provides a distinguishable visualization by incorporating bristle density. For example, bristle lines on the right top area are easily identified, whereas thickness in the line map on the same area is too small to clearly be perceived.

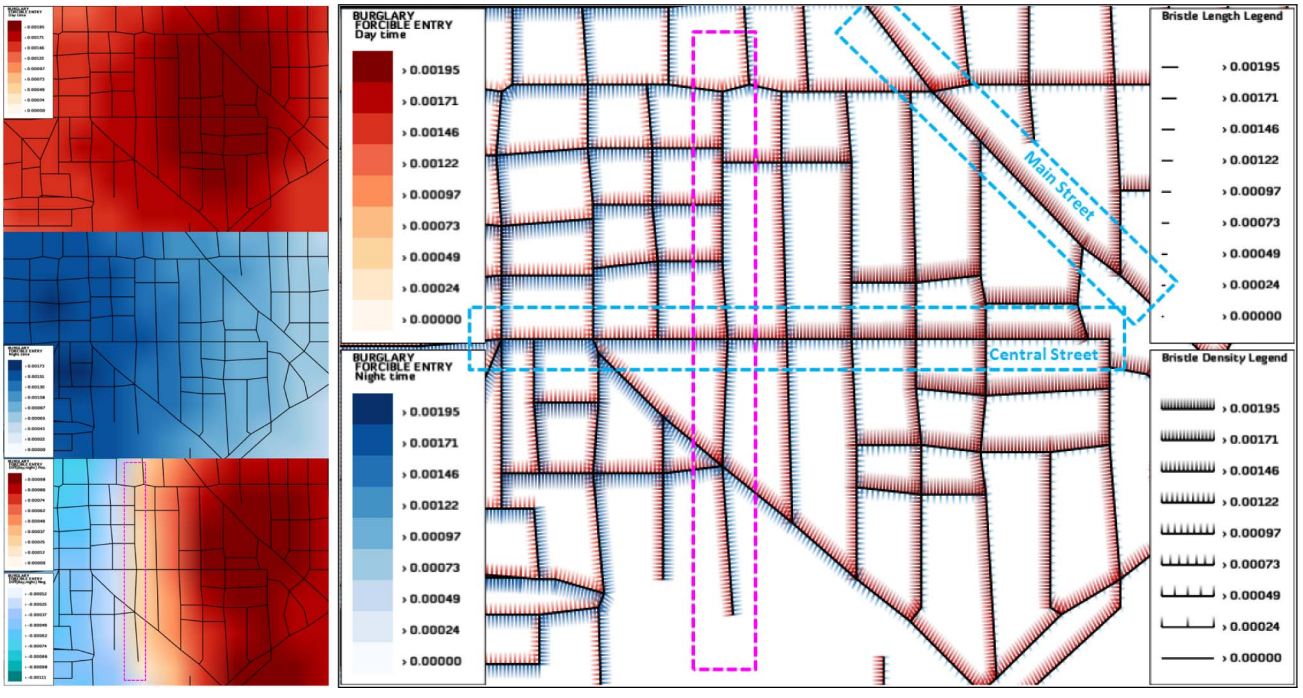


Fig. 4. Encoding daytime versus nighttime variations. (Left column) From top to bottom, color maps showing day, night, and the difference of day and night burglary rates. (Right) Our bristle map separating the burglary rates into their day and night components with opposite orientation along roads. Note that a color map cannot present two components (i.e., both daytime and nighttime burglary rates) at the same location, hence three color maps are needed to see day and night variations simultaneously. Our bristle map can present such information within one bristle map by using different orientations of bristle lines.

is the ability to encode multivariate attributes. One example of this is seen in day versus night time comparison.

Here, one can utilize the orientation to separate two temporal components of a single variable by mapping the temporal components to different orientations of the geographic feature. For instance, it is likely that the rates of data variable will be different with respect to day and night occurrences. We illustrate this visual encoding in Fig. 4. We separate the events into day (6:00 am-6:00 pm) and night (6:01 pm-5:59 am) and map the daytime rates to red and one orientation, and nighttime rates to blue and the other orientation. In Fig. 4(right), we illustrate a bristle map encoding of one variable (burglary) during 2009 where length, density, and color represent the magnitude of the burglary as well as the encoding of day and night parameters is explored as line orientation.

In Fig. 4(right), we show areas of high/low nighttime crime, high/low daytime crime, and combinations there

within. In contrast, a traditional heatmap using a univariate color scheme can only show either daytime crime (Fig. 4(left top)), or nighttime crime (Fig. 4(left middle)). Hence, several heatmaps are needed to see day and night variations as shown in Fig. 4(left column). Viewers must mentally combine the images to locate regions of the map that have high crime levels at daytime and nighttime, thereby increasing their cognitive load.

Another means of reducing the cognitive burden would be to create a heatmap of the difference between night and day. Fig. 4(left bottom) shows the difference of day and night data, and the divergent color scheme shows where high daytime or high nighttime crimes occur. For instance, in Fig. 4(left bottom) the right area indicates higher rates during day, the left area shows higher rates during night, and the border area between the blue and red color schemes only indicates that day and night rates were approximately equal, regardless of them being low or high. Moreover, you need other color maps to explore areas, where one occurs similarly high or low during day and night time.

Bristle map encodings have benefits in this situation. When we explore a daytime versus nighttime bristle map in Fig. 4(right), we see that there exists distinct temporal profiles along the road lines, where we see exclusively dominant areas during either day or night. For instance, see the diagonal road from the top center to the right center (Main Street, Lafayette, IN) showing that daytime burglary dominates along this road. Another observation is made on the horizontal road at the center of the map (Central Street, Lafayette, IN). Along this road, daytime burglary rates increase from left (west) to right (east), whereas nighttime burglary rates decrease from left to right. For the center area in Fig. 4(left bottom), where the blue and red color schemes meet, we also see in Fig. 4(right) that it has relatively equally high rates during both day and night. Such a comparison allows people to understand the differences between the data; however, when subtracting, areas of nearly equal daytime and nighttime crimes will be colored the same. Thus, areas that are safe during both day and night, and areas that are highly dangerous during both day and night will appear the same in the difference color map. In contrast, bristle maps allow viewers to quickly observe trends related to both day and night.

Another example of multivariate encoding using our bristle map is done by separating and/or combining bristle parameters. For instance, bristle density (or length) encodes a variable A, and color encodes a variable B while being presented on one orientation. Similarly, another two variables (C and D) could be encoded and presented on the other orientation. However, this type of parameter combination should be determined carefully so as not to increase viewers' cognitive load. Its effectiveness would depend on several factors such as data type and analysis purpose. In Section 6, we conduct experiments to explore the effectiveness of different parameter combinations.

4.3 Encoding Data Variance

As introduced in Fig. 2, each bristle can include a portion generated for temporal variance of data, see (4). To present the temporal variance of the data over time, we compute both the monthly and yearly mean and variance values. For a given discrete data set during time periods N_T, we first calculate continuous distributions over time. Then, we determine mean and standard deviation values with respect to the underlying data distribution for the entire data set over a given temporal aggregation. Thus, we calculate the mean μ and variance σ values from time varying data K_i, where $i \in [0, N_T - 1]$. Note that μ and σ are computed only once as they represent constant values for a given data set. Mean and variance values for each grid point j are calculated using (7) and (8), respectively. Variance is then used to weight the parameter β in (4) such that given the data magnitude at the current time K_{cur}, we compute the ratio of variance at the current time, $\tilde{\sigma}$ as shown in (9). As such, the parameter t in (4) can be detailed as shown in (10) to represent the length of bristle lines with respect to temporal variance:

$$\mu[j] = \frac{1}{N_T} \sum_{i=0}^{N_T-1} K_i[j] \tag{7}$$

$$\sigma[j] = \sqrt{\frac{1}{N_T} \sum_{i=0}^{N_T-1} (\mu[j] - K_i[j])^2} \tag{8}$$

$$\tilde{\sigma}[j] = \frac{1}{\sigma[j]} |\mu[j] - K_{cur}[j]| \tag{9}$$

$$t = \alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1} = \alpha \times \kappa_{P_1} + \beta \times \left(\frac{\tilde{\sigma}_{P_1}}{\tilde{\sigma}_{max}} \right). \tag{10}$$

Furthermore, the variance term, υ_{P_1}, in parameter t in (4) can also be revised to encode an uncertainty factor by using randomness. We may also encode an uncertainty factor by using color and transparency to enhance the variance component. When using color and transparency, we use a highlight color for the variance component, and then fade out the variance component over the bristle length with a full alpha value for one end point and an alpha value weighted by the variance for the other end point. The constant portion of the bristle is assigned an alpha value of 1 to both end points as it represents an exact data value. Hence, according to the data type and analysis purpose, the encoding of parameter t and the use of the variance portion can be different and should be assessed with respect to the visual message trying to be conveyed. Fig. 5 illustrates the application of encoding the data variance of vandalism with the uncertainty factor. In Figs. 5a and 5c, we use the same color scheme for the constant and variance portions of bristle lines. To enhance the variance component in Figs. 5b and 5d, we highlighted the variance portion in a different color and assigned full alpha values for the constant portion of bristle lines. Figs. 5a and 5b show the same area. In this area, the bristle length shows large fluctuations, indicating a high yearly variance Figs. 5c and 5d show another area. In this area, the bristles are of a nearly constant length, indicating low yearly variance. When considering that the area in Figs. 5a and 5b includes residential areas, while the area in Figs. 5c and 5d includes the downtown Main street, an art theater, and the City Hall in Lafayette, IN, our bristle map shows that the residential areas have higher yearly

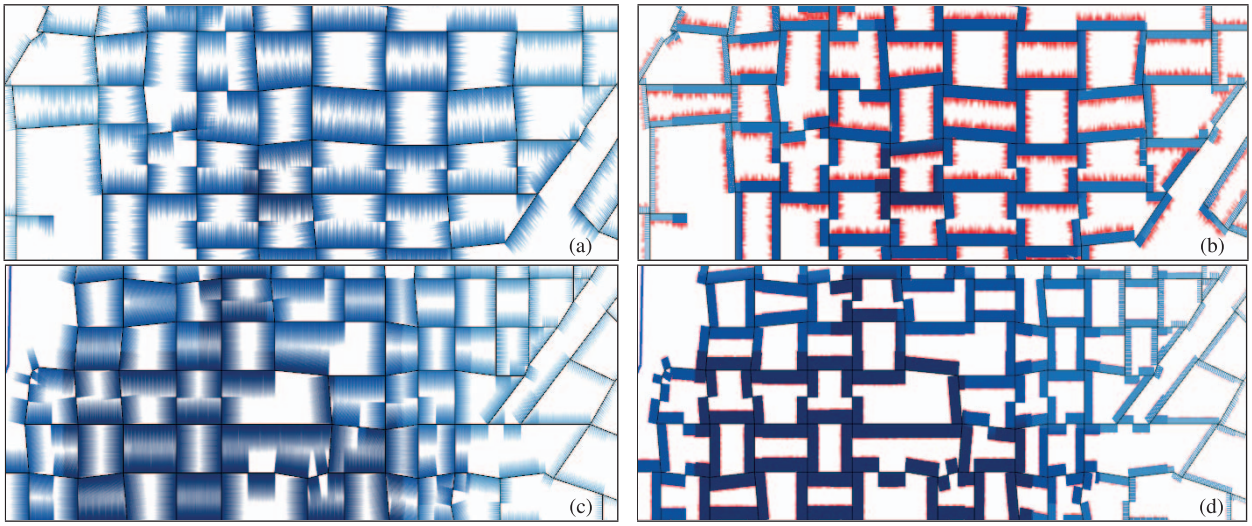


Fig. 5. Encoding data variance of vandalism-graffiti in Lafayette, IN, USA, in 2010 creating an uncertainty aesthetic. Yearly variance of vandalism-graffiti is represented in (a) a residential area and (c) a commercial area without distinguishing the variance component in the bristle length. Parts (b) and (d) show the results using a highlight color for the variance portion and full alpha values for the constant portion of bristle lines. Here, we clearly see that our bristle map can encode the temporal variance and create an uncertainty aesthetic using the variance component.

variance of vandalism (graffiti) when compared to commercial areas.

5 Bristle Clutter Reduction

Although our bristle map can encode various characteristics from multivariate data, it often suffers from clutter around the intersections of road lines. To minimize cluttering, we employ two strategies in our bristle map generation pipeline (Fig. 2): 1) using topology among road lines to determine bristle orientation to minimize clutter and 2) cutting bristle lines crossing neighbor road lines.

5.1 Using Topology

Each bristle map contains an underlying topology of the contextual geographic network that the data re mapped to. In the topology graph, each node is defined as either "outward" or "inward" as illustrated in Fig. 6. Using the topology graph, we choose each segment's bristle line orientation such that the overlap of the bristles at intersections will be minimized, thereby reducing the clutter. If the encoding scheme requires both sides of the edge to contain bristles, then clutter at each intersection is inevitable. However, in cases where bristles map to only one side of an edge, we use the right-hand rule to decide the orientation. Hence, bristle lines on edges connected to neighboring outward and inward nodes are generated in a manner that provides a reasonable reduction in clutter (Fig. 6).

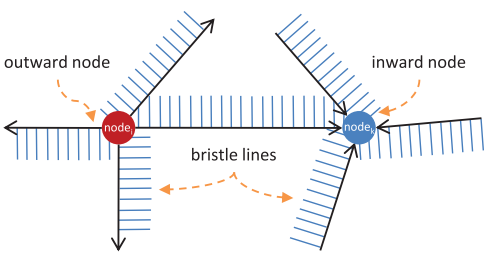


Fig. 6. To minimize clutter, a topology graph consisting of directed edges as road lines and outward (red) and inward (blue) nodes on the intersection of lines is used to decide the bristle line orientation.

Choosing the orientation of bristle lines to minimize overlap can be considered as a 2-coloring problem in vertex coloring; one color presents "outward" while the other presents "inward." Vertex coloring is a well-known graph problem, where no two adjacent nodes share the same color. Moreover, coloring a general graph with the minimum number of colors is known to be an NP-complete problem. In our case, the minimum number of colors should always be 2 but such 2-colorability is not guaranteed for general road lines. While deciding the orientation of bristle lines, we often have undesirable topology generating inevitable overlap of bristle lines. Fig. 7 (upper row) shows such a bad topology example and our strategy to solve this issue. In Fig. 7a, we see two clutter areas caused by an undesirable configuration of neighbor nodes, which guarantee bristle overlap. To solve this, we consider the addition of a virtual node in a topology graph as shown in Fig. 7b, thereby allowing for an orientation switch midway across

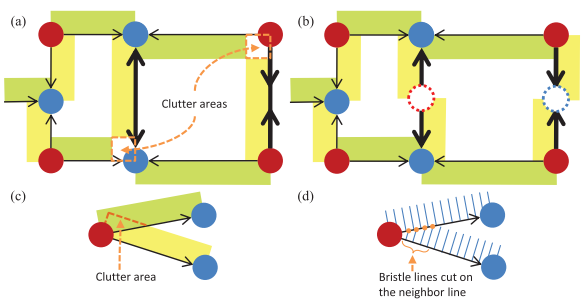


Fig. 7. Two pairs of the cluttering cases and our methods to minimize clutter. Colored box areas on a side of each edge line indicate the orientation for bristle lines. (a) Case 1: bad topology, where two inward nodes (blue) share a line and two outward nodes (red) share a line, generates inevitable clutter. (b) Virtual nodes (dotted circles) are added to split an edge line. (c) Case 2: a small angle between edge lines causes a clutter area. (d) Bristle lines crossing a neighbor edge line are cut on the neighbor line.

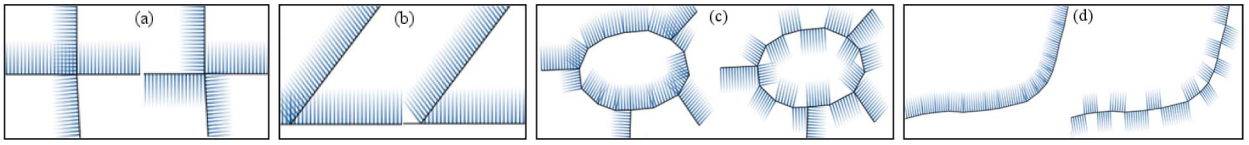


Fig. 8. Before/after image pairs of our clutter reduction. Each pair shows a case of (a) changing bristle orientation using topology, (b) cutting bristle lines crossing neighbor road lines, (c) circular roads, and (d) curved roads.

the edge and reducing the clutter. For neighboring two inward nodes (blue), we add a virtual outward node (red dotted circle) at the road line connecting two inward nodes resulting in splitting bristle lines on the road line. Similarly, a virtual inward node (blue dotted circle) is added for neighboring two outward nodes (red).

5.2 Avoid Crossing Neighbors

Another cluttering case is illustrated in Fig. 7c. When two road lines intersecting with less than a 90 degree angle have bristle lines, some of the bristle lines overlap as illustrated in Fig. 7c. For this case, we forbid bristle lines to cross neighbor road lines by placing the end point of a bristle line on the neighbor road line as shown in Fig. 7d. We first check the intersection of bristle blocks (colored boxes in Fig. 7) for the current road line on which we are generating bristle lines and its neighboring road lines by using the topology graph. If the blocks are intersected, we then check if a bristle line crosses the neighbor road lines by utilizing the intersection algorithm of 2D line segments [38]. This idea is based on the theory of amodal completion (or amodal perception) [39] in psychology that describes how the human visual system completes parts of an object even when it is only partially visible. Although the length of a bristle line represents data magnitude, benefits from cutting the length to avoid clutter dominate the side effects from data misunderstanding that could be caused by clutter. Moreover, when using redundant encoding utilizing bristle length and density as data magnitude, bristle density could help viewers complete parts of the bristle lines. Fig. 8 shows four image pairs before and after applying our clutter reduction strategies. Some improvements could also be considered in the future. For instance, our strategies still generate cluttered bristle lines in cases where road lines are very dense or close to others. We perform experiments in Section 6 to see how people understand the differences before and after clutter reduction. Here, we note that the experiments performed were for comparison and identification tasks. In these task types, line direction (as will be shown in the experiments) had little impact on the user results. However, in a cluster/delineate task in which users are asked to segment the data, the splitting of direction may influence the user's perception of cluster boundaries. As such, we recommend that map designers take caution in employing this scheme and use it only in appropriate map contexts. Future work will explore other schemes and design issues to handle neighbor crossings and influence on map design.

6 Evaluation

To evaluate the effectiveness of our bristle maps, we conducted two quantitative controlled experiments. These studies are both comprised of an introductory session, and a training session. In the first study, five tasks were conducted to evaluate the efficiency of bristle maps compared to existing visualization methods (point, color (kernel density estimated—KDE), and line maps as shown in Figs. 1a, 1c, and 1d) and post-task questionnaires for qualitative feedback. In the second study, two tasks were conducted to evaluate the accuracy of users in estimating values from each of the map types (point, KDE, bristle, and line) as well as evaluating the perceived aesthetics of each image. Prior to each study, a pilot study was also conducted to ensure that each task contains a fair comparison among the techniques.

Participants. In the first study, thirty graduate students (23 males, seven females) in engineering, science, and statistics from our university participated in the study. All participants reported that they had experience in visualizing data on geographical maps using colors or icons (e.g., paper maps, online map services). The experience varied from almost daily (11 participants), 1-2 times a week (17 participants) to 1-3 times a month (two participants). For the identification/accuracy tasks and aesthetic comparisons (Tasks 6 and 7), a secondary study was run on 26 undergraduate students in engineering from our university.

Apparatus. The experiment was performed on a 30″ monitor using our experimental application running on Windows XP, as shown in Fig. 9, where all visualizations were generated with 2,228×1,478 resolution. Each visualization was overlayed with numbered circles as shown in Fig. 9. Participants selected one of the numbers to answer the question in each trial using buttons in the interface panel on the top of the screen. Criminal incident reports collected in West Lafayette and Lafayette, Indiana from 1999 to 2010 were used in each trial, but different types of crimes were selected to generate visualizations in the training phase and in the actual study.

Design. We employed a repeated measure design of tasks incorporating variations of the images shown in Figs. 1a, 1c, and 1d and line maps similar to those of

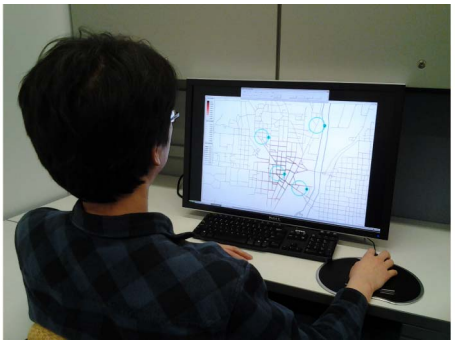


Fig. 9. Example setup for our experiment.

TABLE 2
The Number of Data Sets, Techniques (Cases in Fig. 8 for Task 5), and Trials

	Data sets	Techniques (or cases)	Trials
Task 1	18	5	90
Task 2	18	5	90
Task 3	15	6	90
Task 4	12	4	48
Task 5	7	8	56
Task 6	2	4	24
Task 7	2	4	2

Fig. 3(right). Table 2 shows the number of data sets, techniques (cases as shown in Fig. 8 for Task 5), and trials in each task. For example, in Task 1 we utilized 18 different data sets to compare five different techniques (i.e., point map, color map, line map, and bristle maps using two different encoding schemes). Hence, each participant performed $18\times5=90$ trials in Task 1. In Task 3, we compared six different techniques (i.e., point map, bivariate color map, line maps in two different encoding schemes, and bristle maps in two different encoding schemes) with 15 data sets, resulting in 90 trials. Due to the difficulty of creating good examples to be used from our real crime data, we used fewer data sets in Tasks 4 and 5. In summary, each participant performed a total of 374 trials in Tasks 1 to 5, and it generally took 90 minutes.

Since the design of Tasks 1-5 focused on questions of comparing regions, a secondary study was also conducted. This study was again a repeated measure design of tasks incorporating variations of the images shown in Figs. 1a, 1c, and 1d and line maps similar to those of Fig. 3(right). However, here the subjects were asked to identify the values of regions in the image. Areas of homogeneous visual variables were circled in each image and the subjects were asked to approximate the amount of crime per region. As a final task, the subjects were simultaneously presented with a point map, color map, bristle map, and line map and asked to rank order the images based on their aesthetic values.

For all Tasks, trial order was varied using a magic square method [40] in each task. Completion time and participants' answers were recorded for a quantitative metric. The collected data from each task was subjected to an analysis of variance (ANOVA) test to determine if the average time and accuracy of task completion were significantly different among techniques. A Post-Hoc Tukey HSD test was then performed to determine significance between the techniques. P-values reported in this study come from the resultant Tukey HSD test. Before the study, participants were introduced to our experiment application and the techniques through an introductory session and a training session. During the training session, participants could ask questions and receive guidance in the use of the experiment application and analysis of each visualization. Once the training was completed, participants moved to the actual study. After completing each task (Tasks 1 to 4) participants were asked to answer the questionnaire to rate the efficiency of the techniques using a five-point Likert scale [41]. After completing Task 5, participants were also asked to describe their impression with regards to visual complexity for before and after image pairs applying our clutter reduction. In the questionnaire, we stated that the visual complexity is high if a participant felt any kind of difficulty or confusion in understanding the density, length, and color of bristle lines that encode the underlying data. Finally, after finishing all tasks, participants were asked to rate the overall efficiency among techniques.

Hypotheses. In this experiment, we hypothesized that our bristle maps would be better than or equally as good when compared to the other techniques in terms of task completion time and accuracy. Specifically, we hypothesized that our bristle maps would be better than other techniques as the complexity level of tasks increased from univariate to multivariate. The rationale of this assumption is that the line map and bivariate color map use at most two variables, whereas the several encoding parameters in our bristle map have the potential to create effective encoding combinations. We also hypothesized that our clutter reduction strategies would be useful to minimize cluttering on areas where a large number of bristle lines are created. In our follow-on experiment exploring identification of values, we hypothesized that bristle maps would be as accurate as all other representations in determining values. We also hypothesized that bristle maps would be ranked higher in terms of their aesthetics.

Tasks. We tested seven tasks: three for univariate, bivariate, and multivariate data encoding, respectively, one for temporal variance encoding, one for the clutter reduction, one for accuracy comparisons among the rendering styles, and one for aesthetic comparisons.

In Task 1, when given four regions highlighted in circles on the map, participants were asked to "find the region with the highest crime rate" in different visualizations representing spatiotemporal crime data using point, color, line-T (data encoded in the line (T)hickness), bristle-CLD (a redundant data encoding using (C)olor, (L)ength, and (D)ensity), and bristle-LD (a redundant data encoding using (L)ength and (D)ensity).

In Task 2, four regions were highlighted in circles on the map. Participants were asked to "find the region with the highest crime rates at both (or either) day and night time," using point (encoding day/night time crime rates in different colors), color, line-TO (data encoded as line (T)hickness and using (O)rientation for day/night crime rates), bristle-CLDO (redundant data encoding using (C)olor, (L)ength ,and (D)ensity, and using (O)rientation to indicate day/night crime rates), and bristle-LDO (data encoded using (L)length and (D)density, but in a constant color, using (O)rientation to indicate day/night crime rates). The point map had differently colored points for day and night time crime rates, and two maps (day and night time color maps) were given in different colors for the color map.

In Task 3, four regions were highlighted in circles on the map. Participants were asked to "find the region with the highest crime rates for both (or either) two crimes (crime 1 and 2)," using point map (encoding two crimes in different colors), bivariate color map (Color-B), line-TO (a data encoding using (T)hickness in different colors, and using (O)rientation to indicate crime types), line-CT (encoding crime 1 using (C)olor and crime 2 using (T)hickness),

bristle-LDO (a redundant data encoding using (L)ength and (D)ensity, and using (O)rientation to indicate crime types), and bristle-CD (an encoding using (C)olor to indicate crime 1 and (D)ensity to indicate crime 2, with constant length).

In Task 4, participants were given two regions highlighted in circles on the map. Then, they were asked to "find the region with the highest temporal variance" in different visualizations using point maps, color maps, line maps, and bristle-LDV (a redundant data encoding using (L)ength and (D)ensity, and representing (V)ariance in the variance part of a bristle line). For the point, color, and line maps, multiple images were displayed on the screen to provide visualizations during several years. Our bristle map embedded the variance in the variance part of the bristle length as shown in Fig. 2 (third stage) and Fig. 5 (right column).

In Task 5, given two regions predefined in circles on bristle maps, participants were asked to "answer if crime rates on this given two regions look either different or the same as each other." Fig. 8 shows representative image pairs before and after applying our clutter reduction method. In trials, participants compared each case in Fig. 8 to a base case (i.e., bristle lines on a single straight road).

In Task 6, subjects were presented with a series of images with a single predefined circle, which covered an area consisting of homogeneous visual variables (i.e., identical color, bristle length, thickness, etc.). A univariate encoding was explored, and the Bristle-CLD settings were utilized for the bristle map. Participants were asked to estimate the amount of crime in the area using the provided scale (or scales in the case of bristle and line maps). Time and accuracy of the results were measured.

In Task 7, subjects were presented simultaneously with four images representing the same data set. These images consisted of a point map, a color map, a bristle map, and a line map. Subjects were asked to rank order the images in order of most to least aesthetically pleasing.

7 Results and Discussion

After all tasks were completed, times and answers collected during the study were analyzed using a single-factor ANOVA. A Post-Hoc Tukey HSD test was then performed to determine significance between the techniques. P-values reported in this study come from the resultant Tukey HSD test. For accuracy, the percentage of correct answers was computed.

Task 1. A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with highest crime rates within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps, and bristle maps. There was a significant effect of visualization type on time at the $p < 0.05$ level for the conditions $[F(4, 145) = 35.366, p = 0.0000001]$ and a significant effect of visualization type on accuracy at the $p < 0.05$ level for the conditions $[F(4, 145) = 3266.782, p = 0.0000000006]$. Because statistically significant results were found, we computed a Tukey posthoc test with results reported in Table 3. In Table 3, p-values < 0.05 indicate that groups were statistically different from one another.

TABLE 3
Tukey HSD Results for Task 1

	p-value $<$	Point map	KDE map	Line-T
Time	Bristle-CLD	**.00001**	**.00001**	**.00042**
	Bristle-LD	**.00001**	**.00001**	**.01811**
	p-value $<$	Point map	KDE map	Line-T
Accuracy	Bristle-CLD	**.00001**	.1554	**.01851**
	Bristle-LD	**.00001**	.3214	**.00602**

The result showed that the bristle maps groups were both significantly different than the point, color, and line maps in terms of speed (at the $p < 0.05$ level). Specifically, the bristle map groups average times were 50.7 and 56.6 seconds for the CLD and LD conditions, respectively, which was slightly faster than the Line-T condition at 69 seconds and much faster than the point map condition at 102.6 seconds. However, the color map group was the fastest at 34.6 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group in terms of accuracy (at the $p < 0.05$ level). Specifically, the bristle map groups accuracy ratings were 99.6 and 99.8 percent for the CLD and LD conditions, respectively, which was much higher than the point map condition with accuracy of 41.4 percent. No accuracy differences were found when compared to the other groups. See Table 8 for more specific results.

The comparison between color maps and bristle maps showed that color maps were better than the bristle map in terms of average time, and were not significantly different in terms of accuracy. This shows that bristle maps as a redundant encoding scheme has the same potential to convey data as single parameter encoding schemes; however, traditional schemes such as color maps may allow for a quicker comparison in the univariate case.

Comparing Bristle-LD and Line-T, we saw that the length of the bristle map matches the thickness of the line map. Hence, the bristle density was useful to find answers in Task 1 in terms of completion time and accuracy. Some participants also mentioned bristle density in their qualitative feedback as *"Bristle map is especially good when density of the bristles is also used"* and *"In bristle map, length, and density were more noticeable than color difference."* In this univariate encoding test, the point map showed the worst results and the color map was the best results in terms of time and accuracy as shown in Table 8.

Task 2. A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with highest crime rates at both day and nighttime within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps and bristle maps. There was a significant effect of visualization type on time at the $p < 0.05$ level for the conditions $[F(4, 145) = 2.717, p = 0.032]$ and a significant effect of visualization type on accuracy at the $p < 0.05$ level for the conditions $[F(4, 145) = 89.89, p = 0.0000002]$. Because statistically significant results were found, we

TABLE 4
Tukey HSD Results for Task 2

	p-value <	Point map	KDE map	Line-TO
Time	Bristle-CLDO	**.01713**	.70091	.05943
	Bristle-LDO	**.02024**	.81621	.07166
	p-value <	Point map	KDE map	Line-TO
Accuracy	Bristle-CLDO	**.00001**	.07062	.36692
	Bristle-LDO	**.00001**	.99999	**.01283**

TABLE 5
Tukey HSD Results for Task 3

	p-value <	Point map	KDE-B	Line-TO	Line-CT
Time	Bristle-LDO	**.00009**	**.00515**	.58128	.73239
	Bristle-CD	**.01131**	**.03469**	.15506	.20693
	p-value <	Point map	KDE-B	Line-TO	Line-CT
Accuracy	Bristle-LDO	**.00001**	**.00001**	.27189	**.02771**
	Bristle-CD	**.00001**	**.00001**	.07002	.41194

computed a Tukey posthoc test with results reported in Table 4. In Table 4, p-values < 0.05 indicate that groups were statistically different from one another.

As we hypothesized, the result showed that the bristle maps groups were both significantly different than the point maps in terms of speed (at the $p < 0.05$ level). Specifically, the bristle map groups average times were 86.3 and 87.2 seconds for the CLDO and LDO conditions, respectively, which was slightly faster than the point map condition at 106.2 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group in terms of accuracy (at the $p < 0.05$ level). Specifically, the bristle map groups accuracy ratings were 90.5 and 93.3 percent for the CLDO and LDO conditions, respectively, which was much higher than the point map condition with accuracy of 63.1 percent. See Table 8 for more specific results.

The comparison between color maps and bristle maps showed that color maps were better than the bristle map in terms of average time, and were not significantly different in terms of accuracy. This shows that bristle maps as a reduandant encoding scheme has the same potential to convey data as single parameter encoding schemes; however, traditional schemes such as color maps may allow for a quicker comparison in the univariate case.

Findings also indicated that Bristle-LDO was better than Line-TO in terms of accuracy, whereas Bristle-CLDO was not significantly different from Line-TO in terms of accuracy. This indicated that the bristle density seems to be useful in finding correct answers in Bristle-LDO, but it was not in Bristle-CLDO. Further testing in combinations of visual variables and the ability to determine levels of sparseness will be done in the future.

Task 3. A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with highest crime rates in two types of crimes within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps, and bristle maps. There was a significant effect of visualization type on time at the $p < 0.05$ level for the conditions $[F(5, 174) = 6.655, p = 0.00001]$ and a significant effect of visualization type on accuracy at the $p < 0.05$ level for the conditions $[F(5, 175) = 144.24, p = 0.00000001]$. Because statistically significant results were found, we computed a Tukey posthoc test with results reported in Table 5. In Table 5, p-values < 0.05 indicate that groups were statistically different from one another.

The result showed that the bristle maps groups were both significantly different than the point maps and color maps in terms of speed (at the $p < 0.05$ level). Specifically, the bristle map groups average times were 88.2 and 94.5 seconds for the LDO and CD conditions, respectively, which was faster than the point map condition at 118.3 seconds and the color map condition at 115.3 seconds.

For accuracy, the bristle maps groups were both significantly different than the point map group and the color map group in terms of accuracy (at the $p < 0.05$ level). Specifically, the bristle map groups accuracy ratings were 94.4 and 90.4 percent for the LDO and CD conditions, respectively, which was much higher than the point map condition with accuracy of 26.6 percent and the color map condition with accuracy of 72.6 percent. See Table 8 for more specific results.

Note that we separated parameters for different crime types in Bristle-CD: (C)olor encodes crime 1 and (D)ensity encodes crime 2. Bristle-CD showed a significant effect compared to the bivariate color map as shown in Table 5. However, generation on this type of bristle maps should be selected carefully because one parameter could dominate the other. For instance, when we use color and length to separate two crime data, short bristle length for low crime rates in crime 2 removes bristle lines in dark color for high crime rates in crime 1. In our experiment, we selected color and density for two crimes, with constant length of bristles.

Task 4. A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with high temporal variance within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps, and bristle maps. There was a significant effect of visualization type on time at the $p < 0.05$ level for the conditions $[F(3, 116) = 42.051, p = 0.00001]$ and a significant effect of visualization type on accuracy at the $p < 0.05$ level for the conditions $[F(3, 116) = 42.33, p = 0.00001]$. Because statistically significant results were found, we computed a Tukey posthoc test with results reported in Table 6. In Table 6, p-values < 0.05 indicate that groups were statistically different from one another.

The result showed that the bristle maps groups were both significantly different than the point maps, line maps and color maps in terms of speed (at the $p < 0.05$ level). Specifically, the bristle map groups average time was 48.4 seconds for the LDV condition, which was faster than the point map condition at 194 seconds, the color map

TABLE 6
Tukey HSD Results for Task 4

Time	p-value $<$	Point maps	KDE maps	Line maps
	Bristle-LDV	**.00001**	**.00001**	**.00001**
Accuracy	p-value $<$	Point maps	KDE maps	Line maps
	Bristle-LDV	**.00001**	**.00001**	**.00001**

TABLE 7
Average Rank Ordering by Aesthetics

	Point	KDE Map	Bristle	Line
Average	2.26	2.6	2.79	2.34
Std Dev	1.18	.97	1.19	1.10

condition at 171.8 seconds, and the line map condition at 178.9 seconds.

For accuracy, the bristle maps groups were both significantly different than the point maps, line maps, and color maps in terms of speed (at the $p < 0.05$ level). Specifically, the bristle map groups accuracy rating was 94.7 percent for the LDV condition, which was much higher than the point map condition with accuracy of 53.6 percent, the color map condition with accuracy of 72.6 percent, and the line map condition with accuracy of 75.5 percent. See Table 8 for more specific results.

As we hypothesized, we found that the representation of temporal variance in bristle maps was significantly faster and accurate in terms of both average time and accuracy compared to providing several images of the point, color, and line maps. Moreover, we found that techniques showed the increasing pattern from the point maps to Bristle-LDV as shown in Table 8. This indicates that changes among several images would be better perceived in line patterns than in points or colors.

Task 5. A one-way between-subjects ANOVA was conducted to compare the effect of different map visualizations on a subject's time and accuracy in determining areas with high temporal variance within a given visualization. Conditions varied based on the given visualization, point maps, kernel density estimated color maps, line maps, and bristle maps. There was no significant effect of visualization type on time at the $p < 0.05$ level for the conditions $[F(1, 56) = 0.328, p = 0.569]$ and no significant effect of visualization type on accuracy at the $p < 0.05$ level for the conditions $[F(1, 56) = 0.315, p = 0.315]$. In Task 5, we found that bristle lines with and without clutter reduction did not differ significantly w.r.t. both average time and accuracy for all cases (Fig. 8). This means that the base bristle lines and bristle lines before applying clutter reduction and the base and bristle lines after applying our clutter reduction are perceived similarly by participants. Moreover, when told that the bristle line orientation does not encode data, the opposite orientations of bristle lines on a single straight road caused by virtual nodes (Fig. 7b) did not affect accuracy (87.7 percent). Other cases showed 42-58 percent of accuracy.

Task 6. For Task 6, we hypothesized that subjects would be as accurate as all other representations in determining

TABLE 8
Average Time and Accuracy

	Technique	Average time (seconds)	Accuracy (%)
Task 1	Point map	102.6	41.4
	KDE map	**34.6**	**100**
	Line-T	69	98
	Bristle-CLD	50.7	99.6
	Bristle-LD	56.6	99.8
Task 2	Point map	106.2	63.1
	KDE map	90	93.1
	Line-TO	100.5	87.9
	Bristle-CLDO	**86.3**	90.5
	Bristle-LDO	87.2	**93.3**
Task 3	Point map	118.3	26.6
	KDE-B	115.3	72.6
	Line-TO	**84**	**96.4**
	Line-CT	86.1	86.8
	Bristle-LDO	88.2	94.4
	Bristle-CD	94.5	90.4
Task 4	Point maps	194	53.6
	KDE maps	171.8	61.9
	Line maps	178.9	75.5
	Bristle-LDV	**48.4**	**94.7**

values. In Task 6, we found that bristle maps did not differ significantly w.r.t. accuracy when compared with point map, color map, and line map identification (ANOVA results of p-value $= 0.18093$, $F = 1.63$). However, we found that bristle maps did differ significantly w.r.t. time when compared with point map, color map and line map (ANOVA results of p-value $= 0.0314$, $F = 2.622$). Particularly, we found line maps and heat maps to both be significantly faster than point maps and bristle maps in identifying values (Tukey HSD test value of $p < 0.05$). Overall, these results indicate that in terms of accuracy, all geographical representations were equally useful; however, participants were (on average) over 1 second quicker in value judgments on both line maps and colors maps. This is most likely due to the fact that participants were quicker at making color judgments as compared to counting points and mentally linking multiple variables for the bristle maps.

Task 7. In Task 7, we found that users had a highly variable rating of which image appeared to be more aesthetically pleasing. The average positions and standard deviations are summarized in Table 7. Here, we find that while bristle maps have a slightly higher average ranking, there is no significant difference between the aesthetic ordering. A one-way between-subjects ANOVA was conducted to compare the rankings of map visualizations by subject in determining which visualization was ranked highest in aesthetics. There was no significant effect of visualization type on aesthetics at the $p < 0.05$ level for the conditions $[F(3, 183) = 1.79, p = 0.149]$.

Qualitative evaluation. Fig. 10 shows the results from qualitative feedback. Among the 30 participants, 27 participants (90 percent) agreed or strongly agreed that the bristle map was efficient for day and night time comparison in Task 2, 26 for two color maps and 23 for line map. Twenty-four participants (80 percent) agreed or strongly agreed that the bristle map was efficient for the comparison of two crimes in Task 3, 26 for the line map and 19 for the

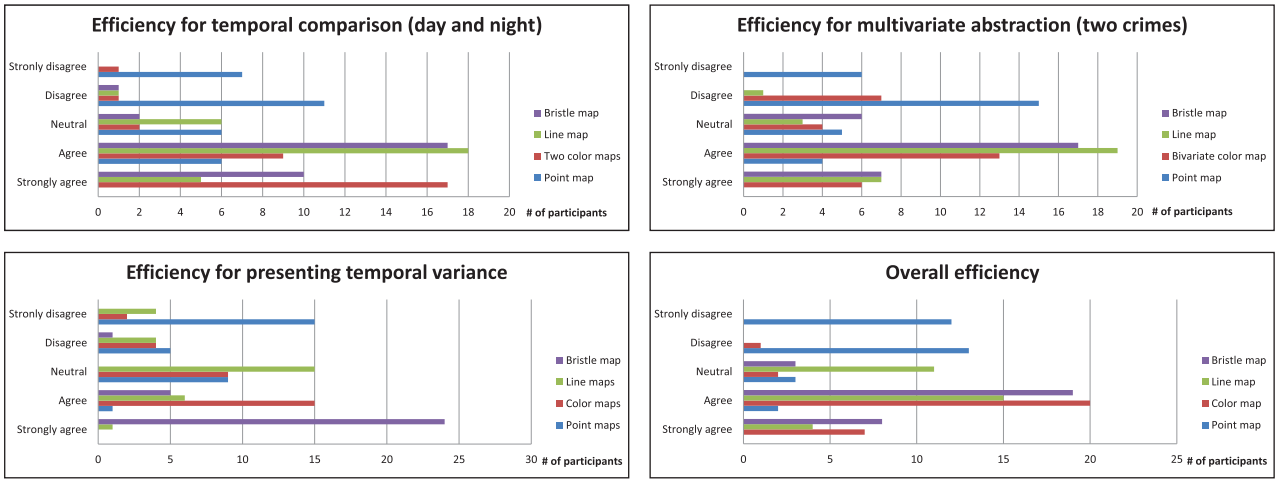


Fig. 10. Results from qualitative feedback for Tasks 2, 3, 4, and overall efficiency.

bivariate map. Twenty-nine participants (96.6 percent) agreed or strongly agreed that the bristle map was efficient for temporal variance representation. In the question for overall efficiency, 27 participants (90 percent) agreed or strongly agreed that bristle maps and color maps were overall efficient, and 19 (63.3 percent) for line maps. For point maps, 25 participants (83.3 percent) disagreed or strongly disagreed.

Participants were also asked to answer visual complexity and preference questions regarding the before (NCR) and after (CR) image pairs applying our clutter reduction. For the circular case (Fig. 8c), 96.5 percent of participants felt that NCR has higher visual complexity and 78.5 percent preferred CR. For the curved case (Fig. 8d), 65.5 percent of participants answered that CR has a higher visual complexity and 64 percent preferred NCR. While both cases use a technically identical clutter reduction algorithm, participants reported different visual complexity and preference for them. This indicates that our clutter reduction could be improved by considering the complexity of the underlying network structure.

Summary and Limitations. As a univariate encoding, the bristle maps were significantly different (in terms of speed and accuracy) than the point, color, and line maps. In the case of the point and line maps, bristle maps use resulted in a higher average correctness and speed; however, the color map for the univariate case had the fastest response and accuracy totals. This seems to indicate that the redundant encoding scheme is actually not beneficial in these cases. As such, use of bristle maps for single variable encoding is not recommended.

With regards to bivariate and multivariate encoding, bristle maps and line maps outperformed color and point maps. This is not surprising as bristle and line maps are able to combine variables into a single image, whereas in the case of point and color maps, the user must mentally combine the two images together. Bristle-(C)LD also showed a significant effect of the bristle density compared to Line-T. As a bivariate encoding, using orientation in bristle maps was not significant compared to two color maps. However, in the comparison with the bivariate color map, Bristle-LDO showed a significant effect in terms of average time and accuracy. As such, we have that Bristle-(C)LDO as a bivariate encoding scheme created a middle level of cognitive load in-between two color maps and a bivariate color map. Bristle maps also showed potential as a multivariate encoding technique in a single view. Based on the results in Task 3, a point map using various colors and a multivariate color map would considerably increase users' cognitive load. In Tasks 1-3, we also observed that there is no significant effect between the bristle maps using the different encodings. The representation of temporal variance in the bristle map was significantly different from other methods. Our results also showed the differences among point, color, and line maps. Participants could better find the region with higher temporal variance when using line maps than using point and color maps. In the qualitative evaluation, 90 percent of the participants agreed or strongly agreed the overall efficiency of bristle maps to find answers. However, users also strongly preferred the color map in these cases as well.

Finally, we found that with regards to accuracy in identifying values, no technique outperformed any others. However, users were significantly faster in identifying values in both the color and line map scenarios. We hypothesize that in both cases the user focused only on the color, whereas in the point map case they needed to count the points and in the bristle map case they needed to reconfirm the univariate value by double checking several of the encoding legends.

Overall, this technique would be recommended when encoding large amounts of multivariate spatiotemporal point data. As the number of point samples increase, aggregation techniques are need to allow for quick summaries of the data, and, as is evidenced by our studies, pure spatial location representation by glyphs results in too much overlap for accurate measurement and evaluation. As the number of variables increase, color map representations allow for the encoding of variables only along a single visual variable (resulting in bivariate color maps or small multiple plots).

In using multivariate encodings, it is extremely important to understand the interaction effects that the visual variables will introduce in one another. Research into the perceptual interactions among different visual variables was performed by Acevedo and Laidlaw [42]. They

measured the perceptual interference of icon size, spacing, and brightness, noting that brightness outperforms spacing and size while being subject to interferences from both spacing and size. Acevedo and Laidlaw also noted that spacing also outperformed size, which contradicted some previous results; however, this result seems to align with our participants noting that the bristle spacing was a useful cue. Their results were reportedly due to the spacing sampling along a sinusoidal curve. The sampling of our bristles follows a uniform pattern within classification bins. Thus, there seems to be sufficient scientific evidence to justify using sparsity as a discriminating variable in the case of the bristle maps; however, further studies on this are warranted. Stone [43] has also studied the effect of size in color perception, noting that color appearance changed dramatically with the size being viewed. As such, it may be better to utilize fewer map classifications (color bins) when using bristle maps to increase the perceptual distance between each color being visualized.

The main limitations of the bristle map technique is that the combinations of data encoding can potentially prove overwhelming for the designer, and a poor choice on variable encoding can result in a suboptimal visualization. In particular, previous studies have provided results that can be used to predict that certain combinations of visual variables will either enhance or impede map reading. For example, the combination of length and density form an emergent property akin to Bertin's definition of grain. Such effects cannot be ignored; however, bristle maps can be encoded to take advantage of such combinations, as shown in Tasks 3 and 4.

Finally, with regards to scalability of the bristle map technique, in areas of dense roadways, different aggregation methods would need to be considered. As the roads become dense, the ability to plot lines of perceptually different length would become untenable. However, a solution to this would be to draw only the most important roads, thereby removing smaller roads from the analysis, or utilizing bristle maps in a focus+context manner.

8 Conclusions

In this work, we have described our novel multivariate data encoding scheme, the Bristle Map. This scheme provides a novel approach for encoding color, length, density, and orientation as data variables and allowing the user to explore correlations within and between variables on a single view. Given the number of parameters available within this encoding, this article has presented only a subset of potential encodings and examples. Here, we have shown the use of encoding bristle lines with redundant information, multivariate attributes for variable comparison, and temporal variance. We also showed a means of potentially encoding data uncertainty. To minimize overlap of bristle lines, we generated a topology graph from underlying geographical line features and employed strategies for clutter reduction. Then, to evaluate the effectiveness of bristle maps, we performed an evaluation study, where we explored different visual encoding combinations within the bristle maps and compared with existing techniques in several tasks. Based on our experiment results, we believe that our bristle map technique has much potential to increase the amount of information that can be visualized on a single map for geovisualization.

Acknowledgments

The authors would like to thank Ahmad M. Razip for his help in setting up a web-based user study environment. This work was supported by the US Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. Jang's work was supported in part by the Industrial Strategic technology development program, 10041772, funded by the Ministry of Knowledge Economy (MKE, Korea). Isenberg's work was supported in part by a French DIGITEO chair of excellence.

References

[1] A. MacEachren, D. Xiping, F. Hardisty, D. Guo, and G. Lengerich, "Exploring High-D Spaces with Multiform Matrices and Small Multiples," *Proc. IEEE Symp. Information Visualization (InfoVis)*, pp. 31-38, 2003, IEEE CS, doi: 10.1109/INFVIS.2003.1249006.

[2] C. North and B. Shneiderman, "Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata," *Proc. Working Conf. Advanced Visual Interfaces (AVI)*, pp. 128-135, 2000, ACM, doi: 10.1145/345513.345282.

[3] C. Weaver, "Cross-Filtered Views for Multidimensional Visual Analysis," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 2, pp. 192-204, Mar./Apr. 2010, doi: 10.1109/TVCG.2009.94.

[4] D.S. Ebert, R.M. Rohrer, C.D. Shaw, P. Panda, J.M. Kukla, and D.A. Roberts, "Procedural Shape Generation for Multi-Dimensional Data Visualization," *Computers & Graphics*, vol. 24, no. 3, pp. 375-384, June 2000, doi: 10.1016/S0097-8493(00)00033-9.

[5] S. Bachthaler and D. Weiskopf, "Continuous Scatterplots," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1428-1435, Nov./Dec. 2008, doi: 10.1109/TVCG.2008.119.

[6] R. Maciejewski, S. Rudolph, R. Hafen, A.M. Abusalah, M. Yakout, M. Ouzzani, W.S. Cleveland, S.J. Grannis, and D.S. Ebert, "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205-220, Mar./Apr. 2010, doi: 10.1109/TVCG.2009.100.

[7] J.J. van Wijk and A. Telea, "Enridged Contour Maps," *Proc. Conf. Visualization (VIS)*, pp. 69-74, 2001, IEEE CS, doi: 10.1109/VISUAL.2001.964495.

[8] R.J. Phillips and L. Noyes, "An Investigation of Visual Clutter in the Topographic Base of a Geological Map," *Cartographic J.*, vol. 19, no. 2, pp. 122-132, Dec. 1982, doi: 10.1179/000870482787073225.

[9] S. Openshaw, "The Modifiable Areal Unit Problem," *Concepts and Techniques in Modern Geography*, vol. 38, Geo Books, 1984.

[10] M. Swink and C. Speier, "Presenting Geographic Information: Effects of Data Aggregation, Dispersion, and Users' Spatial Orientation," *Decision Sciences*, vol. 30, no. 1, pp. 169-195, Jan. 1999, doi: 10.1111/j.1540-5915.1999.tb01605.x.

[11] C.L. Eicher and C.A. Brewer, "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation," *Cartography and Geographic Information Science*, vol. 28, no. 2, pp. 125-138, Apr. 2001, doi: 10.1559/152304001782173727.

[12] R. Spence, *Information Visualization*. Addison-Wesley, 2001.

[13] L. Wilkinson, *The Grammar of Graphics*, second ed. Springer-Verlag, 2005.

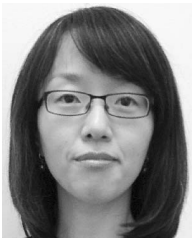
[14] A.M. MacEachren, *How Maps Work: Representation, Visualization, and Design*. Guilford Press, 1995.

[15] S. Chainey, L. Tompson, and S. Uhlig, "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security J.*, vol. 21, no. 1/2, pp. 4-28, Feb.-Apr. 2008, doi: 10.1057/palgrave.sj.8350066.

[16] B.W. Silverman, *Density Estimation for Statistics and Data Analysis (Monographs on Statistics and Applied Probability)*, vol. 26, Chapman and Hall, 1986.

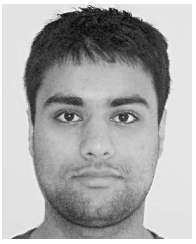
[17] C. Ahlberg and B. Shneiderman, "Visual Information Seeking Using the FilmFinder," *Proc. Conf. Companion Human Factors in Computing Systems (CHI)*, pp. 433-434, 1994, ACM, doi: 10.1145/259963.260431.

[18] Y.-H. Fua, M.O. Ward, and E.A. Rundensteiner, "Structure-Based Brushes: A Mechanism for Navigating Hierarchically Organized Data and Information Spaces," *IEEE Trans. Visualization and Computer Graphics,* vol. 6, no. 2, pp. 150-159, Apr.-June 2000, doi: 10.1109/2945.856996.
[19] A. Dix and G. Ellis, "By Chance: Enhancing Interaction with Large Data Sets through Statistical Sampling," *Proc. Working Conf. Advanced Visual Interfaces (AVI),* pp. 167-176, 2002, ACM, doi: 10.1145/1556262.1556289.
[20] A. MacEachren, "Visualizing Uncertain Information," *Cartographic Perspectives,* vol. 13, pp. 10-19, 1992.
[21] R. Dunn, "A Dynamic Approach to Two-Variable Color Mapping," *The Am. Statistician,* vol. 43, no. 4, pp. 245-252, Nov. 1989, doi: 10.1080/00031305.1989.10475669.
[22] J. Olson, "Spectrally Encoded Two-Variable Maps," *Annals Assoc. Am. Geographers,* vol. 71, no. 2, pp. 259-276, June 1981, doi: 10.1111/j.1467-8306.1981.tb01352.x.
[23] A. MacEachren and D. DiBiase, "Animated Maps of Aggregate Data: Conceptual and Practical Problems," *Cartography and Geographic Information Systems,* vol. 18, no. 4, pp. 221-229, Oct. 1991, doi: 10.1559/152304091783786790.
[24] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey, "Weaving Versus Blending: A Quantitative Assessment of the Information Carrying Capacities of Two Alternative Methods for Conveying Multivariate Data with Color," *IEEE Trans. Visualization and Computer Graphics,* vol. 13, no. 6, pp. 1270-1277, Nov./Dec. 2007, doi: 10.1109/TVCG.2007.70623.
[25] T. Saito, H.N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda, "Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data," *Proc. IEEE Symp. Information Visualization (InfoVis),* pp. 173-180, 2005, IEEE CS, doi: 10.1109/INFOVIS.2005.35.
[26] M. Sips, J. Schneidewind, D.A. Keim, and H. Schumann, "Scalable Pixel-Based Visual Interfaces: Challenges and Solutions," *Proc. 10th Int'l Conf. Information Visualization (IV),* pp. 32-38, 2006, IEEE CS, doi: 10.1109/IV.2006.95.
[27] C. Panse, M. Sips, D. Keim, and S. North, "Visualization of Geo-Spatial Point Sets via Global Shape Transformation and Local Pixel Placement," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 5, pp. 749-756, Sep./Oct. 2006, doi: 10.1109/TVCG.2006.198.
[28] D. Dorling, A. Barford, and M. Newman, "Worldmapper: The World as You've Never Seen It Before," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 5, pp. 757-764, Sep./Oct. 2006, doi: 10.1109/TVCG.2006.202.
[29] P.C. Wong, K. Schneider, P. Mackey, H. Foote, G. Chin, R. Guttromson, and J. Thomas, "A Novel Visualization Technique for Electric Power Grid Analytics," *IEEE Trans. Visualization and Computer Graphics,* vol. 15, no. 3, pp. 410-423, May/June 2009, doi: 10.1109/TVCG.2008.197.
[30] D. Fisher, "Hotmap: Looking at Geographic Attention," *IEEE Trans. Visualization and Computer Graphics,* vol. 13, no. 6, pp. 1184-1191, Nov./Dec. 2007, doi: 10.1109/TVCG.2007.70561.
[31] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D Information Visualization for Time Dependent Data on Maps," *Proc. Ninth Int'l Conf. Information Visualization (InfoVis),* pp. 175-181, 2005, IEEE CS, doi: 10.1109/IV.2005.3.
[32] J. Bertin, *Semiology of Graphics.* ESRI Press, 2011.
[33] J. Tarbell, "Substrate," *Web Site & Simulation,* http://www.complexification.net/gallery/machines/substrate/, 2003, Feb. 2012.
[34] T. Isenberg, "Visual Abstraction and Stylisation of Maps," *Cartographic J.,* vol. 50, no. 1, pp. 8-18, Feb. 2013, doi: 10.1179/1743277412Y.0000000007.
[35] P. Rheingans, "Task-Based Color Scale Design," *Proc. SPIE,* vol. 3905, pp. 35-43, 2000, SPIE, doi: 10.1117/12.384882.
[36] C. Ware, "Color Sequences for Univariate Maps: Theory, Experiments and Principles," *IEEE Computer Graphics and Applications,* vol. 19, no. 5, pp. 41-49, Sep./Oct. 1988, doi: 10.1109/38.7760.
[37] C.A. Brewer, *Designing Better Maps: A Guide for GIS Users.* ESRI Press, 2005.
[38] M. Prasad, "Intersection of Line Segments," *Graphics Gems II,* J. Arvo, ed., pp. 7-9, Academic Press, 1991.
[39] A. Michotte, G. Thinès, and G. Crabbé, *Les Complements Amodeux des Structures Perceptives (Amodal Completions of Perceptual Structures),* Louvain: Institut de Psychologie del'Université de Louvain, France: Studia Psychologica, 1964.
[40] M.S. Farrar, *Magic Squares.* BookSurge Publishing, 1996.
[41] R.A. Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology,* vol. 22, no. 140, pp. 5-55, 1932.
[42] D. Acevedo and D. Laidlaw, "Subjective Quantification of Perceptual Interactions among Some 2D Scientific Visualization Methods," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 5, pp. 1133-1140, Sept. 2006, doi: 10.1109/TVCG.2006.180.
[43] M. Stone, "In Color Perception, Size Matters," *IEEE Computer Graphics & Applications,* vol. 32, no. 2, pp. 8-13, Mar./Apr. 2012, doi: 10.1109/MCG.2012.37.

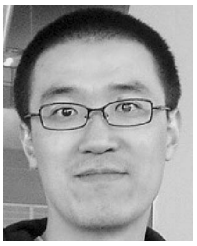


SungYe Kim received the master's degree in computer science and engineering from Chung-Ang University, South Korea in 2000, and the PhD degree in electrical and computer engineering from Purdue University in May, 2012. She is currently a graphics software engineer at Intel Corporation. Prior to this, she was employed as a research engineer at the Electronics and Telecommunications Research Institute from 2000 to 2006. Her research interests include computer graphics, illustrative visualization, visual analytics, and information visualization.

Ross Maciejewski received the PhD degree in electrical and computer engineering from Purdue University in December, 2009. He is currently an assistant professor at Arizona State University in the School of Computing, Informatics & Decision Systems Engineering. Prior to this, he served as a visiting assistant professor at Purdue University and was at the Department of Homeland Security Center of Excellence for Command Control and Interoperability in the Visual Analytics for Command, Control, and Interoperability Environments (VACCINE) group. His research interests include geovisualization, visual analytics, and nonphotorealistic rendering. He is a member of the IEEE.



Abish Malik received the BS degree in electrical engineering from Purdue University in 2009, and is currently working toward the PhD degree in the School of Electrical and Computer Engineering at Purdue University. He is a research assistant at the Purdue University Rendering and Perception Lab. His research interests include visual analytics, correlation, and predictive data analytics.



Yun Jang received the bachelor's degree in electrical engineering from Seoul National University, South Korea in 2000, and the master's and doctoral degree in electrical and computer engineering from Purdue University in 2002 and 2007, respectively. He is an assistant professor of computer engineering at Sejong University, Seoul, South Korea. He was a postdoctoral researcher at CSCS and ETH Zürich, Switzerland from 2007 to 2011. His research interests include interactive visualization, volume rendering, visual analytics, and data representations with functions. He is a member of the IEEE.

David S. Ebert received the PhD degree in computer science from Ohio State University. He is a professor in the School of Electrical and Computer Engineering at Purdue University, the University Faculty scholar, the director of the Purdue University Rendering and Perceptualization Lab, and the director of the Purdue University Regional Visualization and Analytics Center. His research interests include novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually based visualization, illustrative visualization, and procedural abstraction of complex, massive data. He is a fellow of the IEEE and a member of the IEEE Computer Society's Publications Board.

Tobias Isenberg received the doctoral degree from the University of Magdeburg, Germany. He is a senior research scientist with INRIA in France. Previously, he held positions as assistant professor for computer graphics and interactive systems at the University of Groningen, the Netherlands, and as a postdoctoral fellow at the University of Calgary, Canada. He works on topics in interactive nonphotorealistic and illustrative rendering as well as computational aesthetics and explores applications in scientific visualization. He is a member of the IEEE.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 pp109-110
Volume 18, Number 2, April 2013

Guest Editorial: Special Issue on Visualization and Visual Analytics

Aidong Lu*, David Ebert, Jinzhu Gao, Song Zhang, Alark Joshi

This special issue is devoted to the new research addressing challenges in the areas of visualization and visual analytics. Visualization and visual analytics are closely related research areas, both concentrating on developing visual techniques to reveal meaningful information out of various data in real-life applications. Visualization as a field has its roots in Computer Graphics and has become a popular research area over the years. The field of visual analytics is relatively young with a concentration on analytical reasoning facilitated by interactive visual interfaces. In general, visualization and visual analytics research is tightly connected with certain types of data or applications and researchers in both fields strive to discover known or unknown data patterns for domain users.

Since Tsinghua Science and Technology adjusted the scope to information technology in 2011, this is the second special issue on topics of Visualization and Visual Analytics. All the articles published in this special issue are selected through the open call. All the submissions have gone through the blind peer-reviewed process. We finally accepted eight research articles, five from the United States and three from China. The articles cover various topics on visualization and visual analytics. Applications include geo-spatial visualization, flow visualization, molecule visualization, online log visualization, microblogging visualization, image transition, edge-bundling, and insight management.

In the following, we roughly divide the eight accepted articles to three categories: scientific visualization, information visualization, and visual analytics. The readers may find that almost all the visualization articles involve components of data visualization, interactive exploration, and analytical reasoning components. The concentrations of each approach and the details related to visualization applications are different.

A brief overview of the accepted articles is given below.

Scientific Visualization

Climate research produces a wealth of multivariate data. In "An Interactive Visual Analytics Framework for Multi-Field Data in a Geo-Spatial Context", Zhang et al. present a framework for studying multi-field climate data. Several visualization and interaction techniques, such as fixed-window brushing and correlation-enhanced display, are presented and integrated with the Google Earth platform. The system has been tested by a team of climate researchers, who made a few important discoveries using it.

Turbulent flows are intrinsic to many processes in science and engineering, however the complex, non-linear interactions between individual eddies in these flows are hard to identify and quantify across multiple scales. In "Methods to Identify Individual Eddy Structures in Turbulent Flow", Wang et al. present several novel approaches for accurately segmenting individual eddy structures in turbulent flows. These methods can help quantify information of a flow at the level of individual structures and automatically track the evolution and interaction of large numbers of individual vortices in a complicated turbulent flow.

Atmospheric nucleation serves as a significantly

• Aidong Lu is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. E-mail: aidonglu@gmail.com.
• David Ebert is with School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA.
• Jinzhu Gao is with the Department of Computer Science, University of the Pacific, Stockton, CA 95211, USA. E-mail: jgao@pacific.edu.
• Song Zhang is with the Department of Computer Science and Engineering, Mississippi State University, Starkville, MS 39762, USA.
• Alark Joshi is with the Department of Computer Science, Boise State University, Boise, ID 83725, USA.
∗ To whom correspondence should be addressed.
Manuscript received :2013-03-24; accepted : 2013-03-24

important role in many atmospheric and technological processes. In "Similarity-Based 3-D Atmospheric Nucleation Data Visualization and Analysis", Zhu et al. present a data visualization solution with a novel algorithm for calculating similarity between the 3-D molecular crystals to visualize and classify 3-D molecular crystals effectively. The overall performance of the visualization system has been further improved with GPU acceleration.

Information Visualization

Real-time data log visualization is challenge due to the data complexity: it is streaming, hierarchical, heterogeneous, and multi-sourced. In "An Online Visualization System for Streaming Log Data of Computing Clusters", Xia et al. present a two-stage streaming process to visualize the log data generated by computing clusters. The visualization supported by a visual computing processor consists of a set of multivariate and time variant visualization techniques. The effectiveness and scalability of the proposed system framework are demonstrated on a commodity cloud-computing platform.

Microblogging, similar to Twitter, has provided a popular communication scheme for Web users to share information and express opinions. In "Portraying User Life Status from Microblogging Posts", Tang et al. presented an interactive visualization, LifeCircle, to explore behaviors of microblog users. The approach tightly integrates interactive visualization with novel and state-of-the-art microblogging analytics. Data from Sina Weibo has been used in the case studies and the results demonstrate the approach provides a quick summary of user life status for potential personal users and commercial services.

Transition in many information visualization applications can help users perceive changes and understand the underlying data. In "A Study of Animated Transition in Similarity-Based Tiled Image Layout", Zhang et al. investigate the effectiveness of animated transition in a tiled image layout with spiral arrangement. Based on three aspects of animated transition, an integrated solution, called AniMap, is presented for animating the transition between layouts during query processes. The effectiveness of the animated transition solution has been demonstrated by experimental results and a comparative user study.

Edge bundling has been a popular approach as edge is an important visual primitive for encoding data in information visualization research. In "Edge Bundling in Information Visualization", Zhou et al. first provide a survey on edge-bundling techniques for reducing visual clutter problem in visualization. They have reviewed the cost-based, geometry-based, and image-based edge-bundling methods designed for graphs, parallel coordinates, and flow maps. They also describe a number of visualization applications using edge-bundling techniques, discuss the evaluation studies on the effectiveness of edge-bundling methods, and point out some future research directions.

Visual Analytics

Significant progress has been made toward effective insights discovery in visual analytics systems, while managing large amounts of insights generated in visual analytics processes is also important. In "ManyInsights: A Visual Analytics Approach to Supporting Effective Insight Management", Chen and Yang present a multi-dimensional visual analytics prototype, ManyInsights, that integrates several insight management approaches, including insight annotation, browsing, retrieval, organization, and association. This paper also reports a longitudinal case study that has evaluated ManyInsights with a domain expert, realistic analytic tasks, and real datasets.

Acknowledgments

We thank the authors for contributing their papers to this special issue and thank all the reviewers who dedicated their precious time to provide timely and valuable reviews and comments. In addition, we would like to acknowledge the Managing Editor, He Chen, and his staff at Tsinghua University Press for their tremendous help during the production of this special issue.

Editors: Gitta Domik
and Scott Owen

How Visualization Courses Have Changed over the Past 10 Years

G. Scott Owen
Georgia State University

Gitta Domik
University of Paderborn

David S. Ebert
Purdue University

Jörn Kohlhammer
Fraunhofer IGD Darmstadt

Holly Rushmeier
Yale University

Beatriz Sousa Santos
University of Aveiro

Daniel Weiskopf
University of Stuttgart

The past 10 years have seen profound changes in visualization algorithms, techniques, methodologies, and applications. For example, we're seeing

- extensive use of GPUs,
- improved algorithms for flow or volume visualization,
- emphasis on highly interactive visual interfaces,
- the advent and increasing importance of visual analytics,
- an increase in nontechnical students in our courses,
- greater need for professional use of visualization in the workplace, and
- evaluation frameworks for effective visualization.

All this forces alterations to our visualization courses, especially what, how, or whom we teach. A basic problem has always been that we couldn't rely on standard textbooks to frame the mandatory knowledge in this field. This situation is unlike that of computer graphics, in which the community widely acknowledges several standard textbooks. Visualization curricula suggestions—for example, ACM Siggraph's Education Committee recommendations (www.upb.de/cs/vis) or the Visual Analytics Digital Library (http://vadl.cc.gatech.edu)—are partly outdated or incomplete. Computer science curricula guidelines, such as from the IEEE and ACM, also lag in their recommendations of content for this novel, dynamic knowledge area.

Outdated course content recommendations, together with profound changes in the underlying technology and methodology, produce an unstable ground for educators at a time when visual representations have gained great importance in economics, science, and many other areas of society.

To address this issue, under the auspices of the ACM Siggraph Education Committee, we held meetings or workshops at Siggraph 2011 and 2012 and a panel and workshop at Eurographics 2012. At the panel, called "The Changes We Have Made to our Visualization Courses over the Last 10 Years," Holly Rushmeier, Jörn Kohlhammer, David Ebert, Beatriz Sousa Santos, and Daniel Weiskopf discussed how they've changed their courses to reflect current problems and practical solutions.[1] (Slides are at www.upb.de/cs/vis.) Each panelist has many years' experience teaching courses covering topics such as scientific visualization, data visualization, information visualization, visualization techniques, and visual analytics.

Here, we examine the insights gathered at the panel, workshops, and meetings.

Visualization in a Liberal Education

Rushmeier teaches the course Visualization: Data, Pixels, and Ideas in the context of Yale's liberal

 Published by the IEEE Computer Society

arts education. Liberal education is, according to the Association of American Colleges and Universities, "a philosophy of education that empowers individuals with broad knowledge and transferable skills, and a strong sense of value, ethics, and civic engagement."[2] Consequently, Rushmeier has a mixed audience with technical and nontechnical backgrounds for which she must design meaningful course content and assignments. Because of her students' mixed background, her course requires no programming or advanced mathematics.

The students' goals are to

- understand visualization's basic components,
- understand computer graphics tools for producing visualizations,
- recognize bad visualizations, and
- use visualization effectively in discovery and communication.

The course covers

- spatial visualization and projections (3D to 2D);
- motion;
- interaction;
- communication best practices—for example, Edward Tufte's principles; and
- scientific and information visualization.

(The courses of the other educators we'll be looking at also include many or most of these topics.)

Class assignments focus on the students' future work life needs. This means Rushmeier must find engaging, manageable datasets on political, historical, or other issues in the humanities. It also means having the students experiment with tools that allow flexibility in designing visualizations without requiring programming skills. Because there's no programming, the course uses tools such as Excel, Matlab, and a VRML (Virtual Reality Modeling Language) viewer.

Geographic information systems are another tool the students learn to use. These systems are used increasingly in the humanities, including in subjects such as comparative literature (see Figure 1).

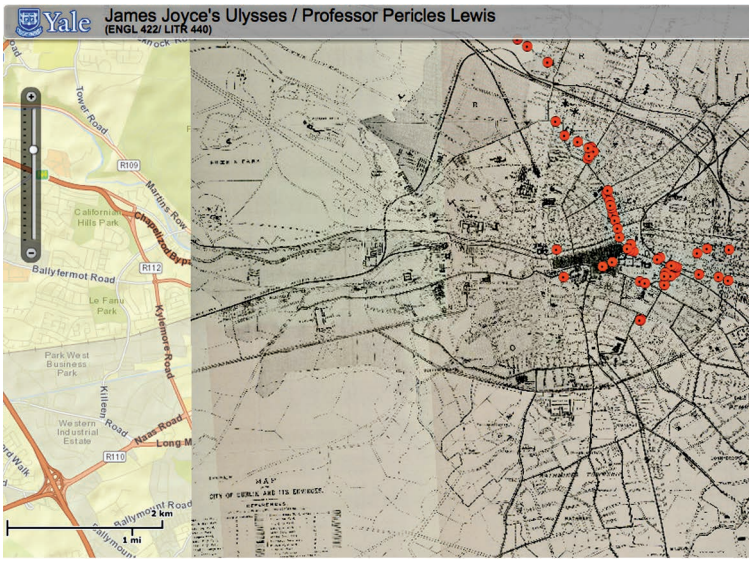


Figure 1. Students in Pericles Lewis's *Ulysses* seminar at Yale University mapped major events in the novel using the addresses in Google, cross-referenced with a map of Dublin in the time of James Joyce.

A Perspective between Research and Business Careers

Kohlhammer sees an increase in students' motivation to learn visualization aspects for their business careers. As in Rushmeier's course, assignments and projects in his Information Visualization and Visual Analytics course at Technische Universität Darmstadt focus on students' later work life by providing hands-on experience with real-world datasets and data types. His courses have a mix of computer science, business informatics, and mathematics students, with some psychology and engineering students. They all have solid programming skills and an affinity for computer graphics.

His course has evolved to include more practical exercises and a strong connection to industry, dealing with areas such as business intelligence, finance (for example, risk analysis), and security. He encourages and supports student involvement in using real-world datasets in global competitions such as the VAST Challenge (Visual Analytics Science and Technology; http://vacommunity.org/VAST+Challenge+2013).

Figure 2 shows a Web-based visual search system for time-oriented research data that Kohlhammer's students developed.[3]

Teaching Visual Analytics: Leveraging Multidisciplinarity

Ebert and Elmqvist teach Introduction to Visual Analytics to students with diverse backgrounds, so the course requires no programming expertise.[4] The students are expected to have a knowledge of one or more of these areas: data analysis, knowledge management, statistics, computer graphics, or visualization. The course consists of group discussions of papers, lectures by the instructors (the course is team-taught), projects, and student presentations of papers.

The projects, which might be individual or group, are particularly important. At least five projects have resulted in conference submissions. Figure 3 shows an example project.

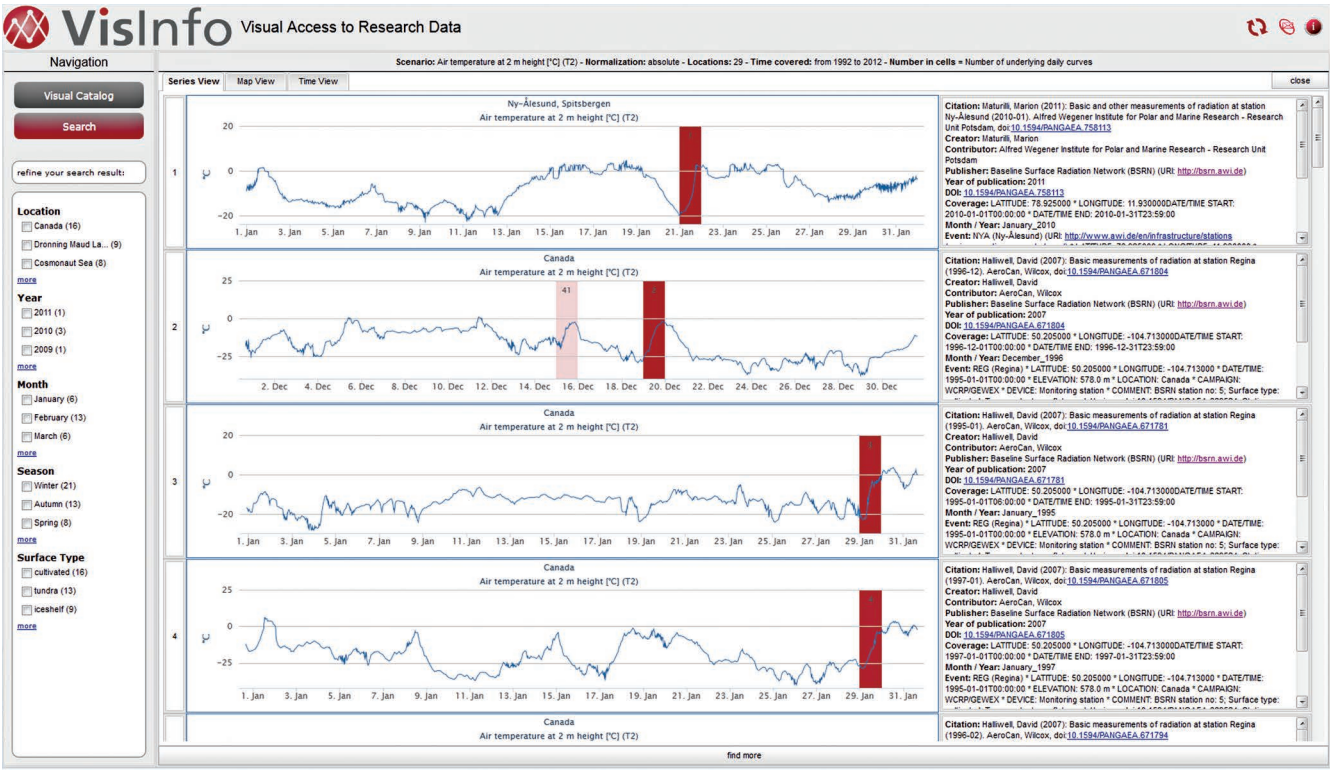


Figure 2. Jörn Kohlhammer's students at Technische Universität Darmstadt developed a Web-based visual search system for time-oriented research data. First, they created a visual catalog of daily temperature patterns based on the self-organizing-maps algorithm (not shown). Then, they presented a results list of documents based on a detailed selection. Metadata facets allowed interactive drill-downs in the results set.

This course, the only one in this article that focuses on visual analytics, has become a model for courses at many universities. Ebert finds a challenge in the fact that field of visualization has become too broad to cover in 15 weeks. So, the university complements this course with other courses—for example, Visualization Techniques, which covers in detail such topics as volume and flow visualization.

Changes in Beatriz Sousa Santos's Visualization Courses

Over the past decade, Sousa Santos has added more material on human characteristics (beyond visual perception), distributed and collaborative visualization, and displays. She also now teaches information visualization courses. Her students read more research papers and perform more evaluation experiments. Her courses include both undergraduate and graduate students with backgrounds in computer science, engineering, and management information systems. They do practical assignments using the Visualization Toolkit.[5] Sousa Santos and her colleagues have also integrated user studies into their courses.[6]

Teaching Visualization at the University of Stuttgart

The University of Stuttgart, where Weiskopf teaches, offers a variety of courses in computer graphics, geometric modeling, image synthesis, and visualization. In particular, there are courses on scientific visualization and information visualization. Additionally, the university uses a two-semester visualization-centered project to teach software engineering.[7] The university also offers the outreach course Introduction to Visualization in Science and Engineering. Students in that course have limited programming experience, so the course is heavily tool based.

The program aims to generate a common basis for computer graphics, visualization, and computer vision and to complement computer science students' typical mathematical and theoretical education. A challenge is the growing need for background knowledge from diverse fields such as mathematics, computer science, human-computer interaction, psychology, data mining, machine learning, and application-specific domains.

The Emerging Areas

From the panel, workshops, and meetings, three distinct areas emerged:

- *scientific or data visualization*, in which the data dimensions usually coincide with physical di-

mensions, such as in medical or remote-sensing scalar or flow data;
- *information visualization*, with typically multidimensional data, such as in finance, business intelligence, or large databases; and
- *visual analytics*, with massive, multisource, multiscale, heterogeneous, and streaming data.

Data preprocessing (for example, filtering, normalizing, and linguistic analysis) and subsequent visual presentations (for example, line graphs, line-integral-convolution images, and cone trees), which both depend on data syntax and semantics, might be different for these areas but also overlap considerably.

All three areas share some learning objectives. Students should be able to

- understand visualization techniques;
- recognize good versus misleading visualizations;
- select appropriate visualization techniques and visual attributes on the basis of the data and task;
- explain selected algorithms underlying visualization techniques for, for example, 2D data, 3D scalar or vector data, time-dependent data, multivariate data, hierarchically structured data, graphs and networks, or data with other structures;
- discuss the handling of unstructured data;
- understand the appropriate manipulation of data before mapping (which differs between data visualization, information visualization, and visual analytics);
- understand the limitations and capacity of human information processing;
- group and describe visualization techniques by some order (for example, by domain, data characteristics, or tasks);
- discuss how scaling (of the data or display) influences visualization techniques; and
- understand the theory and application of evaluation techniques to prove a visualization or interaction technique's success.

Although the three areas have distinct data domains, courses in them must cover the following themes.

The User

This theme includes human information-processing limitations and capabilities as well as an understanding of the tasks users bring to visualization problems.

The Design Stage

This stage describes a careful mapping of data components to visual attributes and the interactivity between users and data as well as between users and visualizations.

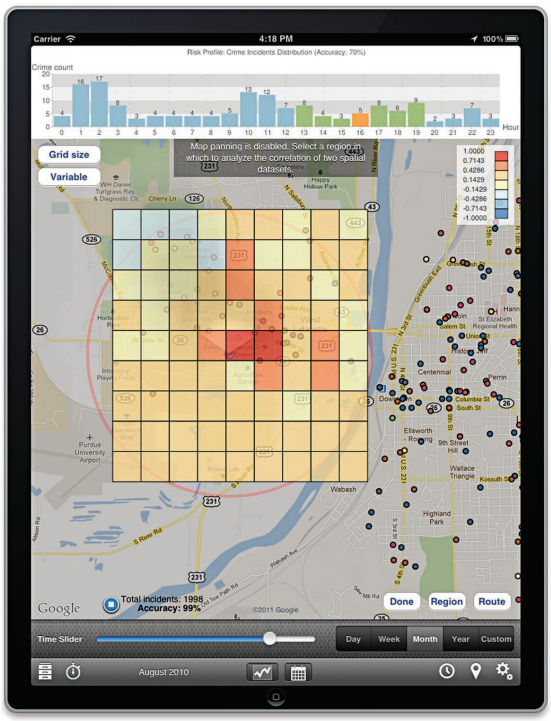


Figure 3. This project, by Ahmad Razip, a student in Niklas Elmqvist and David Ebert's Introduction to Visual Analytics course at Purdue University, correlated bus stops and crime incidence distribution.

Visual Presentation

This includes a wealth of visualization solutions, sorted by data characteristics, application domain, or task and described by their various parameters. Instructors can present this theme at the breadth level by showing and discussing (interactive) visualizations. They can provide breadth-level training by using the available tools, in-depth training by developing interactive visualization techniques on a GPU, or training at any stage in between, depending on the students' qualifications.

Interaction Techniques

Interaction techniques are a requirement for visual analytics. They're also becoming increasingly necessary for data and information visualization, in which GPU techniques can reach the necessary processing speed.

Communication

Visual analytics in particular has stressed production, presentation, and dissemination as part of the visualization process. However, these topics are also important for data visualization and information visualization.

Collaboration

Interactivity aids collaboration, especially in synchronous and local situations. However, collaboration among stakeholders can also involve aspects that are asynchronous and distributed, such as Web-based collaboration technologies.

Evaluation

Evaluation is continuous. It starts with requirements analysis of the visualization problem. It continues with the human-in-the-loop's constant awareness of the software processes proceeding toward the visualization goal. It ends with evaluation to ensure reaching the goal for the specific visualization problem.

One challenge is that more nontechnical students are interested in the courses because of visualization's increasing use in business and industry.

Displays

The variety of different displays' capabilities (size and spatial and tonal resolution) poses problems for visualization techniques, interactivity, and communication. These capabilities must be addressed at least at the mapping or design stage.

Challenges Identified

One challenge is that more nontechnical students are interested in the courses because of visualization's increasing use in business and industry. This necessitates the use of tools and real-world cases and datasets. If a course teaches both technical and nontechnical students, this potential difficulty could actually be an opportunity to approach a visualization problem from the viewpoints of multiple disciplines.[8] For computer science students, the courses should add newer developments such as shader programming and computer vision. One benefit of this is that some student projects might be worthy of conference publication, as has been the case with Elmqvist and Ebert's visual-analytics course.

Another challenge, as Ebert and Weiskopf stated in the panel discussion, is the need for instructors to update their own knowledge in diverse background fields ranging from math, to human-computer interaction and perception, to shader programming. So, this article's references include a few textbooks we use.[9–12]

At the panel, both Rushmeier and Ebert remarked that visualization has become too broad of a field to cover in one semester in suitable depth. So, instructors must decide between the breadth and depth of topics or offer one or more complementary visualization courses.

To help educators respond to the changes occurring in visualization courses, we've compiled a set of materials they can use to update their courses:

- the complete set of slides of the panelists and coauthors in this article,
- previously published articles by Ebert, Weiskopf, and Sousa Santos on their visualization courses (all from *IEEE CG&A*'s Education department), and
- related articles from *IEEE CG&A*'s Education department and other sources.

Links to these materials are at www.upb.de/cs/vis.

Acknowledgments

Thanks to Riccardo Scateni for providing the rooms for the Eurographics 2012 workshop at the University of Cagliari.

References

1. G. Domik et al., "Visualization Curriculum Panel—or the Changes We Have Made to Our Visualization Courses over the Last 10 Years," *Eurographics 2012—Education Papers*, 2012; www.cs.uni-paderborn.de/fileadmin/Informatik/AG-Domik/VisCurriculum/folien/eg2012-panel-Domik-2.pdf.
2. "Liberal Education," Assoc. of Am. Colleges and Universities, 2013; www.aacu.org/resources/liberaleducation/index.cfm.
3. J. Bernard et al., "Irina: A Visual Digital Library Approach for Time-Oriented Scientific Primary Data," *Int'l J. Digital Libraries*, vol. 11, no. 2, 2011, pp. 111–123.
4. N. Elmqvist and D.S. Ebert, "Leveraging Multidisciplinarity in a Visual Analytics Graduate Course," *IEEE Computer Graphics and Applications*, vol. 32, no. 3, 2012, pp. 84–87.
5. P. Dias, J. Madeira, and B. Sousa Santos, "Education: Teaching 3D Modelling and Visualization Using VTK," *Computers and Graphics*, vol. 32, no. 3, 2008, pp. 363–370.
6. B. Sousa Santos et al., "Integrating User Studies

into Computer Graphics-Related Courses," *IEEE Computer Graphics and Applications*, vol. 31, no. 5, 2011, pp. 94–96.
7. C. Müller et al., "Large-Scale Visualization Projects for Teaching Software Engineering," *IEEE Computer Graphics and Applications*, vol. 32, no. 4, 2012, pp. 14–19.
8. G. Domik, "Fostering Collaboration and Self-Motivated Learning: Best Practices in a One-Semester Visualization Course," *IEEE Computer Graphics and Applications*, vol. 32, no. 1, 2012, pp. 87–91.
9. C. Ware, *Information Visualization: Perception for Design*, 3rd ed., Morgan Kaufmann, 2012.
10. M.O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization*, AK Peters, 2010.
11. J.J. Thomas and K.A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE, 2005.
12. M. Bailey and S. Cunningham, *Graphics Shaders: Theory and Practice*, 2nd ed., AK Peters, 2011.

G. Scott Owen *is professor emeritus at Georgia State University's Department of Computer Science. Contact him at sowen@gsu.edu.*

Gitta Domik *is a professor at the University of Paderborn's Faculty for Electrical Engineering, Computer Science, and Mathematics. Contact her at domik@uni-paderborn.de.*

David S. Ebert *is the Silicon Valley Professor of Electrical and Computer Engineering at Purdue University's School of Electrical and Computer Engineering. Contact him at ebertd@purdue.edu.*

Jörn Kohlhammer *heads the Competence Center for Information Visualization and Visual Analytics at Fraunhofer IGD Darmstadt and is a member of the Interactive Graphics Systems Group at Technische Universität Darmstadt. Contact him at joern.kohlhammer@igd.fraunhofer.de.*

Holly Rushmeier *is a professor and the chair of computer science at Yale University. Contact her at holly@acm.org.*

Beatriz Sousa Santos *is an associate professor in the University of Aveiro's Department of Electronics, Telecommunications, and Informatics. Contact her at bss@det.ua.pt.*

Daniel Weiskopf *is a professor of computer science at the University of Stuttgart. Contact him at weiskopf@visus.uni-stuttgart.de.*

Contact department editors Gitta Domik at domik@uni-paderborn.de and Scott Owen at sowen@gsu.edu.

Report from Dagstuhl Seminar 13352

Interaction with Information for Visual Reasoning

Edited by

David S. Ebert[1], Brian D. Fisher[2], and Petra Isenberg[3]

1 **Purdue University, US,** ebertd@purdue.edu
2 **Simon Fraser University – Surrey, CA,** bfisher@sfu.ca
3 **INRIA Saclay – Île-de-France – Orsay, FR,** petra.isenberg@inria.fr

Abstract

From August 26–August 30, 2013 Seminar 13352 was held at Dagstuhl on the topic of "Interaction with Information for Visual Reasoning." The seminar brought together a group of cognitive scientists, psychologists, and computer scientists in the area of scientific visualization, information visualization, and visual analytics who were carefully selected for their theoretical and methodological capabilities and history of interdisciplinary collaboration. During the seminar seven discussion groups were formed during which the role of interaction for visualization was carefully reflected on. We discussed in particular the value, structure, and different types of interaction but also how to evaluate visualization and the idea of 'narrative' as applied to visual analytics. This report documents the program and short summaries of the discussion groups for the seminar.

Seminar 26.–30. August, 2013 – www.dagstuhl.de/13352
1998 ACM Subject Classification H.5.2 [Information Interfaces and Presentation]: User Interfaces – Graphical User Interfaces (GUI), I.3.6 [Computer Graphics]: Methodology and Techniques – Interaction Techniques
Keywords and phrases Interaction, visualization, visual analytics, cognitive science, psychology
Digital Object Identifier 10.4230/DagRep.3.8.151

1 Executive Summary

David S. Ebert
Brian D. Fisher
Petra Isenberg

Scientific and information visualization researchers routinely build and evaluate interactive visualization systems to aid human reasoning. However, this work is often disconnected from the methodological and theoretical tools developed by the cognitive and social sciences to address the complexities of human thought processes. Those tools and methods can help us to understand human perception and understanding of data visualization, but typically do not address how rich interaction with computational processes could be engineered to support better decision-making. Yet, an increasing number of researchers are turning to the question of how to best engineer interaction techniques for visualization and how to best study and understand their influence on cognition, insight formation, and also efficiency and effectiveness of work. The goal of this seminar was to bring together researchers in cognitive science and psychology with researchers in the field of visualization to discuss the value that interaction can bring to visualization, how best to study it, and how research on interaction in cognitive science can be best integrated into visualization tools and systems to the benefit of domain experts or casual users of these tools.

Interaction with Information for Visual Reasoning, *Dagstuhl Reports*, Vol. 3, Issue 8, pp. 151–167
Editors: David S. Ebert, Brian D. Fisher, and Petra Isenberg
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Table of Contents

3 About the Seminar

Participants

The seminar brought together a diverse group of international cognitive scientists, psychologists, and computer scientists in the area of scientific visualization, information visualization, visual analytics, and human computer interaction. All participants (see Figure 1) were carefully selected for their theoretical and methodological capabilities and history of interdisciplinary collaboration. Thirty participants joined the seminar, out of which seven had a background in psychology and the remainder were primarily computer scientists in training. Eleven participants were female and three in total came from industry. For about one third of participants this seminar was their first Dagstuhl event. Figure 2 shows gender balance and country statistics for all participants.

Format

The seminar followed a format largely based on breakout groups. The first day of the seminar involved short introduction slides for each participant with longer 15 minute invited talks from the domains of cognitive science and psychology. Tuesday the first four breakout groups discussed topics on mixed initiative interaction, crowd interaction, the value of interaction, and conceptual structures of interaction. Each breakout group was comprised of participants with mixed backgrounds in computer science and cognitive science/psychology and had the goal to work on a specific problem related to the title of the breakout group. Wednesday, each participant switched to a second breakout group on topics: evaluation of interaction, magical interaction, and visual narrative. Almost all participants then went together on a social event to visit the Völklinger Hütte (see Figure 3). Thursday morning we heard a presentation about DBLP, and continued with the second breakout group sessions. Thursday ended with discussion in the breakout groups about publishable results from the seminar and working towards establishing a publication plan of action. These were then presented in front of the whole group. Two invited talks started the day on Friday which ended with a discussion on publication venues that would be beneficial for both the computer scientists and cognitive scientists/psychologist. Table 1 gives an overview of the seminar schedule and the following list includes the titles of the individual breakout groups.

- Mixed Initiative Interaction
- Crowd Interaction in Visual Reasoning
- Magical Interaction
- Evaluating Interaction for Visual Reasoning
- Conceptual Structures of Interaction for Visual Reasoning
- Visual Narrative
- The Landscape of Explanations for the Value of Interaction for Visual Reasoning

Output

The organizers and participants planned to publish results from the breakout groups as a Morgan Claypool mini series on interaction for which David Ebert is the series editor. Working groups have been invited to publish their results there. Furthermore we are in contact with an international journal for an open-call special issue to further push the importance of interaction for visual reasoning as an emerging topic in the domain of visualization.

Figure 1 Group picture of all participants taken in the sun outside the Dagstuhl chapel.

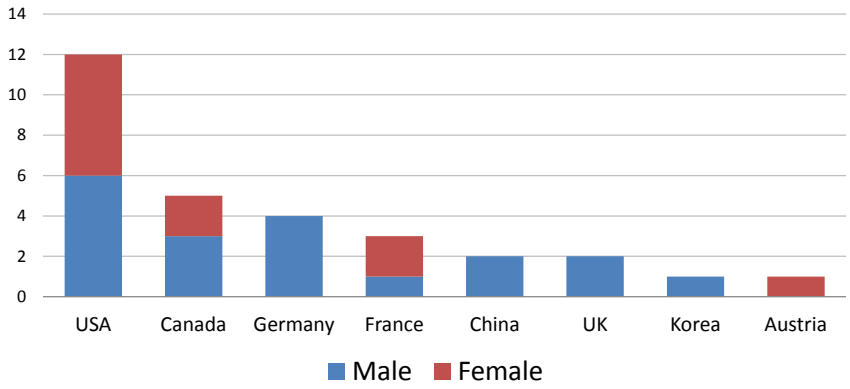


Figure 2 Gender balance and country of participants' home institution.

Figure 3 Excursion to Völklinger Hütte allowed for less structured research conversations and an interesting lesson about the ironworks (picture courtesy of Fanny Chevalier).

Table 1 Schedule of the seminar. Details on talks and breakout groups follows further below.

Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*
Introduction to the Seminar Introduction Participants I	Discussion: Breakout Groups	Discussion: Breakout Groups II	Presentation: DBLP	Invited Talks: Sketching
Introduction Participants II	Breakout Groups I	Breakout Groups II	Breakout Groups II	Discussion on Publication Venus & Closing
Introduction Participants III	Breakout Groups I	Social Event: Völklinger Hütte	Discussion Writing Groups I	
Invited Talks Psychology	Report from Breakout Groups I	Social Event: Völklinger Hütte	Discussion Writing Groups II	

The Dagstuhl team performed an evaluation at the end of the seminar. The responses were primarily positive with the overall quality of the seminar rated with a 10/11. All 16 respondents reported that they agreed or agreed completely that the seminar inspired new ideas for their own work, development or teaching. 15/16 respondents agreed that the seminar inspired joint work and all 16 respondents said that the seminar led to insights from neighboring fields. This is exciting as it shows that the seminar's goal of bridging the gap between two research communities was successfully met. In terms of improvements, several participants suggested to leave more room for (impromptu) talks and that it would have helped to prepare participants more prior to arriving at Dagstuhl. We take these as suggestions for the organization of possible future seminars.

Acknowledgments

We would like to acknowledge the valuable input of Dr. Shixia Liu (Microsoft Research Asia) in the application and preparation of this seminar.

4 Overview of Talks

4.1 Interacting with Information – Overview of Past & Current Work

Simon Attfield (Middlesex University, GB)

Joint work of Attfield, Simon; Okoro, Efeosasere

In this talk I presented myself and my research interests. I discussed how an interest in information interaction focusing on domain areas such as journalism, e-discovery, intelligence analysis and healthcare has led to an interest in sensemaking and how it can be supported through digital design. Something I am currently interested in is information structuring during sensemaking, such as howit can be characterised and the comparative effects of different kinds of structuring, such as narrative and argumentation. For both of these questions I am working with Efeosasere Okoro to develop a relational language for capturing

the semantics of user generated information structures and looking at the comparative effects of different structuring conventions on task performance anduser experience.

4.2 Cognitive Science of Representational Systems

Peter C.-H. Cheng (University of Sussex – Brighton, GB)

In addition to giving a brief overview of Peter Cheng's main areas of research the talk focused on how cognitive science can inform the design of representational systems to support complex problem solving and learning in conceptually rich domains. It seems necessary to combine our understanding of higher cognition, forms of external representations and the nature of knowledge within a Representational Epistemic approach in order to successfully design graphical displays that can fully accommodate the many diverse tasks that are typically found in complex domains. Novel notations and visualisation were briefly presented for event scheduling, personnel rostering, production planning and scheduling, high school algebra, electricity, particle mechanics, probability theory, propositional logic and syllogisms.

4.3 Flexible Perception of Structure in Viz & Education

Steve Franconeri (Northwestern University – Evanston, US)

Selective attention allows us to filter visual information, amplifying what is relevant and suppressing what competes. But recent work in our lab suggests another role – flexibly extracting and manipulating visual structure. Selective attention allows us to group objects with similar features, extract spatial relationships between objects, and imagine manipulations of objects. An understanding of these mechanisms has concrete implications for the design of visualizations across science and education.

4.4 A Sample of Sketching Research in Cognitive Science

Steve Franconeri (Northwestern University – Evanston, US)

Sketching is a tool for visual thinking. It helps people explore information sets that are too large to hold (or process) in working memory. It helps people re-organize information, allowing it to be seen from a new perspective. I presented two case studies of the power of sketching – in the first, sketching reveals how people understand a problem within an image, and in the second, sketching facilitates insight by promoting for aging through visual relationships.

4.5 Human Interactions in Abstract Visual Spaces

Wayne D. Gray (Rensselaer Polytechnic Institute, US)

Human behavior is interactive behavior. Behavior emerges from the interaction of bounded cognition with the natural or designed task environment and task goals. Topics covered included: (a) Interactive routines, (b) The eye-hand span, (c) Analogies for memory, (d) Local, not global optimization, (e) Modeling the whole human (not just the convenient bits!), (f) Modeling pre-attentive and attentive visual processes, (g) Tools for the statistical analysis of visual saliency and similarity, (h) Tools for analyzing eye data, and (i) Possibilities for more indepth talks including *The Cognitive Science of Natural Interaction* and *Elements of Extreme Expertise*.

4.6 Interacting with Visual Representatives

David Kirsh (University of California – San Diego, US)

Topics covered were: a) what is the difference between epistemic and pragmatic actions – and why they matter for visualization theory; b) explain a core(epistemic) interactive strategy used in reasoning with visualizations. This involves mentally projecting a structure onto an external structure, for instance, an illustration, a geometric figure, a manipulable visualization, then physically realizing that mentally projected structure by altering the external structure or visualization, then start this a) perceive mental project structure; b) create structure process over again by mentally projecting onto this newly altered structure; c) explain the importance of the difference between explicit and implicit encoding of information: when does representation A encode information more explicitly than representation B – good visualizations encode the right information more explicitly; d) what does interaction do for us that we could not do as well in our heads?

4.7 Sketching and Embodied Cognition

David Kirsh (University of California – San Diego, US)

I explore two topics that highlight the power of embodied cognition: using the process of making a sketch to help decide whether a stone is a genuine tool shaped by prehistoric humans as opposed to a visually similar stone that occurs naturally. This shows how the process of physically working with a sketch or illustration teaches us more than just looking at the same illustration or sketch. The second topic introduces another instance of 'modeling or simulation to drive cognition.' In dance there is a process called marking that has many of the same virtues of sketching but this time using the body instead of paper and pencil. Marking is a process in dance practice where a dances sketches a phrase, using less energy

and working on aspects of that phrase rather than performing the whole phrase in all its complexity and with all the effort and speed required for full out performance. Our study of marking showed that marking can help a person explore aspects of a movement, one by one; it also allows a person to bring certain elements or dimensions of the phrase into focus. This idea that modeling or imperfectly simulating a complex process can lead to insight that is hard or impossible to obtain by studying the fullprocess is found in practices other than dance and lithic sketching. In painting artists make practice sketches in order to explore elements they want to highlight or get just right. By making simplified models they bring aspects or elements or dimensions of the complex thing they want to paint into better focus, ensuring they capture features they might otherwise have missed. Attached is a paper published on marking and other issues related to interactivity "Embodied Cognition and the Magical Future of Interaction design."

4.8 Toward Systematic Design of Different Interactive Visualization Components

Kamran Sedig (University of Western Ontario, CA)

Visualization tools can support and enhance the performance of cognitive activities such as sense making, problem solving, and analytical reasoning. To do so effectively, however, a human-centered approach to their design and evaluation is required. This presentation highlights a number of different issues that we have been investigating in order to identify some of the main components of interactive visualizations in order to develop a systematic approach for their design and analysis. A few of the issues that are discussed include: Interaction, interactivity, cognitive activities (i. e., reasoning), and visual representations. A recent paper dealing with interaction design for visual representations is also uploaded.

5 Working Groups

5.1 Visual Narratives

Simon Attfield (Middlesex University, GB), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), Catherine Plaisant (University of Maryland – College Park, US), Margit Pohl (TU Wien, AT), Huamin Qu (The Hong Kong University of Science & Technology, HK), Michelle X. Zhou (IBM Almaden Center – San José, US)

As a group we explored the application of the notion of 'narrative' to visual analytics. We defined narrative as, 'A sequence of events connected in a meaningful structure where the connecting principles are time, causation, logic, rationale and/or entity relationships.' We considered that a 'visual narrative' is the visual rendering of this, and that an 'Interactive visual narrative' is a visual narrative where interaction occurs during construction and/or presentation. For the construction phase some key considerations are element selection and

editing and annotating the narrative. For presentation some key considerations are pace, branching, overview and detail, and annotation.

We agreed that narrative can bring structure to information and that this can support comprehension, recall, personal and public audit of an analysis, and also help support analytic systematicity and influence decisions.

We agreed on at least four kinds of narrative as applied to visual analytics. These are:

Stories in the data – Stories which recount temporal and/or causal sequences within data. This is relevant where the data has a temporal dimension or when time is used to 'unfold' data. It is a selective and possibly interpreted account of what the data 'says' structured around temporal associations. It may provide a 'natural' way of thinking about data when it comes to identifying particular kinds of pattern, generating causal explanations, making predictions of the future and for supporting higher-level categorisation (e. g. determining intent of actors during legal cases for crime classification).

The stories of analysis – Stories which communicate the provenance of an analytic outcome by way of a history of the analytic process. It depicts the process through which an analysis was conducted and conclusions reached. Visibility of this story can offer support to the analyst and to others either during or after an analysis for reflecting what was done, auditing and interpreting outcomes. To support construct of this story, data might be gathered during the analysis by automated logging and/or manual annotation. Considerations in abstracting a meaningful narrative include: What are the meaningful agents and objects? What are the meaningful units of activity? How do you detect and represent analytic branching? How do you make a narrative engaging and build in 'Dramaturgie' (crisis/outcome, tension/release)? How do you map from low-level events (as captured by logging) to a meaningful story of progressive sensemaking?

The 'logic' of the conclusion – This is a story which recounts the (probably informal) logic of an analytic outcome. It links elements in terms of premises (both observed and assumed) and conclusions into a coherent argument. It is different from a 'story in the data' and a 'story of the analysis' in that it specifically links elements in terms of the way that one proposition supports another, rather than how one event led to another. It is a sequence of ideas akin to a logical or mathematical proof in which the relations between ideas are determined by relations of implication and not the chronology in which they occurred. An advantage of this kind of story is that its form can make explicit assumptions which might otherwise remain implicit.

Educational narratives – A story generated for pedagogical motives intended for teaching others how to perform or interpret an analysis or how to use complex analytic tools. A key role of story in this case is to lead the student through a series of ideas which may progressively build in a way that is engaging, accessible and memorable. Where the learning supports informed decision making, for example in the case of educating patients, there may be considerations of how presentational elements, such as order, might affect bias.

Finally, we discussed interaction issues as these relate to narrative in visual analytics. We identified three phases of user involvement which deserve consideration. They are: **Data gathering**, raising questions of how this is done; **Narrative construction**, raising questions of how this is done and when; and **Presentation**, raising questions about the provision of overviews for the sake of coherent user mental models, the level of user involvement in interaction (lean forward vs lean back), the pace of presentation, granularity and abstraction of the narrative, and the tension between telling an engaging and representative story.

5.2 Evaluating Interaction for Visual Reasoning

Anastasia Bezerianos (University Paris South, FR), Mary Czerwinski (Microsoft Research – Redmond, US), Brian D. Fisher (Simon Fraser University – Surrey, CA), Steve Franconeri (Northwestern University – Evanston, US), Wayne Gray (Rensselaer Polytechnic Institute, US), Petra Isenberg (INRIA – Saclay, FR), Bongshin Lee (Microsoft Research – Redmond, US), Jinwook Seo (Seoul National University, KR)

One of the main challenges in designing interactive visual analytics systems is to measure the effectiveness of their interaction designs. Traditional performance measures such as task completion time and error rate often fail to demonstrate the value of interacting with visual representations and the effect of interaction on the analysts' reasoning process. Measuring insights is also limited in that it is hard to replicate and quantify, and it does not capture the role of the interactive system in the process that leads to the insightful moments. The role of interaction is thus difficult to tease out with an insight based evaluation. The goal of this paper is to present to the visual analytics community alternative measures related to interaction, human reasoning and analysis processes, borrowed and adapted from the field of cognitive psychology. The article that will be written based on the discussion at Dagstuhl will be structured around the design of a study to evaluate different interaction modalities for visual reasoning. It will discuss both high level questions, such as the formation and evolution of a research question, and low level aspects including the choice of evaluation tasks and methodologies. This article will serve both as an introduction to alternative evaluation measures and methodologies adapted from cognitive psychology, as well as a walkthrough example for researchers on how to formalize research hypotheses and structure evaluations around them.

5.3 The Landscape of Explanations for the Value of Interaction for Visual Reasoning

Sheelagh Carpendale (University of Calgary, CA), Anastasia Bezerianos (University Paris South, FR), Peter Cheng (University of Sussex – Brighton, GB), Brian D. Fisher (Simon Fraser University – Surrey, CA), Steve Franconeri (Northwestern University – Evanston, US), Daniel Keefe (University of Minnesota – Twin Cities, US), Bongshin Lee (Microsoft Research – Redmond, US), Chris North (Virginia Polytechnic Institute – Blacksburg, US)

Our intuition tells us that interaction is a really important factor for visual reasoning. As a community we have generated a wealth of examples of interaction techniques. However, there is little consensus about how to explain how interactions actually create these benefits. By examining a series of diverse interaction techniques in light of visual reasoning we will begin mapping the landscape of explanations of how these interactions add value in terms of visual reasoning. We will use a semi-structure knowledge acquisition process to gather a series of interaction examples that each have three parts: a short stop motion storyboard, an accompanying verbal explanation of the task and system; and, in particular, the creators

explanations of the perceived benefits in terms of visual reasoning. By intentionally both sampling for diversity and analyzing from the perspectives of multiple research disciplines we hope to add richness to this discussion. The contribution of this paper is an exploration the space of alternative explanations to expand our understanding of the value of interaction for visual reasoning.

5.4 Mixed Initative Interaction

Christopher Collins (University of Ontario, CA), Simon Attfield (Middlesex University, GB), Fanny Chevalier (University of Toronto, CA), Mary Czerwinski (Microsoft Research – Redmond, US), Heidi Lam (Google Inc. – Mountain View, US), Catherine Plaisant (University of Maryland – College Park, US), Christian Tominski (Universität Rostock, DE), Michelle X. Zhou (IBM Almaden Center – San José, US)

The group defined mixed initiative interaction as a type of interaction for visual reasoning in which the human analyst and the visualization system both are active participants in the interaction. In a traditional interaction scenario, the visualization software is reactive, responding to inputs from the analyst. In MI interaction, the system would play a more active role, for example, making suggestions about appropriate views or next steps in the analysis process. There are then two directions of interaction: human to system, e.g. applying filters, making selections, loading new data; system to human, e.g. suggesting views, suggesting next steps, automatic highlighting of potentially interesting part of a view.

Mixed initiative interaction has been studied for several years, but remains on the periphery of mainstream visualization research. Systems in this area are often called 'smart visualization' or 'intelligent user interfaces'. However, it seems the community is skeptical due to the cost of error: if a smart visualization system suggests a particular representation type, or an analysis process which is inappropriate to the data or current task requirements, then an analyst could become frustrated, or, worse, may come to incorrect conclusions, biased by the underlying interaction model. Other challenges in this research include being able to gather appropriate and sufficient user data to create a model of the user, such as understanding their level of experience, preferences, prior domain knowledge, etc. As this sort of data is difficult to gather and often inconclusive, we focussed our discussion on MI interaction possibilities in scenarios where we do not have prior knowledge about the analyst. Our discussions lead to a list of factors which can be used to evaluate the success of system-initiated interaction prompts:

- Are they timely? Are suggestions provided at the right time or do they interrupt the analyst's flow?
- Does the system take initiative sparingly? If the system takes the initiative too often, the analyst may become fatigued and ignore suggestions.
- Are system suggestions appropriate? Is the system suggesting views, prompts, or other cues which enhance the analysis experience and potential for insight? Or do they lead to incorrect conclusions about the data?
- Is the provenance of system suggestions transparent? Can the human analyst understand why the system makes any given suggestion?

Scenarios where MI interaction may be useful include a new analyst using a system for the first time and requiring tutorial-style guidance. In this scenario, the system may not know much about the characteristics, prior knowledge, interaction styles and preferences of the analyst and has to provide assistance based on characteristics of the data and the current interaction session, and perhaps a crowd-driven model of the way other analysts have used the system.

Human to System Interaction (Human Initiative)

We called the types of traditional interactions, such as selecting data items, panning and zooming a view, "*explicit* interactions". Newer forms of interaction, such as 'model steering' by repositioning items in a visualization to indicate prior knowledge about their relatedness, are also important inputs to a mixed initiative system.

Where we focused our discussion was on new forms of cues which may be gathered by the system in a mixed initiative interaction model to improve the quality and timeliness of prompts and suggestions. We called these inputs 'implicit interactions'. We enumerated the following list of potential 'implicit interactions' which could be tracked by a visualization system and used to decide when and how to take initiative:

- Dwell time (eye gaze, touch, or mouse cursor)
- Facial gestures
- Highlighting / copying behaviour
- Repeated actions
- Body position / gestures (proxemics)
- Thrashing—changing actions / direction
- Emotional indicators
- Physiological indicators
- Mouse signatures
- Keyboard signatures
- Repositioning items on the screen
- The history of what they have explored already (the analysis process)

These implicit interactions could provide a wealth of data to a mixed initiative system, but would also have drawbacks which need to be investigated, including privacy concerns, potential for reinforcing actions (encouraging 'tunnel vision'), or ambiguity of the meaning of the indicators. For example, physiological and behavioural indicators of excitement and annoyance may be quite similar. Which implicit interactions would be most important and how they could work together to create a profile of the analyst state are areas of future research inspired by our discussions.

System to Human Interaction (System Initiative)

Others have researched system-initiated interaction driven by user profiles and data characteristics, so we targeted our discussions on the types of feedback a system could provide based on analysis of implicit interaction data. The timing of system-initiated interaction is crucial: ideally it is timely, does not interrupt the flow of analysis and human-initiated interaction, and is appropriate to the data and task. Design decisions to consider in future work include: (a) how to present interaction and analysis suggestions, (b) how to reveal the provenance of guidance (why a suggestion is made by the system), (c) how to encode the confidence level the system has in a suggestion, etc. We recommended that system feedback

could also be subtle or implicit. For example, if the system senses the analyst is "lost" or "stressed", rather than asking "are you stressed?" it could simply provide additional on screen help, adjust or simplify the interface, and suggest alternative views.

We explored a variety of system responses which may be appropriate if the indicators strongly point to the need for system-initiated interaction. Specifically, we looked at possible responses to various detected emotional states, such as frustration, confusion, boredom, interest, and engagement. System responses may include: show more views like the current view, show views different from the current view, show what other people (all people / people like me / experts) did in similar situations, offer help, or simplify the view (remove a data dimension or perform aggregation). MI interaction should be flexible and perhaps system-initiation should be turned off automatically after the analyst does not acknowledge or use system suggestions over a long period of time.

To conclude our meetings, the group brainstormed about a paper outline reporting on our mixed initiative interaction for visual reasoning ideas, and assigned next steps to the participants.

5.5 Conceptual Structures of Interaction for Visual Reasoning

Kelly Gaither (University of Texas – Austin, US), David S. Ebert (Purdue University, US), Thomas Ertl (Universität Stuttgart, DE), Hans Hagen (TU Kaiserslautern, DE), Petra Isenberg (INRIA Saclay, FR), Tobias Isenberg (INRIA Saclay, FR), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), Margit Pohl (TU Wien, AT), Kamran Sedig (University of Western Ontario, CA)

Interaction is a fundamental element of successful visualization methods and tools. In visualization, interaction can support many low-level and high-level tasks and goals, can support different representation and interaction intends, and can be realized by different techniques. The specific incarnations of the interaction design, however, are driven by the specific application domain, by the tasks being supported, by the type of data being analyzed, by the specific representations being chosen, by potential limitations of computability, and by the needs and requirements of the users. The question that we aimed to analyze is if we can identify general principles of interactions that bridge different domains and are common among tasks, data types, and representations. Can we formulate or propose a language or schema of interaction that is common for most if not all visualization tools and methods, potentially with different dialects?

5.6 Magic Interactions with Information for Visual Reasoning

Daniel Keefe (University of Minnesota – Twin Cities, US), Sheelagh Carpendale (University of Calgary, CA), Peter Cheng (University of Sussex – Brighton, GB), Fanny Chevalier (University of Toronto, CA), Christopher Collins (University of Ontario, CA), Tobias Isenberg (INRIA Saclay, FR), David Kirsh (University of California – San Diego, US), Heidi Lam (Google Inc. – Mountain View, US), Chris North (Virginia Polytechnic Institute – Blacksburg, US), Kamran Sedig (University of Western Ontario, CA), Christian Tominski (Universität Rostock, DE), Xiaoru Yuan (Peking University, CN)

Today, there is much excitement around the concept of "natural user interfaces." The interest is sparked in part by the widespread availability of multi-touch devices, including smart phones and tablets. However, the trend is not limited to these new commercial devices; a variety of recently developed user interface techniques that enable seemingly more direct ways of interfacing with computers have been dubbed "natural." Will these natural interactions define the future of computing? As user interface designers, and in particular as designers and researchers interesting in supporting users as they reason about super-complex information, we have to ask, is "natural" actually the right target? Do we really want to design natural interactions or do we want something else? How about "supernatural" or even "magical" interactions? Our Dagstuhl working group found that the more we thought about the systems and human-computer interfaces that have most influenced us or impacted our work, the more we recognized that (at least the first few times we used these systems) they all felt magical. Some examples include: (1) clicking and dragging a drawing of a cartoon character who then responds "intelligently" by changing his pose in direct response to the user's input, understanding how to move as if by magic; (2) Browsing video data by clicking directly on characters in the video rather than using a slider; (3) Bumping mobile devices to transfer files; and (4) Selecting 3D point clouds just by drawing a 2D lasso. All of these interactions have "the power of apparently influencing the course of events by using mysterious or supernatural forces" and "a quality that makes something seem removed from everyday life, esp. in a way that gives delight" – two properties taken directly from the definition of the word magic. Grounded in findings from the cognitive science research community, we developed several explanations for when and why "magical interactions" seem to work well, including the notions of a different cognitive cost structure for natural vs. magical interactions, superpower and amplification, context/temporal appropriateness, and working with underspecified and imprecise data or applications. Based on these insights, we call for a new research focus that moves beyond "natural user interfaces" and instead targets magic interactions with information.

5.7 Crowd Interaction in Visual Reasoning

David Kirsh (University of California – San Diego, US), Huamin Qu (The Hong Kong University of Science & Technology, HK), Jinwook Seo (Seoul National University, KR), Xiaoru Yuan (Peking University, CN)

How can we harness the intelligent capacity of crowds to reason visually about a topic? In recent years real-time posting of images, video and tweets by citizens in the midst of a natural disaster has provided better information to first responders than information from helicopters and their own officials on the ground. That new source of local information is just starting. It presents a great opportunity for data mining but also for visual reasoning. There are major challenges. How should we enable *intelligent aggregation* of visual and other qualitative data? How should we make *visual and textual summaries* to communicate ideas? How might we use groups to *preprocess data* to support interactive visualizations?

The basic methods of crowd sourcing in use to date are far removed from our goal of using crowds to create meaningful visualizations. Currently, crowds are used for judging the plausibility of statements, for making value judgments, for giving aggregate views, but as yet not for creating visual narratives, intelligent visualizations, or preprocessing qualitative data.

In our group we identified problems that need to be solved to harness crowd power for visual reasoning and qualitative argumentation. Beside aggregation and intelligent summarizing we considered questions of coordination: how should the identification and disbursement of tasks be intelligently managed? Can this be done automatically or can we use Turkers to make themselves into a self-organizing system? How can Turkers be unleashed on a problem and generate requests that when answered in the right way are amenable to visualizing complex problems?

Participants

- Simon Attfield
Middlesex University, GB
- Anastasia Bezerianos
University Paris South, FR
- Sheelagh Carpendale
University of Calgary, CA
- Peter C.-H. Cheng
Univ. of Sussex – Brighton, GB
- Fanny Chevalier
University of Toronto, CA
- Christopher Collins
University of Ontario, CA
- Mary Czerwinski
Microsoft Res. – Redmond, US
- David S. Ebert
Purdue University, US
- Thomas Ertl
Universität Stuttgart, DE
- Brian D. Fisher
Simon Fraser Univ. – Surrey, CA
- Steve Franconeri
Northwestern University – Evanston, US
- Kelly Gaither
University of Texas – Austin, US
- Wayne D. Gray
Rensselaer Polytechnic, US
- Hans Hagen
TU Kaiserslautern, DE
- Petra Isenberg
INRIA Saclay – Île-de-France – Orsay, FR
- Tobias Isenberg
INRIA Saclay – Île-de-France – Orsay, FR
- Daniel Keefe
University of Minnesota – Duluth, US
- David Kirsh
University of California – San Diego, US
- Jörn Kohlhammer
Fraunhofer IGD – Darmstadt, DE
- Heidi Lam
Google Inc. – Mountain View, US
- Bongshin Lee
Microsoft Res. – Redmond, US
- Chris North
Virginia Polytechnic Institute – Blacksburg, US
- Catherine Plaisant
University of Maryland – College Park, US
- Margit Pohl
TU Wien, AT
- Huamin Qu
The Hong Kong University of Science & Technology, HK
- Kamran Sedig
Univ. of Western Ontario, CA
- Jinwook Seo
Seoul Nat. University, KR
- Christian Tominski
Universität Rostock, DE
- Xiaoru Yuan
Peking University, CN
- Michelle X. Zhou
IBM Almaden Center – San José, US

Decision Support and Operational Management Analytics

David Ebert
Purdue University
ebert@purdue.edu

Brian Fisher
Simon Fraser University
BFisher@sfu.ca

Paul Kantor
Rutgers University
Paul.Kantor@rutgers.edu

Carolyn Watters
Dalhousie University
cwatters@dal.ca

Taken in isolation, algorithmic "data sciences" approaches and human-centred "visual analytics" methods hold great promise for operationalizing archival datasets and streaming real-time data in support of strategic and operational decision-making across a broad range of human activities. When used in conjunction, computational and human-centred visual information systems have even greater potential.

This minitrack builds upon earlier HICSS minitracks on visual analytics, mobile computing, and digital media at scale to focus more closely on the translation between real-world decision-making and computational methods. Papers selected for this minitrack include two **application-focused studies:** "Early Warning of Impending Oil Crises using the Predictive Power of Online News Stories" *and* "Simulation-based Visual Layout Planning in Advanced Manufacturing"

A second two papers focus on **computational methods and toolkits** whose design is influenced by application studies: visualization methods and toolkits such as "Aperture: An Open Web 2.0 Visualization Framework" *and* "Enhancing Scatter Plots Using Ellipsoid Pixel Placement and Shading"

The final two papers discuss the **translational research** process itself: "The Role of Reasoning in Visual Analytics" *and* "Case Study: Successful Deployment of Industry-University Collaborative Visual Analytics Research"

Together, these papers address the range of research, development, and training activities that are necessary for the development and introduction effective use of information systems to support the effective use of large and complex datasets for real-world decision-making. The methods discussed in this minitrack will have applications in other cognitive tasks that include scientific research, business analytics, health sciences, environmental science and engineering R&D.

The operational aspects of this track also have a range of applications in coordination, command and control of complex human activities such as disaster relief, law enforcement, and anti-terrorism add the constraints of real-time performance and distribution of planning to the challenges faced. The common element in all of these situations is the need to respond to the challenges and opportunities provided by the increasing availability of diverse data to optimize the performance of organizations.

We will wrap up the session by inviting attendees to join in a stimulating panel discussion about the value of combining data and visualization to advance the state-of-the-art for these and related problems. We hope you will join us in this important discussion that will shape the future of research on Visual Analytics in the new era of BIG DATA.

1530-1605/12 $26.00 © 2012 IEEE
DOI 10.1109/HICSS.2013.342

Learning and Law Enforcement: How Community-Based Teaching Facilitates Improved Information Systems

Kaethe Beck
Purdue University
kaethe@purdue.edu

Scott Beamon
Indiana State Police
sbeamon@iifc.in.gov

Edward Delp
Purdue University
ace@purdue.edu

David Ebert
Purdue University
ebertd@ecn.purdue.edu

Abstract

To keep pace with the dynamic environment of information systems, it's necessary to prepare the next generation of the workforce for entry into this atmosphere. Department of Homeland Security Center of Excellence: VACCINE has partnered with the INGang Network, a component of the Indiana Intelligence Fusion Center to facilitate the best preparation for students and assist with some information system issues in the law enforcement industry. Exposing students to real-world applications not only facilitates the problem-solving process with respect to information systems, it also makes the student aware of the prevalent system issues. By participating in community-based teaching with law enforcement officers, the students gain a better understanding of the end user which allows them to design better information systems. There is an initial learning curve associated with integrating in any community, having a close working relationship with INGang has lowered the barrier to entry, ultimately allowing for a better information system to be created.

1. Introduction

Established in July of 2009, the Visual Analytics for Command, Control, and Interoperability Environments Center (VACCINE), along with its co-lead, Rutgers University, has served as the Department of Homeland Security's (DHS) Center of Excellence in Command, Control and Interoperability. VACCINE's mission focuses on creating methods and tools to analyze and manage vast amounts of information for all mission areas of homeland security, including our first responders and law enforcement officials. VACCINE accomplishes its mission through an integrated program of research, education and outreach, spanning the disciplines of visualization and computer graphics, engineering, computer science, geographic information systems, cognitive psychology, information technology, and emergency management and public safety. In pursuit of the center mission, VACCINE has partnered with over 40 first responder and law enforcement entities in order to ensure that our software tools enhance their ability to obtain, process, and gain insight from information. This problem driven research approach not only ensures real-world solutions and enhances transition to practices, but also provides greater research challenges. Our approach also greatly enhances the educational experience of our students; in initiating these partnerships, the graduate students of VACCINE are provided the opportunity to witness first-hand, the struggles with information systems currently in place for first responders and law enforcement officers. While VACCINE has numerous partners engaged in this problem-driven research, the Indiana Fusion Center and Indiana Gang Intelligence Network (INGang Network) have worked closely with our students to explain the struggles associated with information sharing and solution development.

VACCINE focuses on the research, development, and deployment of interactive visual analytic environments for communicating and disseminating information and deriving insight from the massive data deluge. Visual analytics is defined as the science of analytical reasoning facilitated by interactive visual interfaces[1]. Part of the mission of the center is to help decision makers make sense of the sea of text, sensor, audio, and video data by developing powerful analytical tools and interactive visual decision making environments that enable quick, effective decisions as well as effective action and response based on available resources. VACCINE integrates data and analysis into interactive visual displays to enable users to make discoveries, decisions, and plan action; in the project with the INGang network, that interactive visualization and analysis tool took shape in the form of GARI.

2. Background

While community-based teaching is nothing new, the concept is being applied to novel fields and disciplines such as software engineering and design[2]. Historically, community-based teaching

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.609

has most often been seen in the medical field. In 1993, *Tomorrow's Doctors* recommended an increase in community based teaching for undergraduates entering the medical profession[3]. This endorsement in correlated to a change in the methodology of preparing students for an occupation in the medical field, and is now common practice.

In applying this same approach to software engineering and system design, similarly to medical students, the individuals learn how to cope with patients or in this case, end users. Students routinely report they enjoy the community based teaching and feel they are better prepared for the workforce. In providing the experience of an internship without the time constraints, students can continue progress towards their degree, while gaining the real-world experience so many corporations and entities value – and for good cause. The practical application the students experience during this community based teaching requires that they increase their communication, leadership, and interpersonal relationship skills. In the field of software engineering or computer science, this can mean the difference when competing for a sought after position. Many universities require a speech or communications course in the engineering and computer science fields in an attempt to secure this skill among graduates whose stereotype is reserved and more comfortable communicating through machines. There is virtue in the ability to explain technical jargon in a language that individuals with a non-technical background can understand. Being able to speak technically and explain difficult concepts on a general level is a coveted skill. In participating in community based learning with the INGang network, these students have built their skill set to be better communicators, leaders, and ultimately, better system designers.

VACCINE has developed a strong partnership with the Indiana State Police, the Indiana Fusion Center, and in particular with the INGang Network. The INGang Network, which is run through the state fusion center, was developed in order to have a cohesive network to share information regarding gangs and gang activity among law enforcement officers. The creation of INGang is a perfect example of the issues that law enforcement officers encounter with respect to information systems and why it is critical to train the future system designers well through community based instruction.

Knowledge and information transfer among law enforcement officers is not facilitated by the current databases and information systems available. As such, the Indiana Fusion Center and State Police created a network of individuals dedicated to sharing gang related information. The creation of INGang is for the sole propose of transferring and propagating information. This is a prevailing theme across law enforcement agencies. There is no standardized database or record management system; each department or entity has selected their own, some have even custom designed systems to meet their individual needs. VACCINE graduate students have the opportunity to work with the law enforcement officials to assist them with their information systems and in particular, extracting information from them in a useful, productive manner.

3. Related Work

The concept of community-based teaching is common in a number of disciplines such as education or medical professions – for good cause. It is a common belief that there is no better training than experience. It is, however, not unheard of in other science disciplines. Purdue University's College of Engineering has created a program called EPICS – Engineering Projects in Community Service. The idea of EPICS is to provide students an opportunity to see how their work can impact the community, and teach them something along the way[4]. The projects can be quite long term (up to 3.5 years in some cases), and are almost always multidisciplinary in nature. The program went national in 1997 in order to reach a wider range of students. In this pursuit, Butler University attempted to adapt the current structure of the program in 2001 to apply to Computer Science, specifically for the field of Software Engineering (a group of students who are likely to develop information systems in the future)[5]. Butler was in the process of creating a software engineering degree and they wanted the students to have the opportunity to act with real customers on a deeper level than an internship would provide. They also wanted this relationship to be a longer-term time commitment than internships traditionally last. In this manner, the students could experience what a career in software development would mean.

After completing the first year of this program at Butler, the students were surveyed as a method of determining the effectiveness of community-based teaching. As expected, they found the majority of the students valued the experience – 92% even found that the EPICS program had a substantial impact on their customer awareness[5]. This is significant in the field of information systems. In order to design a productive, successful information system, understanding the end user or customer base is a critical step. If students feel that community-based teaching makes them more aware of the end user, that is one step closer to developing a better information system.

The original EPICS program at Purdue University also surveys its students regularly. In 2012 a survey was distributed to every alumnus that had a registered address in the school's database. Of the 2500 or so solicitations to participate, 528 surveys were completed. The alumni who participated varied in their majors and number of semesters involved in the EPICS program. From the survey, more than 70% of the alumni felt that the EPICS program had "some, large, or very large" impact on that individuals performance as an employee once they entered their professional field. Additionally, 20% claimed that EPICS actually influenced their selection of a career and went on to explain how[4].

4. GARI

The INGang Network and VACCINE have entered a two year pilot program in order to collaboratively develop tools designed to facilitate the transfer of information among the INGang Network. Our initial tool in its testing phase is the Gang Graffiti Automatic Recognition tool, or GARI[6]. This application was designed to catalog and categorize gang graffiti images. Gangs and gang violence are a major issue across the country. With some 33,000 violent gangs encompassing 1.4 million members, the ability to convey information among law enforcement officials is a necessity[7]. Gang graffiti has unique symbols, colors, and methodology to indicate threats, meetings, or other messages. By creating a tool which can essentially be used as a repository for information to decode gang graffiti, VACCINE students and law enforcement officers have created a new information system to facilitate this transfer of knowledge. VACCINE students were integral to designing the system. In working closely with INGang, the students were able to experience for themselves the struggles and frustrations of the current information systems and better understand the issues facing end users.

After learning of the issues with sharing and categorizing information related to gang activity, the students could see that the current information system would not suffice to provide the appropriate information and, moreover, would not allow for that information to be easily shared among law enforcement officials. GARI allows for the capture of graffiti images on Android and iPhones (expected late summer 2013). The application then uploads the images to the GARI server that, if so requested, will run an algorithm to test the image against other known images in the system in order to find similar images. Each image can be tagged with various levels of information – the meaning of a symbol, the associated gang, officer comments or notes, etc. The tool will then allow any other member of INGang who has installed the software to view the images and annotation. There is a map tool for looking at the geographic distribution and placement of graffiti – the software will even use the GPS location of the phone to display the graffiti images within a selected radius.

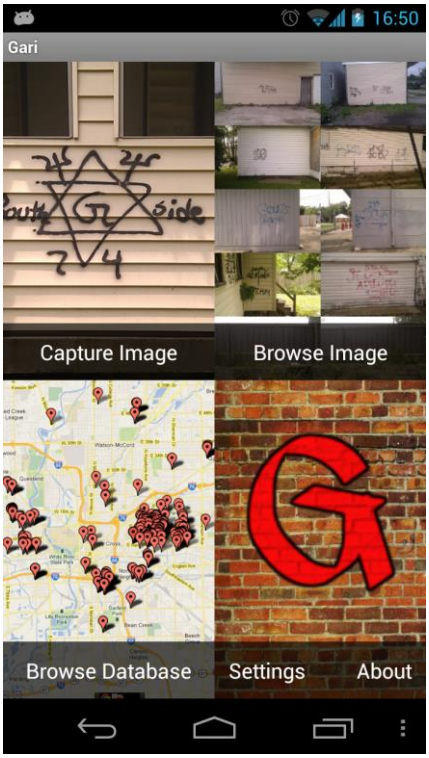


Figure 1: GARI app main screen

5. Student Experience

In order to look at the future of information systems, it's critical to look at the individuals who will be developing those systems. VACCINE students have a unique experience in their education as they are expected to produce a deliverable and deal with an end user/customer, in this case the INGang Network. These very same students will recall the frustration and issues they encountered and overcame when examining the different record management systems from county to county, or attempting to install a program designed to tunnel to a server on a secure network. These experiences have made them aware of the common and prevalent issues that propagate throughout information systems that are critical to safety.

6. Conclusion

Based on the experience in interacting with the INGang network in combination with research related

to community-based teaching, the approach appears to be an excellent method of preparing students for issues they will encounter and could avoid by appropriately designing information systems. As information systems can determine so many facets of productivity in the workforce, it behooves educators charged with preparing the next generation of professionals, to find the best method of educating and training students to develop and design these systems. The VACCINE partnership with the INGang Network arose from the need for information sharing outside of the existing systems, as they were not well designed. By allowing the students to facilitate a new system, the INGang network has also trained a better information system expert. This initial step is the start of the a Midwest partnership that will explore challenges at agencies of various sizes and structures in order to ensure the students are well-rounded in their exposure to the information systems available in the law enforcement arena while exposing them to any number of issues from a lack of network to severely limited bandwidth. With this increased exposure and practical experience, the students are well prepared for a globally competitive, highly distributed workforce.

7. Acknowledgements

This work is supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003.

8. References

[1] J. J. Thomas, K. A. Cook, and National Visualization and Analytics Center, *Illuminating the path*. Los Alamitos, Calif.: IEEE Computer Society, 2005.

[2] J. A. Cantor, "Experiential Learning in Higher Education: Linking Classroom and Community. ASHE-ERIC Higher Education Report No. 7.," ERIC Clearinghouse on Higher Education, Graduate School of Education and Human Development, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 10036-1183 ($18, plus $3.75 postage and handling)., ISBN-1-878380-71-0, 1995.

[3] K. Coleman and E. Murray, "Patients' views and feelings on the community-based teaching of undergraduate medical students: a qualitative study," *Fam. Pract.*, vol. 19, no. 2, pp. 183–188, Apr. 2002.

[4] J. L. Huff, W. C. Oakes, and C. B. Zoltowski, "Work in progress: Understanding professional competency formation in a service-learning context from an alumni perspective," in *2012 Frontiers in Education Conference Proceedings*, Los Alamitos, CA, USA, 2012, vol. 0, pp. 1–3.

[5] P. K. Linos, S. Herman, and J. Lally, "A service-learning program for computer science and software engineering," in *Proceedings of the 8th annual conference on Innovation and technology in computer science education*, New York, NY, USA, 2003, pp. 30–34.

[6] "The Graffiti Code Breaker | DiscoverMagazine.com," *Discover Magazine*. [Online]. Available: http://discovermagazine.com/2012/sep/25-the-graffiti-code-breaker#.Ub0qFvm1G6M. [Accessed: 16-Jun-2013].

[7] "2011 National Gang Threat Assessment," *FBI*. [Online]. Available: http://www.fbi.gov/stats-services/publications/2011-national-gang-threat-assessment/2011-national-gang-threat-assessment. [Accessed: 16-Jun-2013].

Multi-aspect visual analytics on large-scale high-dimensional cyber security data

Victor Y Chen, Ahmad M Razip, Sungahn Ko, Cheryl Z Qian and David S Ebert

Information Visualization published online 28 May 2013
DOI: 10.1177/1473871613488573

Info Vis

Article

Information Visualization
0(0) 1–14

DOI: 10.1177/1473871613488573
ivi.sagepub.com
SAGE

Multi-aspect visual analytics on large-scale high-dimensional cyber security data

Victor Y Chen[1], Ahmad M Razip[2], Sungahn Ko[2], Cheryl Z Qian[3] and David S Ebert[2]

Abstract

In this article, we present a visual analytics system, SemanticPrism, which aims to analyze large-scale high-dimensional cyber security datasets containing logs of a million computers. SemanticPrism visualizes the data from three different perspectives: spatiotemporal distribution, overall temporal trends, and pixel-based IP (Internet Protocol) address blocks. With each perspective, we use semantic zooming to present more detailed information. The interlinked visualizations and multiple levels of detail allow us to detect unexpected changes taking place in different dimensions of the data and to identify potential anomalies in the network. After comparing our approach to other submissions, we outline potential paths for future improvement.

Keywords

Interactive visual analytics, semantic zooming, pixel oriented, multivariate visualization, geospatial analysis, interaction design

Introduction

We designed and developed a visual analytics (VA) system "SemanticPrism" to address the large-scale, high-dimensional cyber situation awareness problem arisen by the VAST 2012 Mini-Challenge 1.[1] The challenge is a "big data" problem. A large enterprise network, named the Bank of Money with approximately 1 million machines, generated approximately 160 million multidimensional data logs (e.g. geographic location, time, activity, policy, machine class/function, and number of network connections) in 2 days. For proper analysis, the analyst must be able to see and compare all of these different dimensions at multiple granularities (e.g. enterprise to individual machines in individual offices). To meet these requirements, we developed the VA system SemanticPrism to visually analyze the given data from three perspectives: spatiotemporal distribution of machines and their health, overall temporal trends, and pixel-based IP blocks. All these visualizations are interlinked and provide 2–4 levels of semantic zooming. The analyst can not only grasp the overall situation of the enterprise network, but also drill down to read more detailed information of regions, offices, and even the level of individual computers. With SemanticPrism's comprehensive visualizations and interaction, we were able to discover all anomalies hidden within the large dataset and won the award of "Outstanding Integrated Analysis and Visualization."

[1]Department of Computer Graphics Technology, Purdue University, West Lafayette, IN, USA
[2]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[3]Department of Art and Design, Purdue University, West Lafayette, IN, USA

Corresponding author:
Victor Y Chen, Department of Computer Graphics Technology, Purdue University, West Lafayette, IN 47907, USA.
Email: victorchen@purdue.edu

While designing the SemanticPrism, rather than simply solving this particular challenge, we tried to explore a more general approach to face real-life large-scale high-dimensional datasets. In this article, we review related literature and then discuss the design considerations of SemanticPrism in terms of scalability, dynamic situation awareness, visualization, and novelty. Furthermore, we compare our approach to other submissions and outline potential paths for future improvement.

Previous research

Previous study has explored a variety of approaches to handling the problems "big data" create. As "big data" are often complex, multidimensional, and multivariate, many studies have discussed different methods of visualizing large datasets. Fua et al.,[2] for example, described the use of hierarchical parallel coordinates to visualize large multivariate datasets. Keim[3] and Oelke et al.[4] have shown the use of pixel-based visualization to fit the huge data space into a small screen space. Keim et al.[5] also developed a hybrid technique that is scalable with "big data" visualization. Some approaches also include using multiple linked views to visualize "big data."[6] A summary of recent visualization techniques of large multivariate datasets is available by Keim et al.[7] Looking beyond the academic community, a number of commercial VA systems that process big data already exist within the marketplace and have been evaluated.[8]

We believe that analyzing this high-dimensional data requires multiple linked views from different perspectives and different levels of granularities and detail. Zoomable user interfaces (ZUIs)[9] allow users to work with a large virtual space and navigate through it by zooming. Semantic zoom[10] lets the user see different representations of the data at different zoom levels. Weaver[11] stated that "semantic zoom is a form of details on demand that lets the user see different amounts of detail in a view by zooming in and out." This method has been widely used to provide smooth analysis experiences through interaction (such as in malicious network objects[12] and relational data[13]). We utilize semantic zoom in this system to visualize the different properties of big data at different granularities.

As the number of connected machines and the possibility of network attacks increased, many experts have developed various network monitoring tools that include a number of different visualization techniques. One of the more popular methods of visualizing network data is to use graph-oriented visualizations[14] where machines are mapped to nodes and links connecting those nodes with different characteristics, such as thickness and color, represent relations among nodes. Boschetti et al.,[15] for example, implemented a graph-oriented approach to monitor network traces and detect anomaly. Iliofotou et al.[16] proposed the use of traffic dispersion graphs (TDGs) as a way to monitor and analyze network traffic by modeling the social behavior of hosts. In many systems, different types of node placement algorithms have been used. Among those, the force-directed graph drawing method[17] and bipartite algorithms[18,19] have been widely adapted.[20,21]

A different visualization approach is the pixel-based visualization approach, and it provides some advantages over the traditional graph-oriented visualization of the computer network. Oelke et al.[4] and Keim[3] introduced the pixel-based (pixel-oriented) visualization technique to maximize the screen space for visualizing large amount of data. In this technique, the entire visualization space is equally divided into squares or rectangles, called pixels, where each data element is assigned. Then, a predefined color map is applied to represent the range of the data attributes. The pixel-based visualization technique has been widely used in various applications and research in which the datasets are very large and multivariate. Borgo et al.[22] presented how the usability of the pixel-based visualization varies over different tasks and block resolutions. Oelke et al.[4] studied visual boosting techniques for pixel-based visualization such as halos and distortion. Ziegler et al.[23] presented how the pixel-based visualization helps analysts gain insight for long-term investments. Ko et al.[24] demonstrated how sales pixel matrices can be used for analyzing competitive advantages of companies. Panse et al.[25] discussed the effectiveness of the technique in PixelMap when the datasets consist of very large number of points. In our study, we employ the pixel-based visualization method to explore the IP address space of the challenge data.

According to MacEachren and Kraak,[26] geospatial datasets are fundamentally different from other kinds of information in at least three ways: structured spatial variables, meaningful location names, and emergent behaviors. To visualize the special geospatial aspects of the data, the popular technique with many geographical information systems (GIS) is applied in order to plot points on a geographical map. This technique has been employed in a variety of domains, such as crime mapping,[27] public health,[28] and social science.[29] Other techniques to visualize large spatial datasets include PixelMaps,[5] which is designed to combine both clustering and pixel-based visualization to plot points on the map. This technique copes well with dense geographic data and prevents data point

overlaps. In this study, we used the point-plotting technique that has more expressive power of identifying individual offices inside a region.

SemanticPrism system

Although the data provided for this challenge are artificially generated, this challenge simulates a real scenario. The SemanticPrism system was designed from the beginning, not only to solve this particular challenge, but rather explore a general approach to achieve cyber situation awareness in a real-life scenario while facing large-scale multidimensional datasets.

Dataset and tasks

The data from the challenge include a geographical map (an image file), a KML (Keyhole Markup Language, an XML notation for expressing geographic data) data file to define regions, and two large spreadsheet tables. One table contains basic information of all computers, including its IP address, business unit, facility, latitude and longitude, machine class (server, automated teller machine (ATM), or workstation), and machine function. The other table contains 160 million records of computer status logs. Each record contains information of a computer's IP address, number of connections (NOCs), policy status, activity flag, and the log timestamp. The policy violation status is a discrete data measurement of the health status from normal to severe (labeled from 1 to 5 to indicate severity). The activity flag categorizes the machines by different types of activities (labeled 1–5). The NOC log is a discrete data value (range from 1 to 100 in the current dataset). The data provide a 2-day snapshot of the health status of all computers in the whole organization with 15-min intervals (192 total time periods).

We first brainstormed several fundamental tasks and their consequent data queries based on the need of cyber security situational awareness: (1) See the geospatial distribution of computers and tell whether there exists any spatiotemporal status pattern. This task requires the system to query computer logs grouped by offices and time periods. (2) Visualize the trends of computer status at different granularities from the overall network to individual machines. The system has to count the numbers of computers of different statuses over time. Also, the NOCs need to be aggregated to get the maximum and average values. (3) Study the spatiotemporal pattern of status over the IP address space. The challenge data do not provide the structure of the network. In the Internet, computers' IP addresses can be classified as Class A–C based on their four 8-bit numbers. Such classes partially reflect the network structure. Computers within the same block (especially a Class C block) are likely to be in one subnetwork. Data also need to be aggregated based on Class C IP blocks. (4) Investigate individual computers through their log history. It requires searching the full history logs of a computer through its office or IP block. The data should be indexed by the computers' offices and IP blocks to speed up data query.

Data transformation and aggregation

Our first challenge was to transform the large-scale data and make it efficient for interactive analysis. Even when using a MySQL database, the direct querying of such a large dataset remains inefficient and can take hours to provide an aggregation number (e.g. the total number of computers of a given status). To speed up the data query and enable a responsive system performance, we created additional indices and aggregated data into new tables. The process to treat the data as a data cube, precompute the aggregation values along necessary dimensions, and store the aggregated values into several tables is in line with the online analytical processing (OLAP) approach.[30] Precomputed aggregated values include the number of computers for each policy status and activity and the maximum and average NOC at a given time for each Class C IP blocks. Querying and processing the data for a group of time series curves only take a small fraction of 1 s; so, most interactions in the system can produce instant results.

For this challenge, the data provided are a static file, meaning aggregation is only done once. In a real-life implementation, such an aggregation and preprocessing could be performed while collecting data on the fly.

System structure and development platform

Although the raw data from the challenge only cover 2 days and are 8 GB in size, in essence, these data are streamed and can eventually become truly big data as time goes on. In order to correctly and effectively manage these big data, SemanticPrism uses a client–server architecture designed as a web application. Clients in the front end visualize the data but do not retain a copy of the whole dataset.

As a prototyping system for research purposes, we built the system with Adobe Flash, PHP, and MySQL. The client-side application uses Flash that is currently supported by most web browsers and is an efficient platform for an application with rich interactions and dynamic graphics.[31] The web server runs PHP to process data requests. The communication between Flash and PHP is done through action message format (AMF). The PHP web application is hosted in a shared server. The MySQL database server resides in

an Intel Xeon 3.0-GHz server, with 64 GB of memory. The Flash client can run smoothly on a notebook computer (Intel 2.6 GHz Core 2 Duo CPU with 4GB RAM).

Visualization and interaction design

The choice of visualization and interaction design should be based on the nature of data and the problems faced. We wanted SemanticPrism to run on a notebook computer to allow maximum freedom of working location. With limited screen space, the analyst should be able to navigate through different dimensions of data, drill down to investigate details, become aware of significant changes, and identify anomalies. To enable exploration of the data from different data dimensions, we chose to use a multiple linked views approach: different types of information are visualized using the geospatial map, time series curves, and pixel-oriented visualization views. Each of the views has multiple visualizations to present different levels of details. We chose semantic zoom as the basic interaction technique to navigate through these visualizations.

Geospatial–temporal visualizations

The default view of SemanticPrism is a geographic visualization with a time slider designed to visualize the computer status at a given time (Figure 1).

Offices around BankWorld are marked on the map as square dots. Different icons are used to distinguish types of offices: small squares represent regular branch offices, squares with one boundary line represent the regional headquarters, and squares with two boundary lines represent headquarters and data centers. The map provides an overview of the most critical information at the current time. Different colors are applied onto the squares to indicate the maximum policy violation status of the computers within the office at that particular time. The colors, varying from yellow to orange to dark red, represent the maximum policy violation status (from 1 to 5) for all computers in the office. The reason the most severe computer health is shown, instead of the average health status, is to draw the analyst's attention to a problem the very first moment the problem arises. This color setting is consistent in representing policy status across different visualizations and functions in SemanticPrism. If there is no log from an office at a certain time, it means that all computers are off-line. In this case, the office is represented by the black color.

To update the statuses of all offices to a different time period, the analyst can drag and slide the time slider to a new time mark along the bar (Figure 1). Additionally, the analyst can input the desired time (period number, ranging from 1 to 192) in the time input slot or use the time step forward/backward button to advance to the next time slot or roll back to the previous slot.

Layers on the map

The SemanticPrism map (Figure 1) uses several layers to stack different dimensions of information together. The analyst can selectively turn on or off these layers.

The dataset's Bank of Money is a global organization spanning eight time zones of BankWorld. In order to present the different time zones and local times for the distributed offices, we provide an overlay time zone layer. It is a half-transparent layer with vertical strips of gray shades. Time zones within the early morning or late night are in darker shades to hint there is less sunlight. Although at night machines tend to be less active, we also wanted to draw the analysts' attention to the fact that some crucial attacks might take place during such time periods.

When there are many offices in a relatively small area, the area can be cluttered with many dots of offices. To improve that condition, we created two layers to highlight offices containing computers with a selected policy status or activity status. The analyst can select a policy from the policy drop-down menu to turn on the policy layer. With one policy selected, any office with a computer that belongs to this selected policy status will be highlighted as a blinking red/blue square. The size of the blinking square reflects the number of computers with the selected policy status. The blinking effect is good at drawing the analyst's attention even if the dot is small. Similarly, the analyst can turn on the activity layer through the activity menu (Figure 1(b)). The system uses orange/green blinking squares to highlight all offices that have computers with the selected activity. Correspondingly, the sizes of the squares reflect the numbers of computers that are involved in such an activity. As time progresses, the change in size and location of blinking squares indicates to the analyst the trends of the policy status and activity.

If both blinking layers are turned on, the blinking squares become messy and hard to read. To effectively reduce the visual clutter and visualize the policy status and activities together, we use a Kernel density estimation (KDE)[32] heat map as an alternative method to visualize the geospatial distribution of a selected policy status or an activity. The heat map is computed based on the density of computers matching the selected status in an area using a clustering algorithm. The heat map uses blue shades: the darker the shade, the more computers there are that match the selected status. This layer is stacked under the blinking layer, as shown

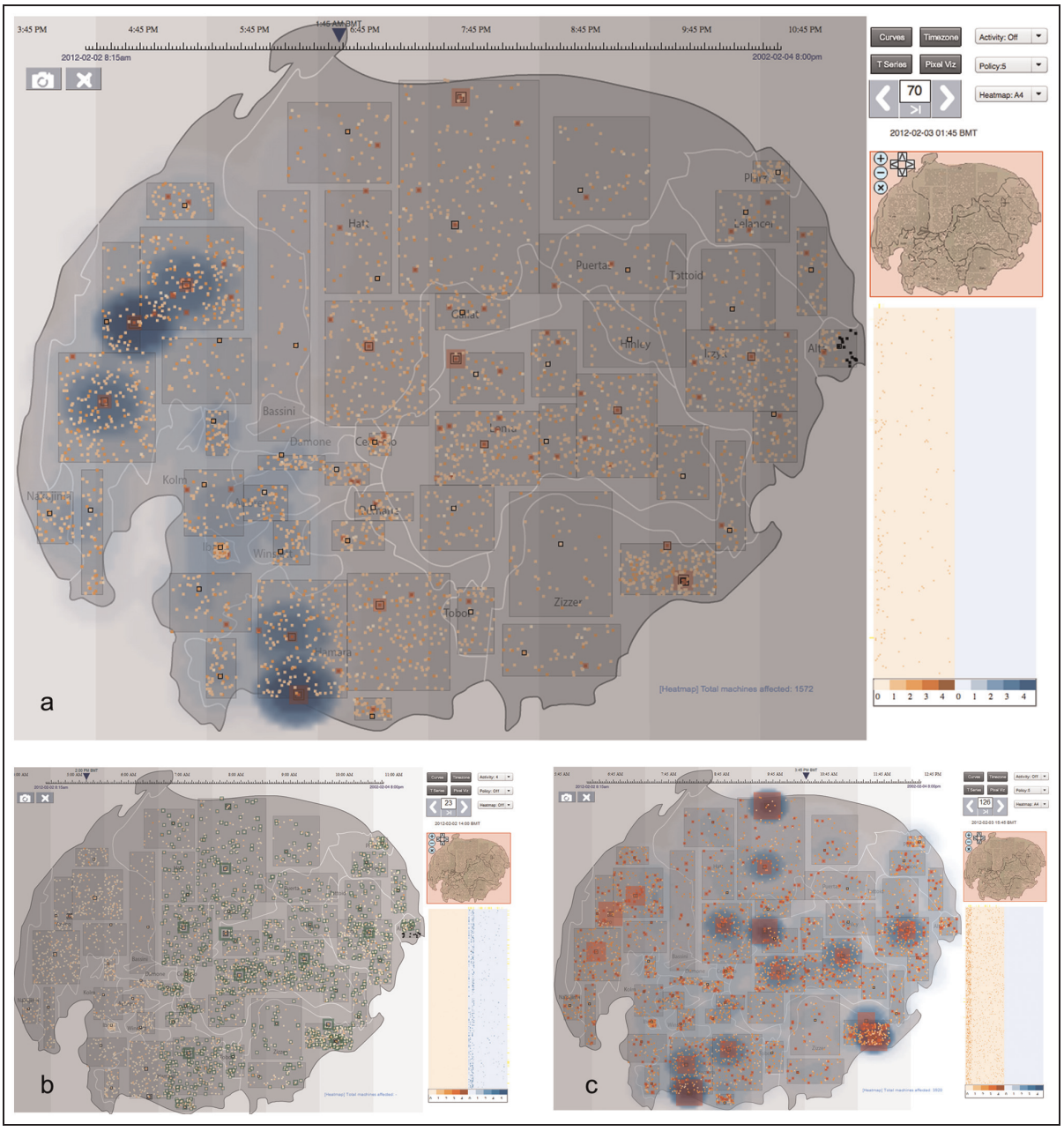


Figure 1. SemanticPrism map view: (a) the policy layer with the activity heat map at 1:45 a.m. BMT, 2 March 2012; (b) with only the activity layer on at 2 p.m. BMT, 2 February 2012; and (c) the policy layer with the activity heat map at 3:45 p.m. BMT, 2 March 2012).

in Figure 1(a). With the combination of the two layers (KDE and blinking), the analyst can read the policy status and activity at the same time.

Zoom and navigate

Through the right side's navigation panel, the analyst can zoom in/out and pan to navigate in the map. A red square shade (right corner of Figure 2(b) and (c)) is used in the panel to indicate current visible area of the map. While zooming in, the space among office dots increases, which can potentially be used to display more information.

In SemanticPrism, the analyst can drill down and investigate the data at different levels of detail (Figure 2) through semantic zooming.[10] Depending on the size of available space, an office is dynamically visualized in one of the following four levels:

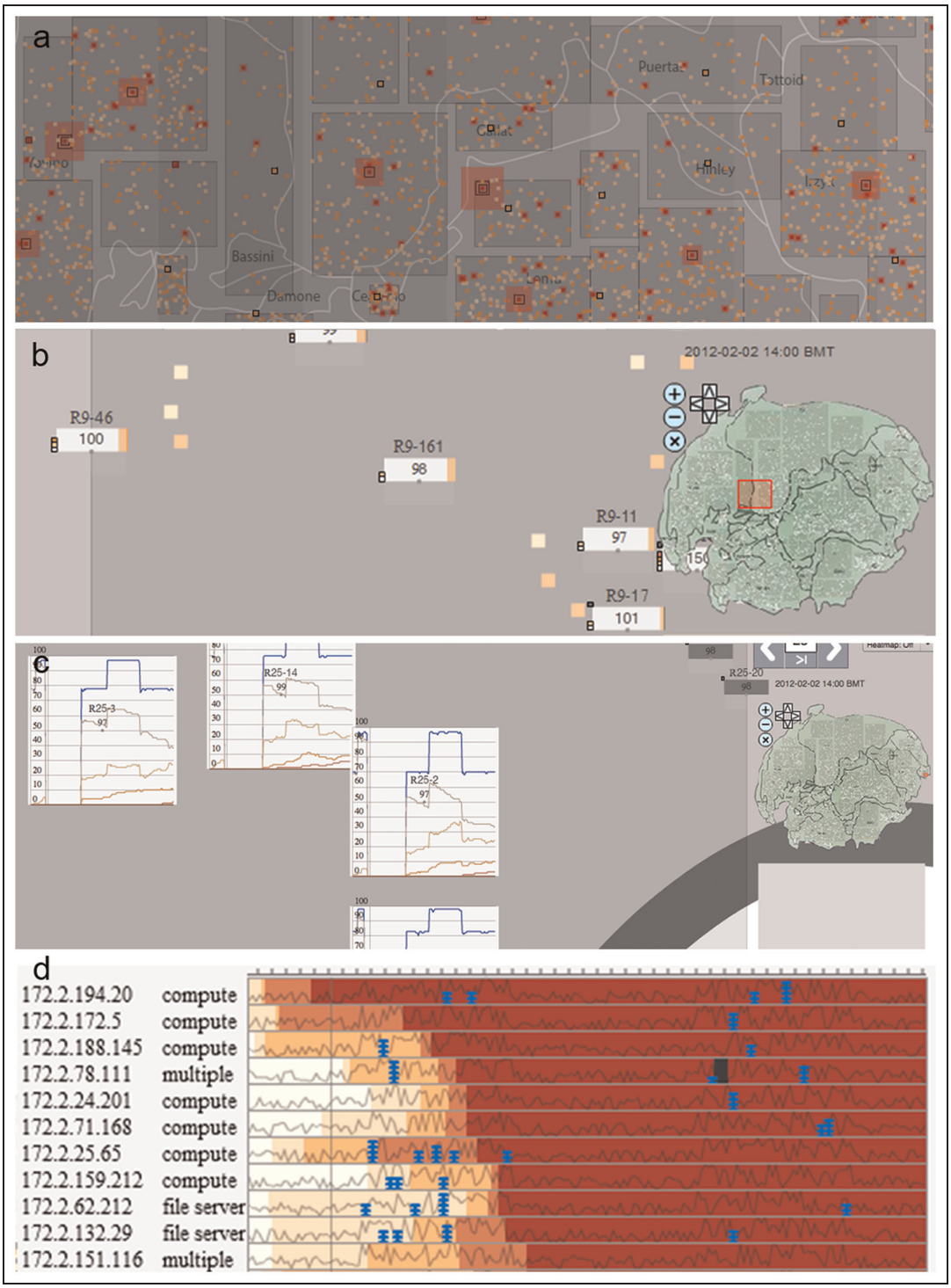


Figure 2. Four levels of semantic zooming on the map. (a) level 1 - offices as dots (b) level 2 - offices as bars to show percentages of problematic computers (c) level 3 - curves to show trends of problems (d) level 4 - history of every computer in one office.

.

Level 1 allows the analyst to visualize an office as an individual dot when using the default full map view or when the space is still quite dense after zooming in (Figure 2(a)). In the full map view, the offices at some areas appear so dense that the square dots may overlap. While zooming in, the squares are also enlarged,

but at a smaller ratio (square root of the screen ratio), to make the spaces among the offices larger. Thus, the density decreases and provides better readability to distinguish these little office dots.

Level 2 (Figure 2(b)) uses a horizontal color bar to show the percentage of computers with different policy statuses, including those that are currently off-line. In this visualization, the analyst may misread those policy statuses with only a very small percentage of computers and assume that they do not exist. To avoid this problem, we used vertical squares to mark if computers with certain policy statuses exist. Also, the analyst can see the office name and total number of computers in the office. At this level, the size of the bar remains the same during zoom in until the space is big enough to show Level 3 details.

Level 3 (Figure 2(c)) indicates the growth curves of all policies in the office where the *x*-axis presents the temporal direction, and the *y*-axis shows the number of computers. This graph contains six different curves, displaying numbers from Policy 1 to Policy 5, and the total numbers of online computers. The standard policy colors are used to distinguish different curves. The size of the graph (200 pixels $\times$ 200 pixels) remains the same when zooming in.

Level 4 illustrates the history status of each individual computer within the office (Figure 2(d)). The history of a computer's policy status is visualized as shades of a red bar. The curve in the middle of the bar shows the NOCs. The computer's activities are visualized as blue bars with stacked horizontal lines. The number of lines represents the activity number. Activity 1 (normal status) is omitted. The analyst can use this visualization to read the finest details of a specific computer in a specific office.

Zooming in/out and panning change the screen dramatically. When the analyst's eyes are focused on one area and there is a sudden change in the visualization, change blindness[33] might take place. The analyst may lose his focus. To avoid that, we integrated animation to permit a more gradual zoom in/out and allow saccadic eye movements[34] to catch up with the changes. Zooming out creates a reverse effect of zooming in. Detailed views will be shrunk until the offices become square dots.

Apart from using the navigation panel, the analyst can directly interact with the map to pan and zoom in/out. Scrolling the mouse's middle wheel zoom in/out of the map. A left mouse drag pans the view. Clicking on an office will directly open Level 4 details of an office. Clicking on the boundary of a region will open the pixel-based visualization of all offices within that region. These offices are laid out in a rectangular array to let the analyst see all offices simultaneously (Figure 4(b)).

Time series curves. While designing the system's information query process, we followed Shneiderman's information-seeking mantra:[35] "overview first, zoom and filter, then detail-on-demand." The SemanticPrism time series curves (Figure 3) provide an overview of the growth trends of policy statuses, activities, and NOCs over the given time period. The default curve view (Figure 3(a)) presents the growth of policy statuses and activities of one class of computers. The analyst can choose to use either a linear or logarithmic scale to draw the curve. The linear scale can intuitively show the overall growth trend. But because of the large number of overall computers, it is hard to read the curve at the early phase of an attack when there are only a few computers affected. The logarithmic scale (Figure 3(b)) boosts the small numbers by adjusting the curves to help the analyst to catch that first moment when a computer is violating a policy.

The time series curve visualization can also be "zoomed in" conceptually. The analyst can narrow down by applying a combination of filters to select certain computer class, computer functions, activities, policy statuses, and NOC to visualize the trends of affected computers.

In SemanticPrism, the user can also create multiple panels (Figure 3(c)), with each containing curves generated by different filters. The analyst can then compare different curves side by side to investigate further.

Pixel-based visualizations. The classification of IP addresses can partially reflect the organization's network structure. Within these data, we also noted that computers within a single office with one class (server/workstation/ATM) belong to the same level Class C. To visualize such an IP address space, we incorporated a pixel-based visualization of IP blocks (Figure 4) to analyze computers in a more detailed classification.

By default, the pixel-based visualization contains one panel showing the selected policy status and activity (the red/blue square on the right side of Figure 1). The selection is done through the same drop-down menu of highlighting policy status and activity. The analyst can expand it to show five panels. Each panel shows the number of computers within an IP block that are affected by each activity and policy (Figure 4(a)). In each of these five panels, the red side shows policy status and the blue side is for activity. Each pixel represents a group of computers in a particular Class C block. The *x*-axis consists of the IP's Class B block (ranging from 172.1 to 172.56), and the *y*-axis consists of the values of Class C blocks (ranging from 0 to 255). The colors of the pixels encode the number of computers that carry the selected policy status or activity flags in the C block.

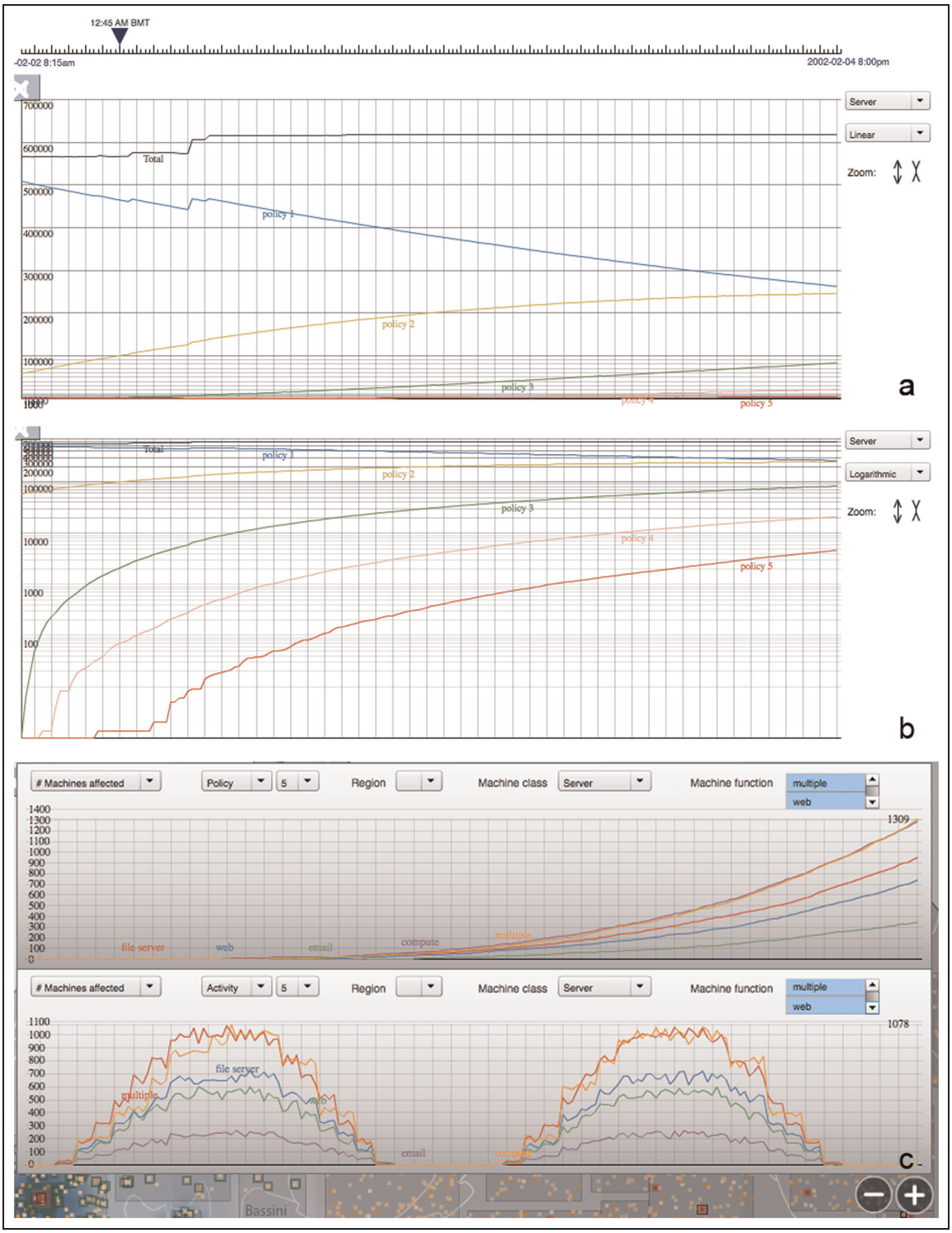


Figure 3. Time series curves in SemanticPrism: (a) linear curve of servers, (b) curves in logarithmic, and (c) dynamic panels to show the curves of policy statuses and activities of selected types of computers.

The IP block pixel-based visualization has three levels of semantic zooming. Hovering the mouse pointer over the inside area of the panel will evoke a zoom-in lens to show enlarged pixels. Clicking of a pixel brings up the time series curve of that C block. Clicking on the bottom *x*-axis of the panel will evoke the system to show the time series curves of all C blocks within the selected B block. The user can choose to see the

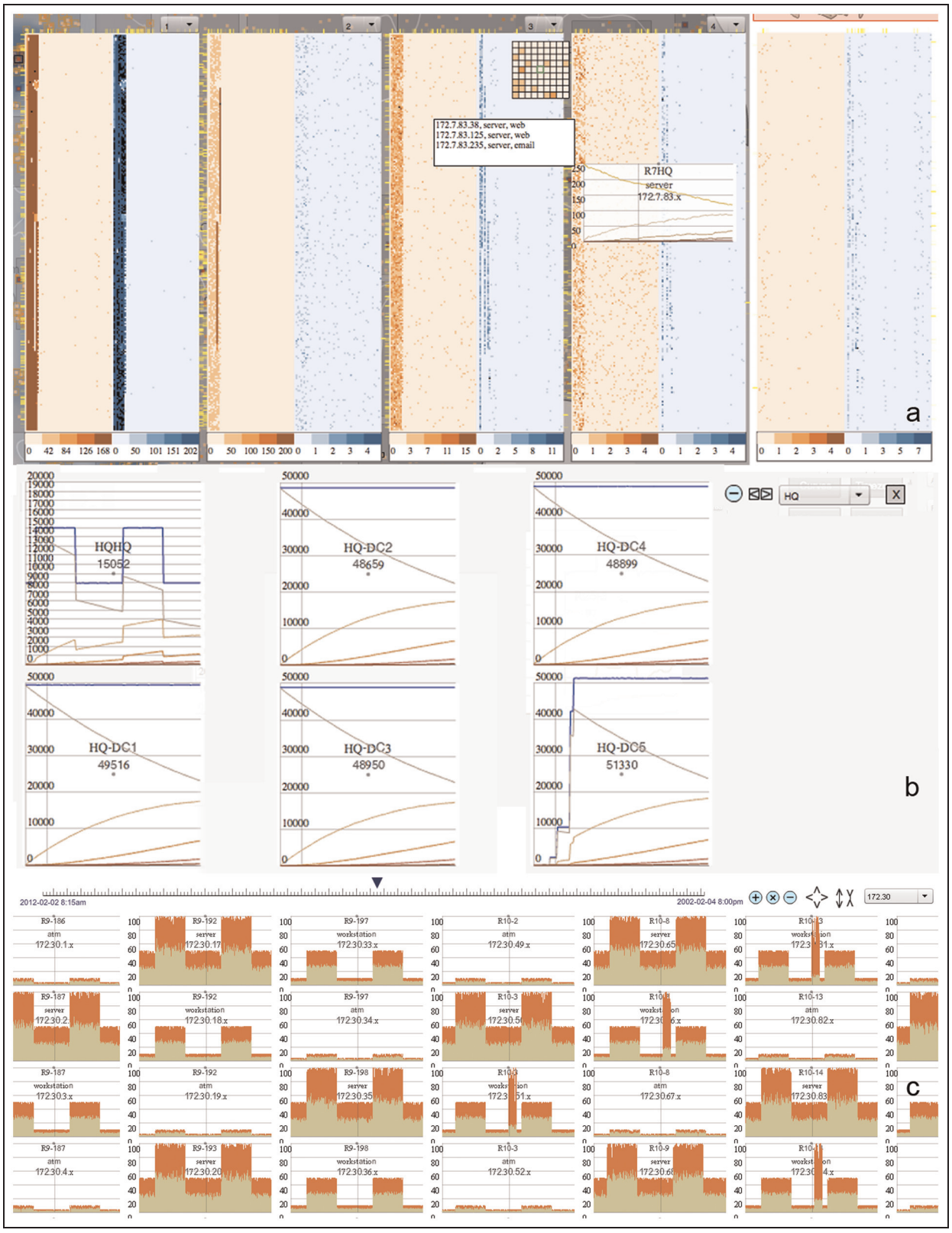


Figure 4. Pixel-based visualizations of IP blocks: (a) five default policy/activity pixel–based visualization panels, (b) offices in a region (Headquarter), and (c) zoom in to show the NOC graphs of all Class C blocks in one Class B IP block.

curves of policy statuses (Figure 4(b)), activities, or NOCs (Figure 4(c)). The analyst can also see Level 4 individual computer histories in the C block (similar to Figure 2(d)) through clicking. The visualization enables the small approach to multiple comparison[36] where analysts can easily investigate the differences of trends in multiple data. For example, in Figure 4, it is easy to notice that some C blocks have an abnormal spike in the middle of the visualization, which is different than the regular on–off office hours. This leads us to investigate abnormal network connections at night.

Situational awareness analysis

In a real-life context, it is essential that immediate actions be taken to prevent the expansion of damage. In SemanticPrism, each of the three visualizations has been used either individually or collectively to support situational awareness.

Detection of problematic computers at an early stage

It is critical to detect the first occurrence of a certain activity or policy violation. The time series curve view accurately presents the statuses over time, including both previous and future growth trends, though it does not pinpoint the location. With the help of the time slider, both the map view and the IP block pixel view can hint to the analyst when and where the first computer fell into policy status 4 or 5. The map view then locates the office, a perspective that is necessary if the prevention action has to be taken on-site. With both the map view and IP pixel view, the analyst can drill down to find the particular computer (IP address), a step that is necessary for remote access and repair. The quickest way to find the accurate occurrence time and location of the problematic computers is to use both the curve view and the map view: anchor the exact time of the first occurrence from the curve view and then switch to the map to find the exact office.

Although it does not appear in the 2012 challenge data, one potential threat worth monitoring is that of abnormal activities happening at off-office hours (e.g. invalid logins or intrusions). Such anomalies can be easily seen from the map by turning on the activity highlight layer and the time zone layer. As time goes on, we can clearly observe how offices become more idle when their local time slots reach the off-working time at 6 p.m. and then become active again at 8 a.m. (Figure 1(b)). The curve view can also help to monitor abnormal activities (Figure 3(c), bottom). Such activities are normally rare. The logarithmic method is useful to show even one instance of occurrence.

Overall trend of policy violations and activities

The curve view shows the growth trend of policy statuses and activities over the time period. By looking at the curves of different classes or regions, the analyst can clearly tell the growth or working pattern of the computers (Figure 3(c)). However, the analyst cannot see the spatial pattern of the spread. With the map view, as time passes, we can see that there are more and more blinking squares, and squares getting larger and larger, which indicate that there are more and more computers under high-risk policies (Figure 1(c)). With the pixel-based visualization, we can see that more and more IP blocks are affected (Figure 4(a)).

We assumed that the IP pixel view could indicate to the analyst the spread patterns of policy violation over the network (e.g. computers within the same IP block or neighboring blocks get affected first), but we have not found any good evidence for this within the current dataset.

Outages in Region 25

Among our three visualizations, the map view accurately shows that there were multiple office outages in Region 25 (black dots on the right side of the map in Figure 1(a)). With the time slider, the analyst can see the development of the outage range and how the offices then recovered (e.g. black dots in Figure 1(c)). Since this was caused by a hurricane, the accurate geospatial locations of these affected offices are useful to indicate the natural disaster's affecting range.

The IP block visualization is not effective in discovering this outage because many computers are turned off at night. The number of hurricane-affected computers does not appear to be distinct or large, so this outage has not been reflected clearly on the time series curve view.

Addition of servers to Headquarter Datacenter 5

The combined use of the time series curve view and pixel-based visualization has helped us to find that many servers were added to the Headquarter Datacenter 5. We can observe a major jump in the number of servers in the curve view (the black total curve in Figure 3(a)). However, SemanticPrism cannot directly link these data change to the individual offices. We have to manually examine offices through the pixel-based visualization. In the zoomed-in view of offices by regions, we can see that there is a significant jump of the number of servers within Datacenter 5 (Figure 4(b)).

Abnormal NOCs at night

Apart from finding the addition of new servers, the pixel-based visualization of all Class C IP blocks also helped to identify another anomaly: abnormal NOC at night. The map view did not provide a way to visualize NOC. Most computers are turned off regularly at night, so the change of NOC is not obvious on the time series curves.

Using the three visualizations to analyze the data, we came to realize that anomalies were usually identified by a combination of different visualizations. No visualization method is really universal. As different components of a system, these visualizations cover the weaknesses of each other and should be used as elements of a tool kit to detect problems in large-scale data.

Feature discussions

Considering the limitation of individual visualizations and the strength of their integration, we started to review the overall pros and cons of SemanticPrism. The following discussions are based on the comments from reviewers of our submission and external questions raised from the VAST workshop presentation and system demonstration.

Scalability

The SemanticPrism has the potential to attack much larger data. The scale of these challenge data can be expanded in several dimensions. The first expandable dimension is time: the data can be extended to months or even years. Other possible data expansions include the addition of more offices, more computers, or more types of activity or policy statuses within the logs.

SemanticPrism's map view only displays the current status. To fit the enlarged time span, the timeline bar might be edited so that the analyst can zoom in/out and navigate semantically. The number of offices will significantly affect the map view as a result of the higher density of office dots. Theoretically, our map can only present a limited number of offices effectively. Since our system uses blinking effects to show alerts, the analyst can still easily see the problematic area even if the map has been fully covered by offices. The heat map layer is fully scalable (O(1)) since it permits the computation and visualization of the relative density of both offices and computers. Our semantic zoom technique can also work with higher office density. While zooming in, the office dots are enlarged at the square root of the screen ratio, which makes their proportions within the screen space much smaller and generates more spaces among the offices. Thus, even if there are many offices close to each other, the analyst still can distinguish each individual office by zooming in further.

The IP block pixel-based visualization will expand (O(n)) when there are more computers (more IP blocks). Since one C block contains up to 254 IP addresses, the number of IP blocks is much smaller (~1/255) than the number of computers.

The scale of time series curves follows the temporal scale as O(n) since the *x*-axis is the time dimension. We can overcome the time span problem by compressing the X direction and use interactions to zoom in and slide along the X direction to see details of the curve.

At last, the system can be converted to handle streaming data. We can change the server side component to only query data for the current moment (or within a certain period of time range), which will help to solve the large time span problem.

SemanticPrism employs layers to permit recognition of multiple properties. If there are an increased number of properties, more layers can be added, but the efficiency of readability may decrease. To improve the analysis, in the future, we might adopt a measurement that integrates multiple properties, such as the "concern level assessment (CLA),"[37] to combine multiple individual layers into one comprehensive layer.

Query data on demand

With SemanticPrism, data are stored in an external database. Only when it is necessary, the client-side Flash application can send a request to the server to fetch a small amount of data. Therefore, a significant size increase of the overall dataset will not necessarily slow down the performance of SemanticPrism. While one is working in the map view and zooming in, only the specific offices within the display area will be checked to see whether they should be expanded to display the next level of details. At Level 2 zooming, the maximum number of offices that will be displayed simultaneously in the screen is 600 (each office needs a space of 30 pixels $\times$ 60 pixels), with each office requiring just six integers for the number of computers under all policy statuses. At Level 3 zooming, the maximum number of offices in one screen is 30 (each office needs a space of 200 pixels $\times$ 200 pixels), with each office fetching data for its time series curves (an array of 192 periods with 6 flags = 1152 integers).

In the pixel-based visualization, if the screen space cannot show all of the items, the user can pan the screen to read the rest. Similarly, the client-side Flash only queries data for those elements that are within the display area.

Visualization design

One of our reviewers indicated that SemanticPrism did not invent any new visualization. The map with multiple layers, the time series curves, and the pixel-based visualization are all very common methods. However, other reviewers stated that these common visualizations are among the most suitable ways to identify anomalies in this challenge. We chose these visualizations because they are very intuitive and the analyst can easily adopt and identify the problems without taking much training.

In this challenge, one main task is to detect anomalies (e.g. virus infection) as quickly as possible. The percentage of problematic computers can be extremely low. Some traditional visualization techniques that can present both the quantity (e.g. use area or length to represent number of problem computers) and the quality (e.g. use colors to represent the policy statuses) may be inefficient to alert the analyst the problem since the size/length of the graph is too small to notice. Therefore, we used several ways to boost the visualizations. In the time series curves, the logarithmic way of drawing the curve boosts small numbers. In the geospatial map, the blinking dots draw the analyst's attention even when the dots are tiny. At Level 2 of office details, we used separate icons to indicate the existence of different policy statuses, as well as using length to represent the percentage of each policy status.

Adobe Flash as the development platform

We would like to spend more time exploring the visualization and interaction design, rather than devoting too much time on coding in the limited competition period. We choose Flash as the platform due to its rich support on graphics, animation, and interaction. Although it is not popular with the development of scientific tools, we found it to fit well with quickly developing functional prototypical systems for research purposes.

Our VAST 2011 challenge submission[38] was also built upon Flash. It was the only submission that used animation to vividly demonstrate the flow of people in different locations. We spent most of our time designing the visualization and interaction. The implementation is relatively simple and straightforward.

Applications built by Flash's ActionScript 3 are compiled and run in the Flash player's virtual machine. The performance is acceptable for our systems. When searching for anomalies with SemanticPrism, most of our interactions are smooth and responsive, and its response time largely falls in the appropriate time limit suggested by Shneiderman.[39] Thanks to our client–server structure and the query-on-demand design, the client-side Flash does not need to handle extremely large amount of data. Most tasks take much less than the acceptable 2-s response time,[39] and many interactions, such as a semantic zoom in the map view, can respond instantly.

Inspirations from other submissions

In the VAST 2012 challenge workshop, we had an opportunity to see other award-winning solutions. All of them are creative and inspiring. This is actually one of the most valuable components in the challenge for us.

Dudas et al.[40] presented a solution that integrates OLAP operations[30] into VA. They used a matrix to display multiple histograms simultaneously. The analyst can perform OLAP operations (drill down, roll up, slicing, and dicing) to manipulate the matrix of curves. When compared to our time series curves, their solution appears superior in two ways. First, the matrix uses two-dimensional (2D) array, which can show five dimensions (column, row, x, y, and stack) of information and concurrently display many histograms. Second, the OLAP operations allow the analyst to generate many curves with simple interactions. In our system, the analyst has to manually select and combine filters to generate the curves. To generate curves containing exactly the same amount of information, our system needs more interactions.

Kachkaev et al.'s[41] solution used a single line to visualize the status change of an office by time. Colors of the pixel in the line present the maximum NOC, the maximum policy status, or the activity flag. These single lines are then grouped into regions. This approach inspired us to note that another level of semantic zooming can be added in our pixel-based visualization. Our first zooming level uses one pixel to display a C block at one time (Figure 4(a)). Then, the next level directly jumps to a 2D histogram (Figure 4(b) or (c)). Kachkaev's one-dimensional (1D) method can be used in between our single-pixel view and 2D curve. At our current second level of curves, we can only display curves for all offices in one region or all C-level IP blocks in one B-level block. This 1D method is compact enough to place a more detailed temporal overview of many regions/C blocks into one screen.

Choudhury et al.'s[37] submission used a machine-inferred variable "CLA," which contains inference rules that embody adductive inferences from parameters including machine class and function, policy status, activity flag, NOC, and time of the day, to compute the concern level of the computers. Our system visualizes different parameters separately. Although our integration of multiple visualizations allows the analyst to see multiple parameters at once, a comprehensive understanding of the combination of parameters is difficult.

SemanticPrism uses simple nodes in the map to show offices. To avoid overlapping, the nodes' sizes are tiny and identical and sometimes hard to read. Pabst's[42] system rearranged office locations to align with grids at the overview. Only when zooming in are the offices redrawn in full precision. Such an arrangement is superior for more effectively using the screen real estate, because the office nodes are bigger and easier to read.

Conclusions and future development

Developed as a VA system to solve 2012 VAST challenge, SemanticPrism has successfully detected all the anomalies. As Cook et al.[43] pointed out, among all the 2012 challenge submissions, traditional visualizations were well applied, although not many new visualization technologies were invented. Our visualization techniques are traditional and popular, so users can understand them well without training. The integration of visualizations is innovative and effective. The three main techniques covered the weaknesses of each other and had been used as a tool kit to detect problems. For large data with many dimensions, the data may have different characteristics at each dimension, and using a single visualization technique will be hard to fully represent the data. The query-on-demand data handling behind semantic zoom view made it possible to create this lightweight client-side application to solve complicated VA tasks in large-scale datasets. Our system design executes Shneiderman's information-seeking mantra pretty well: the "overview to detail levels" not only exists within each type of visualization, but also links across different types. The time series curves act as an overview. The analyst filters and drills down from either the pixel-based visualization or the geospatial map. However, currently, SemanticPrism lacks the capability to automatically indicate potential anomalies and suggest appropriate zooming areas. Detection of anomalies in the "overview" curves simply relies on the analyst's visual judgment. The analyst also has to identify the suitable zooming level and search around. Early stage problems may be ignored when the signs in the curves are too subtle. Having the system to suggest zoom level and indicate anomalies in concerned areas automatically is another potential research direction for SemanticPrism.

After solving the VAST 2012 challenge 1, we started to consider whether the SemanticPrism, as a VA tool, may grow healthily to solve realistic cyber security problems. The underlying design principle may be generalized to help other large-scale high-dimensional data domains. From our standpoint, we can see that it may develop in two directions. First, SemanticPrism may allow us to find a better, natural way to integrate different visualization approaches (linking and hinting at each other), which may eventually lead to new types of visualization and interaction techniques that are more efficient for high-dimensional data analytics. Second, this tool could enhance the current client–server structure in order to allow it to solve other more complex geotemporal VA problems in real-time large-scale datasets.

Acknowledgements

The authors would like to thank Yinghuan Peng, Abish Malik, Sohaib Ghani, and Steve Visser for their valuable help and thoughtful feedback.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

This study was supported in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

1. VAST. *VAST challenge 2012: challenge descriptions*. Available at: http://www.vacommunity.org/tiki-index.php?page=VAST%20Challenge%202012%3A%20Challenge%20Descriptions.
2. Fua Y-H, Ward MO and Rundensteiner EA. Hierarchical parallel coordinates for exploration of large datasets. In: *Proceedings of the conference on Visualization '99: celebrating ten years*, 24–29 October 1999, pp.43–50. Los Alamitos, CA: IEEE Computer Society Press.
3. Keim DA. Designing pixel-oriented visualization techniques: theory and applications. *IEEE T Vis Comput Gr* 2000; 6(1): 59–78.
4. Oelke D, Janetzko H, Simon S, et al. Visual boosting in pixel-based visualizations. *Computer Graphics Forum* 2011;30(3):871–880.
5. Keim DA, Panse C, Sips M, et al. Pixelmaps: A new visual data mining approach for analyzing large spatial data sets. In: *Proceedings of the Third IEEE International Conference on Data Mining*. 19–22 November 2003, pp. 565–568. Melbourne, FL: IEEE Computer Society Press.
6. Guo D, Chen J, MacEachren AM, et al. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE T Vis Comput Gr* 2006; 12(6): 1461–1474.
7. Keim DA, Panse C and Sips M. Information visualization: Scope, techniques and opportunities for geovisualization. In: *Exploring Geovisualization* (eds J Dykes, A Maceachren and M Kraak), Oxford, UK.: Elsevier; 2005, pp.23 -52.
8. Zhang L, Stoffel A, Behrisch M, et al. Visual analytics for the big data era—a comparative review of state-of-

the-art commercial systems. In: *Proceedings of IEEE symposium on visual analytics science and technology*, 14–19 October 2012, pp. 173–182. Seattle, WA: IEEE Computer Society Press.
9. Bederson BB and Hollan JD. Pad++: A zooming graphical interface for exploring alternate interface physics. In: *Proceedings of the 7th annual ACM symposium on User interface software and technology*, 2–4 November 1994, pp. 17–26. Marina Del Rey, CA: ACM Press.
10. Perlin K and Fox D. Pad: An alternative approach to the computer interface. In: *Proceedings of the 20th annual conference on computer graphics and interactive techniques*, 2-6 August 1993, pp. 57–64. Anaheim, CA: ACM Press.
11. Weaver C. Building highly-coordinated visualizations in improvise. In: *IEEE Symposium on Information Visualization (INFOVIS 2004)*. 10–12 October 2004, pp.159–166. Austin, TX: IEEE Computer Society Press.
12. Conti G, Grizzard J, Ahamad M, et al. Visual exploration of malicious network objects using semantic zoom, interactive encoding and dynamic queries. In: *Visualization for Computer Security, 2005.(VizSEC 05)*, IEEE Workshop on. 26 October 2005, pp. 83–90. Minneapolis, MN: IEEE Computer Society Press.
13. Woodruff A, Olston C, Aiken A, et al. DataSplash: a direct manipulation environment for programming semantic zoom visualizations of tabular data. *J Visual Lang Comput* 2001; 12(5): 551–571.
14. Becker RA, Eick SG and Wilks AR. Visualizing network data. *IEEE T Vis Comput Gr* 1995; 1(1): 16–28.
15. Boschetti A, Salgarelli L, Muelder C, et al. TVi: A visual querying system for network monitoring and anomaly detection. In: *Proceedings of the 8th International Symposium on Visualization for Cyber Security*. 20 July 2011, pp. 1. Pittsburgh, PA: ACM Press.
16. Iliofotou M, Pappu P, Faloutsos M, et al. Network monitoring using traffic dispersion graphs (tdgs). In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 24–26 October 2007, pp. 315–320. San Diego, CA: ACM Press.
17. Eades P. A heuristic for graph drawing. *Congr Numer* 1984; 42: 149–160.
18. Ball R, Fink GA and North C. Home-centric visualization of network traffic for security administration. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*. 25–29 October 2004, pp. 55–64. Washington, DC: ACM Press
19. Yin X, Yurcik W, Treaster M, et al. VisFlowConnect: Netflow visualizations of link relationships for security situational awareness. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, 25–29 October 2004 pp. 26–34. Washington, DC: ACM Press.
20. Muelder C, Ma KL, Bartoletti T. A visualization methodology for characterization of network scans. In: *Visualization for Computer Security, 2005.(VizSEC 05)*. *IEEE Workshop on*, 26 October 2005, pp. 29–38. Minneapolis, MN: IEEE Computer Society Press.
21. Mansmann F, Meier L and Keim D. Graph-based monitoring of host behavior for network security. In: *Visualization for Computer Security, 2007.(VizSEC 07)*. *IEEE Workshop on*. 29 October 2007, pp. 187–202. Sacramento, CA: Springer.
22. Borgo R, Proctor K, Chen M, et al. Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *IEEE T Vis Comput Gr* 2010; 16(6): 963–972.
23. Ziegler H, Nietzschmann T and Keim DA. Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. In: *Information Visualisation, 2008. IV'08. 12th International Conference*. 8–11 July 2008, pp. 287–295. London, UK: IEEE Computer Society Press.
24. Ko S, Maciejewski R, Jang Y, et al. MarketAnalyzer: An interactive visual analytics system for analyzing competitive advantage using point of sale data. *Computer Graphics Forum*. 2012;31(3):1245–1254.
25. Panse C, Sips M, Keim D, et al. Visualization of geospatial point sets via global shape transformation and local pixel placement. *IEEE T Vis Comput Gr* 2006; 12(5): 749–756.
26. MacEachren AM and Kraak MJ. Research challenges in geovisualization. *Cartogr Geogr Inf Sci* 2001; 28(1): 3–12.
27. Malik A, Maciejewski R, Collins TF, et al. Visual analytics law enforcement toolkit. In: *Technologies for Homeland Security (HST)*, 2010 IEEE International Conference on, 8–10 November 2010, pp. 222–228. Waltham, MA: IEEE Xplore.
28. Sopan A, Noh ASI, Karol S, et al. Community health map: A geospatial and multivariate data visualization tool for public health datasets. *Government Information Quarterly* 2012; 29(2):223–234.
29. Podnar H, Gschwender A, Workman R, et al. Geospatial visualization of student population using Google™ Maps. *J Comput Sci Coll* 2006; 21(6): 175–181.
30. Codd EF, Codd SB, Salley CT. *Providing OLAP (On-Line Analytical Processing) to User Analysts: An IT Mandate*. White Paper, EF Codd & Associates, 1993.
31. ADOBE. Adobe - Rich Internet applications. Available at: http://www.adobe.com/resources/business/rich_internet_apps/ (accessed 30 April 2013).
32. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theor Probab Appl* 1969; 14(1): 153–158.
33. Simon HA. *The sciences of the artificial*. 3rd ed. Cambridge, MA: MIT Press, 1996.
34. Hoffman JE and Subramaniam B. The role of visual attention in saccadic eye movements. *Atten Percept Psycho* 1995; 57(6): 787–795.
35. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In: *Proceedings of IEEE symposium on visual languages*, 3–6 September 1996, pp. 336–343. Boulder, CO: IEEE Xplore.
36. Javed W, Elmqvist N. Stack zooming for multi-focus interaction in time-series data visualization. In: *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, 2–5 March 2010, pp. 33–40. Taipei, Taiwan: IEEE Xplore.
37. Choudury S, Kodagoda N, Nguyen P, et al. M-Sieve: A visualisation tool for supporting network security

analysts. In: *Vast 2012 MC1 Award: "Subject Matter Expert's Award*, 14–19 October 2012, pp. 165–166. Seattle, WA: IEEE Society.
38. Chen YV, Qian ZC and Zhang L. Mobile analymator: Animating data changes on mobile devices. In: *Proceedings of IEEE Conference on Visual Analytics Sciences and Technology*, 23–28 October 2011, pp. 311–312. Providence, RI: IEEE Society.
39. Shneiderman B. *Designing the user interface: strategies for effective human-computer interaction*. Boston, MA: Addison Wesley, 1986.
40. Dudas L, Fekete Z, Gobolos-Szabo J, et al. OWLAP - using OLAP approach in anomaly detection. In: *IEEE Conference on Visual Analytics Science and Technology*. 14–19 October 2012, pp. 167–168. Seattle, WA: IEEE Society.
41. Kachkaev A, Dillingham I, Beecham R, et al. Monitoring the health of computer networks with visualization. In: *IEEE conference on visual analytics science and technology*, 2012, pp. 169–170. Seattle, WA: IEEE Society.
42. Robert P. Business forensics HQ. VAST challenge MC1 award: "Good Visualization". In: *IEEE Conference on Visual Analytics Science and Technology*. 14–19 October 2012, pp. 161–162. Seattle, WA: IEEE Society.
43. Cook KA, Grinstein G, Whiting M, et al. VAST challenge 2012: Visual analytics for big data. In: *IEEE Conference on Visual Analytics Science and Technology*. 14–19 October 2012, pp. 151–155. Seattle, WA: IEEE Society.

VASA: Interactive Computational Steering of Large Asynchronous Simulation Pipelines for Societal Infrastructure

Sungahn Ko, Jieqiong Zhao, Jing Xia, *Student Member, IEEE*, Shehzad Afzal, Xiaoyu Wang, *Member, IEEE*, Greg Abram, Niklas Elmqvist, *Senior Member, IEEE*, Len Kne, David Van Riper, Kelly Gaither, Shaun Kennedy, William Tolone, William Ribarsky, David S. Ebert, *Fellow, IEEE*

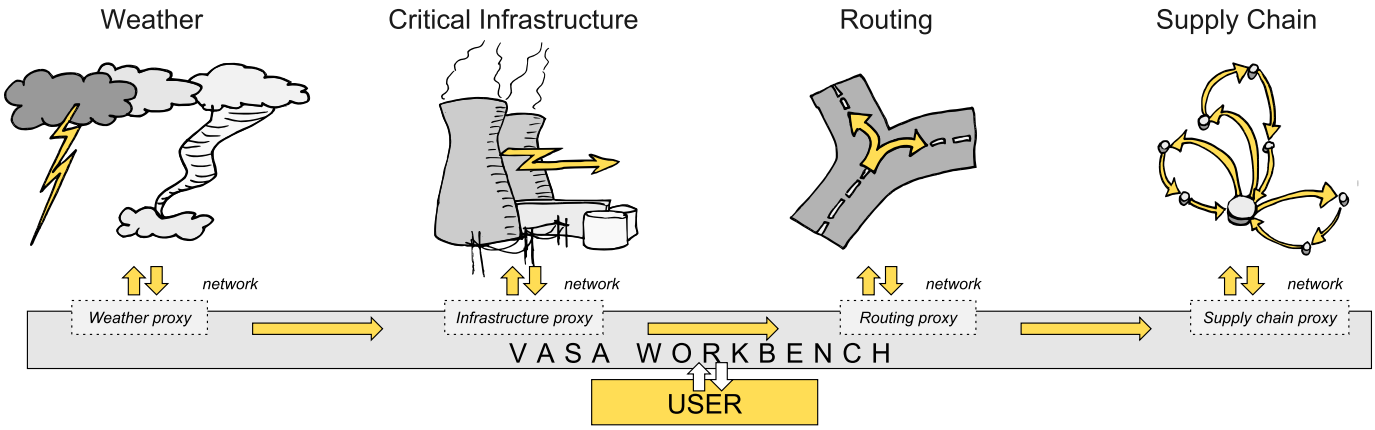


Fig. 1. Conceptual overview of the VASA system, including four simulation components for weather, critical infrastructure, road network routing, and supply chains, as well as the VASA Workbench binding them together.

Abstract—We present VASA, a visual analytics platform consisting of a desktop application, a component model, and a suite of distributed simulation components for modeling the impact of societal threats such as weather, food contamination, and traffic on critical infrastructure such as supply chains, road networks, and power grids. Each component encapsulates a high-fidelity simulation model that together form an asynchronous simulation pipeline: a system of systems of individual simulations with a common data and parameter exchange format. At the heart of VASA is the Workbench, a visual analytics application providing three distinct features: (1) low-fidelity approximations of the distributed simulation components using local simulation proxies to enable analysts to interactively configure a simulation run; (2) computational steering mechanisms to manage the execution of individual simulation components; and (3) spatiotemporal and interactive methods to explore the combined results of a simulation run. We showcase the utility of the platform using examples involving supply chains during a hurricane as well as food contamination in a fast food restaurant chain.

Index Terms—Computational steering, visual analytics, critical infrastructure, homeland security.

1 Introduction

Highways, interstates, and county roads; water mains, power grids, and telecom networks; offices, restaurants, and grocery stores; sewage, landfills, and garbage disposal. All of these are critical components of our societal infrastructure that help run our world. However, the complex and potentially fragile interrelationships connecting these components also mean that this critical infrastructure is vulnerable to both natural and man-made threats: twisters, hurricanes, and flash floods; traffic, road blocks, and pile-up collisions; disease, food poisoning, and major pandemics; crime, riots, and terrorist attacks. How can a modern society protect its critical infrastructure against such a diverse range of threats? How can we design for resilience and preparedness when perturbation in one seemingly minor aspect of our infrastructure may have vast and far-reaching impacts across society as a whole?

Simulation, where a real-world process is modeled and studied over time, has long been a standard tool for analysts and policymakers to answer these very questions (e.g., applications for modeling the real world [10]). Using complex simulations of critical infrastructure components, expert users have been able to create "what-if" scenarios, calculate the impact of a threat depending on its severity, and—last but not least—study optimal mitigation measures to address them. In fact, analysts have gone so far as to name "simulation as the new innovation" [35]: instead of endeavoring to produce the perfect solution once and for all, this new school of thought is to create a whole range of possible solutions and determine the optimal one using modeling and simulation. For example, during the Obama reelection campaign, it was reported that Organizing for Action data analysts ran a total of 62,000 simulations to determine voter behavior based on data from social media, political advertisements, and polling [43]. Basically, the philosophy with big data analytics driven by simulation is not to get the answer perfectly right, but to be less wrong over time [34]. Put differently, while it would be inappropriate to state—as others have done [2]—that big data will never somehow overtake theory, it is clear

- *Sungahn Ko, Jieqiong Zhao, Shehzad Afzal, Niklas Elmqvist, and David S. Ebert are with Purdue University in West Lafayette, IN, USA. E-mail: {ko, zhao413, safzal, elm, ebertd}@purdue.edu.*
- *Jing Xia is with Zhejiang University in Hangzhou, China. E-mail: xiajing@zjucadcg.cn.*
- *Xiaoyu Wang, William Tolone, and William Ribarsky are with University of North Carolina at Charlotte in Charlotte, NC, USA. E-mail: {xiaoyu.wang, ribarsky}@uncc.edu.*
- *David Van Riper, Len Kne and Shaun Kennedy are with University of Minnesota in Minneapolis, MN, USA. E-mail: {vanriper, lenkne, kenne108}@umn.edu.*
- *Greg Abram and Kelly Gaither are with University of Texas at Austin in Austin, TX, USA. E-mail: {gda, kelly}@tacc.utexas.edu*

Submitted to IEEE VAST 2014. Do not redistribute.

that large-scale simulation is a new and powerful tool in our arsenal for making sense of the world we live in.

Applying simulation to the scope of entire critical infrastructures—such as transportation, supply chains, and power grids—as well as the factors impacting them—such as weather, traffic, and man-made threats—requires constructing large *asynchronous simulation pipelines*, where the output of one or more simulation models becomes the input for one or more other simulations arranged in a sequence with feedback. Such a *system-of-systems* [12, 30] (SoS) will enable leveraging existing high-fidelity simulation models without having to create new ones from scratch. However, this approach is still plagued by several major challenges that all arise from the complexity of chaining together multiple simulations in this way: (C1) *monolithic simulations* that are designed to be used in isolation, (C2) *complex configurations* for each model, (C3) *non-standard data exchange* for passing data between them, and (C4) *long execution times* for each individual simulation that are not amenable to interactive visual analytics.

To address these challenges, we present **VASA** (Visual Analytics for Simulation-based Action), a visual analytics platform for interactive decision making and computational steering of these types of large-scale simulation pipelines based on a visual analytics approach. The VASA Workbench application itself is an interactive desktop application that binds together a configurable pipeline of distributed simulation components. It eables the analyst to visually integrate, explore, and compare the inter-related and cascading effects of systems of systems components and potential final alternative outcomes. This is acheive by visualizing both intermediate and final results from the simulation components using a main spatiotemporal view as well as multiple secondary views. The tool provides an interface for the analyst to navigate in time, including stepping backwards and forwards, playing back an event sequence, jumping to a particular point in time, adding events and threats to the timeline, and initiating mitigation measures. Moreover, it allows them to select between or combine different ensemble outputs from one simulation to be fed to other SoS components and explore consequences. Using this interface, an analyst could for example add a weather event (e.g., either an existing hurricane from a historical database, the union of several ensemble output paths, or simulation of a new one) to a particular time, and then step forward a week to see its impact on roads, the power grid, food distribution, and total economic impact in southern United States.

The simulation components provide the main functionality to the VASA platform. Each simulation component communicates with the Workbench using a representational state transfer (REST) API that standardizes the data and parameter exchange. The data flows and parameters passed in the pipeline can be configured using the Workbench application using a graphical interface. Furthermore, the Workbench also includes a local *simulation proxy* for each remote simulation component that provides real-time approximations of each simulation model to enable using them for interactive visual discourse. This feature also provides the computational steering functionality of the Workbench: after configuring a simulation run in an interactive fashion, the analyst can launch the (possibly lengthy) execution from the Workbench. The Workbench then provides tools to manage the simulation pipeline, for example to prematurely shut down a simulation component to accept a partial result, or to skip a particular run.

Our work on the VASA project has been driven by stakeholders interested in supply chain management of food systems, with an initial working example of a food production to restaurant system. For this reason, other than the VASA Workbench application and the protocols and interfaces making up the platform, we have also created VASA components for simulating weather (including storms, hurricanes, and flooding), the power grid, supply chains, transportation, and food poisoning. We describe these individual components and then present an example of how the VASA platform can be used to explore a what-if scenario involving a major hurricane sweeping North Carolina and knocking out a large portion of the road networks and power grid. We also illustrate how the tool can be used to simulate food contamination outbreaks and how this information can be used to track back the contaminated products to the original distribution centers.

2 Background

Visual analytics [38], can be a powerful mechanism to harness simulation for understanding the world. Below we review the literature in visual analytics for simulation and computational steering, as well as appropriate visual representations for such spatiotemporal data.

2.1 Simulation Models

The potential for applying visual analytics to simulation involves not only efficiently presenting the results of a simulation to the analyst, but also building and validating large-scale and complex simulation models. For example, Matkovic et al. [27, 28] show that visual analytics can reduce the number of simulation runs by enabling users to concentrate on interesting aspects of the data. Maciejewski et al. [23] apply visual analytics techniques to support exploration of spatiotemporal models with kernel density estimation and cumulative summation. This work was extended to a visual environment for epidemic modeling and decision impact evaluation [1]. Similarly, Andrienko et al. [5] propose a comprehensive visual analytics environment that includes interactive visual interfaces for modeling libraries and supports selection, adjustment, and evaluation of such modeling methods. Our work is different from this prior art in that our approach combines multiple components in a simulation pipeline, where each stage in the pipeline produces visualization for analysis.

Supply chain management is also a multi-decisional context where what-if analyses are often conducted to capture provenance and processes of supplies. Simulation is recognized as a great benefit to improve supply chain management, providing analysis and evaluation of operational decisions in the supply process in advance [37]. With the IBM Supply Chain Simulator (SCS) [9] and enterprise resource planning (ERP), IBM is able to visualize and optimize nodes as well as relations in the supply chain [20]. Perez also developed a supply chain model snapshot [31] with Tableau. However, existing visualizations of supply chain are mostly limited to either local supply nodes or a metric model rather than managing the overall supply process.

2.2 Computational Steering

Computational steering refers to provoding user control over running computations, such as simulations. Mulder et al. [29] classify uses of computational steering as model exploration, algorithm experimentation, and performance optimization. Applications include computational fluid dynamics (CFD) [13], program and resource steering systems [40], and high performance computing (HPC) platforms [7].

For all of the above applications, the user interface is a crucial component that interprets user manipulation for reconfiguration of data, algorithms, and parameters. Controlling, configuring, and visualizing such computational steering mechanisms is an active research area. Waser et al. proposed World Lines [41], Nodes on Ropes [42], and Visdom [33] as well as an integrated steering environment [33] to help users to manage *ensemble simulations*—multiple runs of the same or related simulation models with slightly perturbed inputs—of complex scenarios such as flood simulations. In the business domain, Broeksema et al. [8] propose the Decision Exploration Lab (DEL) to help users explore decisions generated from combined textual and visual analysis of decision models rooted in articifical intelligence.

2.3 Spatiotemporal Data

Spatiotemporal visual analytics systems enable users to investigate data features over time using a visual display based on geographic maps [3]. In these systems, color, position, and glyphs display features of different regions by directly overlaying the data on the map.

Many approaches to visual analytics for spatiotemporal data exist. Inspired partly by a survey by Anselin [6], we review the most relevant ones below. Andrienko and Andrienko [4] use value flow maps to visualize variations in spatiotemporal datasets by drawing silhouette graphs on the map to represent the temporal aspect of a data variable. Hadlak et al. [16] visualize attributed hierarchical structures that change over time in a geospatial context. Fuchs and Schumann [15] integrate ThemeRiver [17] and TimeWheel [39] into a map to visualize spatiotemporal data. Ho et al. [18] present a geovisual analytics

framework for large spatiotemporal and multivariate statistical flow data analysis using bidirectional flow arrows coordinated and linked with a choropleth map, histogram, or parallel coordinates plot. Our approach is different from those in that our system provide a visual analytics environment for managing and analyzing the results from multiple types of simulations.

Some approaches enable analysis of spatially-distributed incident data. Maciejewski et al. propose a system for visualizing syndromic hotspots [24, 22] while Malik et al. [25] develop a visualization toolkit utilizing KDE (Kernel Density Estimation) to help police better analyze the geo-coded crime data. The latter system is extended to a visualization system [26]where historic response operations and assessment of potential risks in the maritime environment can be analyzed. In our work we also employ KDE for visualizing spatial distrubution of ill people who consumed contaminated food in a supply chain.

3 Design Space: Steering System-of-System Simulations for Modeling Societal Infrastructure

Computational steering is defined as user intervention in an autonomous process to change its outcome. This approach is commonly utilized in visual analytics [38] to introduce a human analyst into the computation loop for the purpose of creating synergies between the analyst and computational methods. In our work, the autonomous processes we are studying are simulation models (often based on discrete event models) that are chained together into asynchronous simulation pipelines where the output of one or several simulations becomes the input to one or several other simulations. Such a simulation pipeline is also a *system-of-systems* [12, 30] (SoS): multiple heterogeneous systems that are combined into a unified, more complex system whose sum is greater than its constituent parts. Synthesizing all these components yields the concept of visual analytics for *steering system-of-system simulations*: the use of visual interfaces to guide composite simulation pipelines for supporting sensemaking and decisionmaking. In this work, we apply this idea to modeling societal infrastructure, such as transportation, power, computer networks, and supply chains.

In this section, we explore the design space of this concept, including problem domains, users, tasks, and challenges. We then derive preliminary guidelines for designing methods supporting the concept.

3.1 Domain Analysis

A wide array of problem domains may be interested in creating large-scale system-of-system simulation pipelines for studying impacts on societal infrastructure. Our particular domain is for business intelligence for supply chain logistics in the fast-food business, but we see multiple potential applications (each with a specific example):

- **Supply chain logistics:** Impact of large-scale weather events on the distribution of goods (particularly perishables, e.g., food).
- **Public safety:** Crime, riots, and terrorist attacks on critical infrastructure, such as on roads, bridges, or the power grid.
- **Food safety:** Incidence, spread, and causes of food contamination, often due to weather (power outage) or transport delays.
- **Cybersecurity:** Societal impact of cybersecurity attacks, such as on power stations, phone switches, and data centers.

3.2 User Analysis

The intended audience for computational steering of simulation models using visual analytics are what we call "casual experts": users with deep expertise in a particular application domain, such as transportation, supply chain, or homeland security, but with limited knowledge of simulation, data analysis, and statistics. Their specific background depends on the problem domain; for example, they may be business or logistics analysts for supply chain applications, police officers for public safety, and homeland security officials for food safety and cybersecurity. Because of this "casual" approach—a term we borrow from Pousman et al.'s work on casual information visualization [32]—our intended users are motivated by solving concrete problems in their application domain, but are not necessarily interested in configuring complex simulation models and navigating massive simulation results.

Even if our primary user audience is these casual experts, it is very likely that the outcome of a simulation steering analysis will be disseminated to managers, stakeholders, or even the general public [38]. Thus, a secondary user group for consuming our analysis products is laypersons with an even more limited knowledge in mathematics, statistics, and data graphics.

3.3 Task Analysis

Based on our review of the literature (Section 2) as well as feedback from domain experts, we identify a preliminary list of high-level tasks for steering system-of-system simulations for societal infrastructure:

- Increasing *preparedness* for potential scenarios;
- Improving the *resilience* of an organization; and
- Planning for *mitigation and response* to a situation.

3.4 Challenges

Modeling the real world is a tremendously difficult and error-prone process. However, we leave concerns about the fidelity, accuracy, and quality of a simulation to research within the simulation design. Rather, in this subsection we concern ourselves with the challenges intrinsic to connecting multiple individual simulation models into large-scale pipelines. In the context of simulation steering for such pipelines, we identify the following main challenges:

C1 **Monolithic simulations:** While individual high-fidelity simulation models exist for all of the above components and threats, these models are monolithic and not designed to work together.

C2 **Complex relationships:** Each high-fidelity simulation model consists of a plethora of parameters and controls that require expertise and training, which is exacerbated when several such models are combined into a single model.

C3 **Non-standard data:** No standardized data exchange formats exist for passing the output of one simulation model, such as for weather, as input to another model, such as supply chain routing.

C4 **Long execution times:** Most state-of-the-art, high-fidelity simulation models require a non-trivial execution time, often on the order of minutes, if not hours. Such time frames are not amenable for real-time updates and interactive exploration.

C5 **Uncertainty and fidelity:** Chaining together multiple simulations into a pipeline may yield systematically increasing errors as uncertain output from one model is used as input to another. This is compounded by the fact that heterogeneous simulation models may have different levels of fidelity and accuracy.

3.5 Design Guidelines

Based on our review of the problem domain, users, and tasks above, as well as the challenges that these generate, we formulate the following tentative guidelines for designing visual analytics methods for steering system-of-system simulation pipelines:

G1 *Simulations as standardized network services:* Distributing simulation models as network services avoids the trouble of integrating a monolithic design with another system (C1) and automatically provides a data exchange format (C3). The simulations also become decoupled, which means they can be parallelized and/or distributed in the cloud to manage long execution times (C4).

G2 *Simulation proxies for interactive response:* Meaningful sensemaking in pursuit of one of the high-level tasks in Section 3.3 requires real-time response to all interactive queries. This means that long execution times (C4) of simulation models in the pipeline should be hidden from the user. We propose the concept of a *simulation proxy* as an approximation of a remote simulation service that is local and capable of providing real-time response at the cost of reduced (often significantly) accuracy.

G3 *Visual and configurable relationships:* The interactive visual interfaces routinely employed in visual analytics may help to simplify and expose the complex configurations necessary for many high-fidelity simulation models (C2), even for non-expert users.

G4 *Partial and interruptible computational steering:* Once an analyst has configured a simulation run using simulation proxies (G2) and visual mappings (G3), the full simulation pipeline must be invoked to calculate an accurate result. A full-fledged simulation run may take minutes, sometimes hours, to complete. The computational steering mechanisms provided by the software should provide methods for continually returning partial results [14] as well as interrupting a run halfway through.

G5 *Visual representations of both intermediate and final results:* To fully leverage the power of visual analytics, we suggest using interactive visual representations of simulation results. Such visualizations should be used for both intermediate data generated by a simulation component anywhere in the pipeline—which would support partial results and interrupting a run at any time—as well as for the final results. All visual representations should be designed with uncertainty in mind (C6), and providing intermediate visualizations should also help in exposing propagation of increasing error. Finally, it may also be useful to use visual representations for the approximations created by simulation proxies (G2), but these should be clearly indicated as such.

4 VASA: Overview

As previously described, our VASA system is a distributed component-based framework for steering system-of-system simulations for societal infrastructure. Figure 1 gives a conceptual model of the system architecture. At the center of the system is the VASA Workbench (Figure 2), a user-driven desktop tool for configuring, steering, and exploring simulation models, impacts, and courses of action. The workbench provides a visual analytics dashboard based on multiple coordinated views, an event configuration view, and a computational steering view. The workflow of the workbench revolves around initiating, controlling, analyzing, exploring, and handling events from the remote simulation components as well as the local simulation proxies.

Within the dashboard, events are displayed in a selectable calendar view (a) where each event's name, dates and a user-selected representative attribute (e.g., storm's maximum wind speed) are shown. The selected events from (a) are listed based chronologically in the event viewer (b) where a user can select times for investigation. In (b-1), various options are provided, including initiating simulations (e.g., cyberattack, storm simulations, distribution re-routing), selecting combinations of events (union, intersection, difference), selecting event visualization modes (polygons, contours), and chronological playback.

Users can fix a time within an event for comparison (right-clicking on a event's black rectangle) and a red mark is shown in the upper right corner of the associated rectangle(b-2), and the impact is shown in the main geospatial view (d-1). We provide a legend window (c) for selected properties (e.g., distribution centers, restaurants, power plants and other infrastructures) and the geographical view (d) provides the simulation results including event evolution, routing paths, and impacts on critical infrastructures. A food delivery schedule to each store within a supply chain is provided in (e) where the x-axis presents corresponds to different restaurants while the y-axis represents different food processing centers or different types of foods. Here, the darker the red, the larger the quantity of the delivered food. The quantity information is provided in a tooltip that helps a user to estimate possible losses. This view enables traceback analysis (e.g., which type of food was contaminated from which processing centers, how much contaminated food was delivered to which store) for food contamination incidents.

5 VASA: Components

Our current VASA suite consists of four simulation components that implement the VASA interface: components for weather, critical infrastructure, routing, and supply chains. We review each of these next.

5.1 Weather Component

In order to provide clients with a one-stop source for weather data, we implement a server that asynchronously amasses data from various online sources and presents it to clients through a RESTful web interface. This provides access to various data through a singly authenticated service that provides consistent and convenient APIs for data acquired from many sources.

5.1.1 Simulation Model

For example, a collaboration of several research centers runs the ADCIRC model during hurricane season off the east and gulf coasts of the U.S. When storms are present, these models are run every four hours, producing ADCIRC-formatted datasets at fixed intervals forward from the initial times. These results are made publicly available using THREDDS and OPeNDAP for cataloging, discovery and data access. When this data appears, we import it onto a VASA server, and provide a simple RESTful API to access the data in convenient multiresolution formats. Similarly, NOAA produces wind-speed probabilities along the tracks of storms as contours at 34, 50, and 64-knot levels. This data is also imported asynchronously onto the VASA service and provided through the VASA RESTful API.

5.1.2 Simulation Proxy

The proxy in this component has two roles. The first role is to prepare all event data sets from the remote event server. Therefore, the system first checks for new updates from the server. If there is a new update, it retrieves the data and saves it on the local workbench for faster loading. The second role is to visualize new status of an event on the date that a user selected and notify the status change of the event to other proxies. An example status change is a user changing the start date of a hurricane in the event viewer. When this happens, the proxy visualizes a new status of the hurricane on the date and notifies this change to other components, which initiates each proxy's work (e.g., estimating an area without power and impassable roads).

A user can select the hurricane visualization type either as polygons or contours for estimation by clicking a button as shown in Figure 2 (b-2, the last button). In the polygon mode, two probability models (blue with two different opacities) are projected as shown in the magnification view in Figure 2. Here, the smaller polygon means an expected path with high probability, and a larger one presents an expected path with low probability. When a user fixes a hurricane, the hurricane turns red for comparison to other paths (of other hurricanes). For example, in Figure 2 the path of Hurricane Irene on August 24, 2011 is projected (blue) and the path of Hurricane Sandy in October 27, 2012 is presented in red for comparison.

In the contour mode, hurricanes are drawn using three different sizes of contours, each of which represents mean areas in different wind speeds (e.g., Hurricane Irene in our simulation model has 64 knot highest wind speed at the innermost contour, and 34 knot lowest wind speed at the outermost contour as shown in Figure 6). To utilize different wind speeds in simulation steering, a user can set up a threshold for infrastructures (e.g., a power generation unit is disabled if the wind hitting the plant has speed higher than 34 knot). In addition, a user can apply one of the contours for a time. For example, Figure 6 (top-right) presents which power generation units are affected when a contour with 34 knot hits the area. Here red circles represent affected restaurants and red circles present the impacted power generation units supplying electricity to those restaurants.

5.1.3 Implementation Notes and Performance

From the client's point of view, the VASA API consists of URLs that encode procedures and parameters that, when issued, return JSON objects containing the results. This provides a very simple interface for use both by browser-based visualization UIs that use AJAX to issue requests asynchronously, and other native platforms that provide equivalent access through language-specific interfaces.

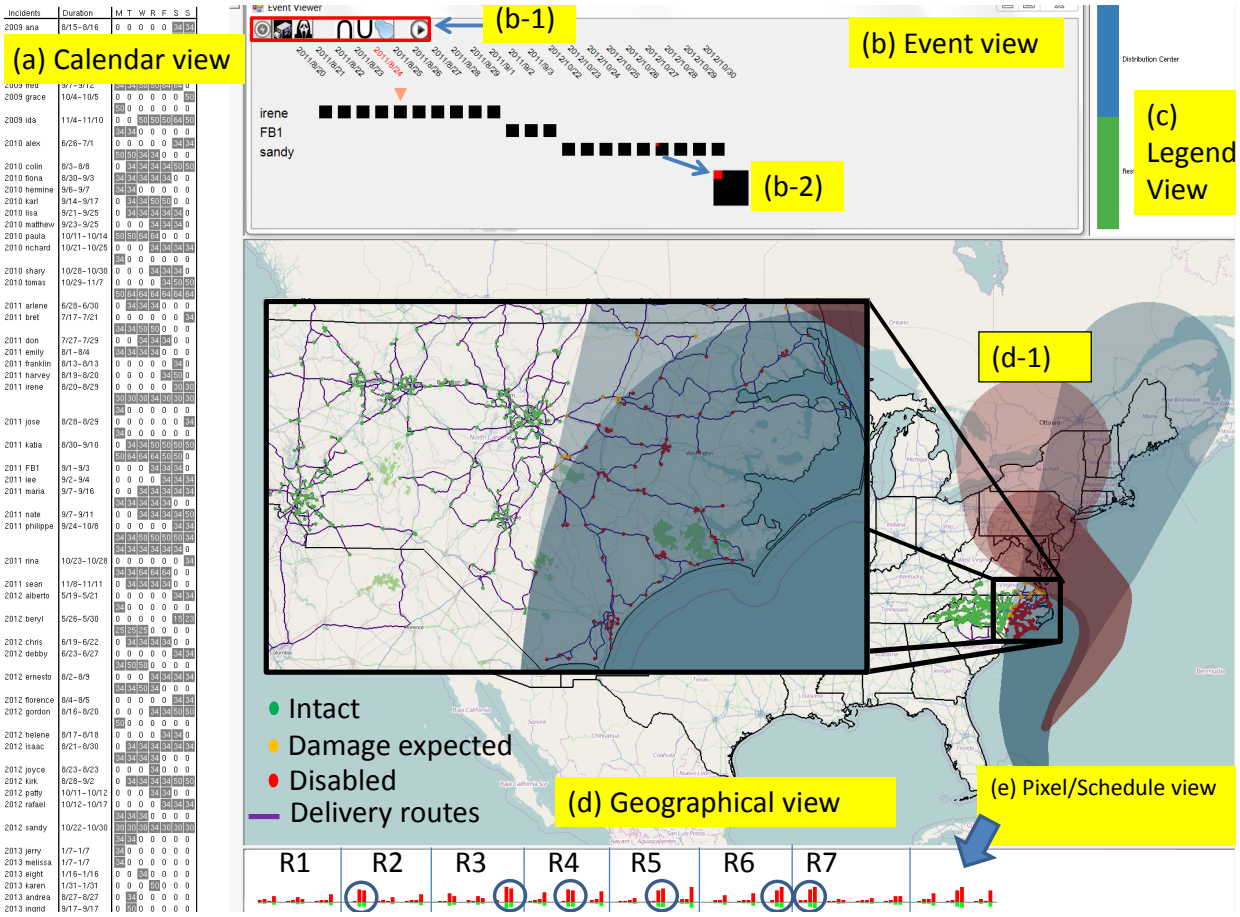


Fig. 2. Multiple coordinated views in the VASA Workbench. (a) Calendar view with available events (e.g., weather, food poisoning, cyberattack, etc). (b) Event timeline for configuring events. (b-1) Event buttons. (b-2) Fixed event. (c) Map legend. (d) Geographical map. (d-1) Hurricane (red). (e) Pixel/Schedule view showing food deliveries. Each area divided by a blue line means a route that visits 3–4 restaruns, 3 time a week. This view also can be used for pixel-based visualization.

5.2 Critical Infrastructure Component

Widespread emergencies such as hurricanes, flooding, or cyberattacks will often affect multiple societal infrastructures. High winds and flooding from a hurricane, for example, could knock out parts of the power grid, the effect of which would cascade to traffic signals, the communications network, the water system, and other infrastructures. The flooding might simultaneously make parts of the road network impassable. These breakdowns would affect critical facilities such as schools, hospitals, and government buildings. For longer-lived disasters, food distribution might break down due to power outage, route disruption, or other cascading effects. The purpose of VASA's critical infrastructure component is to simulate how such external emergencies, modeled in other components, will impact critical infrastructure.

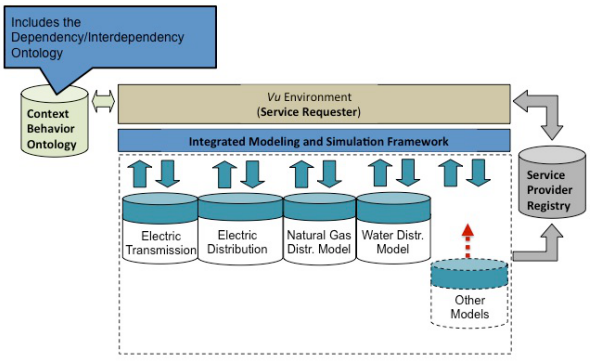


Fig. 3. Vu environment showing the modular structure where different simulation submodels can be inserted.

5.2.1 Simulation Model

To capture these complex, multifarious, and dynamic effects, we developed a simulation model that takes into account the interrelationships between critical infrastructure systems. The simulation is built within the Vu environment (Figure 3), which provides a rule-based framework for integrating multiple infrastructure components at a high level. This results in an interdependency ontology. Thus, for example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a school were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down. Likewise, telecommunication nodes would switch to backup power that might also shut down after its prescribed duration. There could also be outages due to power load imbalances at other points in the grid.

These interlaced critical infrastructures are captured in a set of networks, with each node having a set of properties according to its category and the edges providing a dependency rule according to the category and state of the connected nodes. Relations between networks is captured by edges between nodes in the two networks. The timings of interdependencies and state changes are set according to a universal clock, so that any simulation of cascading effects evolves over time and space (since nodes are geographically located). The rules for networks and interdependencies are set in consultation with experts (in the case of the power grid, for example) or through consultation of the appropriate literature for an infrastructure. However, some of the interdependencies are not directly known, even by experts, since measures or simulations linking some infrastructures have never been done or validated. In this case, we define plausible rules that produce outcomes consistent with experience. This is in fact an advantage of the

Vu approach in that it permits investigating interdependencies at a high level with a quick adjustment of the model and turnaround of results. These interdependencies can then be studied in depth as needed.

The Vu approach will not capture detailed interdependency effects, transients, or complex interactions. However, it has been shown to be quite successful as a high-level simulator for a range of applications. To achieve this, it is modular, allowing new modules to be dropped in (Figure 3) and, as soon as the interdependency ontology is set up, simulations to be run. This makes it easy to configure simulations for several different infrastructures and even involving other types of simulation models, such as population and economic models.

5.2.2 Data

Our prototype system currently uses data from the state of North Carolina, and the data collection and organization process involves locating and identifying components of the various infrastructures for the state. We use publicly available data sources, in some cases identifying infrastructure components by indirect means. For example, comprehensive information about the electrical grid is closely held by the utility companies. However, we have shown our results to utility company officials and received confirmation as to their high level accuracy.

The infrastructures we currently model include the electric grid; the communications network including TV stations, radio stations, cellular switch controls, and cell towers; transportation facilities including airports, bus terminals, rail lines and terminals, bridges, tunnels, and ports; the road network including main and secondary roads; natural gas pipelines and pumping stations; critical facilities including fire stations, police stations, schools, hospitals, emergency care facilities, manufacturing locations, government buildings, and hazmat facilities.

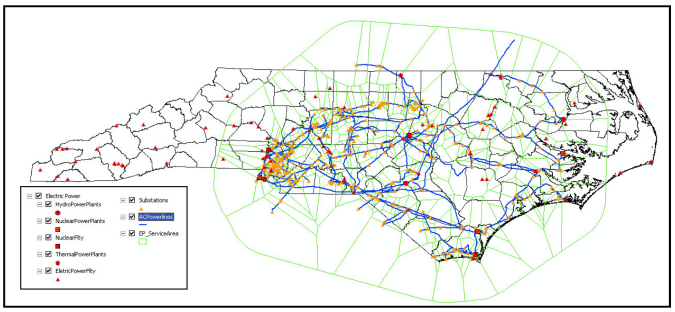


Fig. 4. Power transmission grid with parts of the distribution network.

Figure 4 shows the electric grid, which includes the complete transmission network down to substations for both North and South Carolina. Parts of the distribution network are also included, especially for critical installations. Figure 5 shows the transportation network for North Carolina, including roads, airports, rail lines, etc. For the purposes of VASA, we have also added store and distribution center locations for a large food chain in North and South Carolina. These facilities are linked to the power grid and road networks.

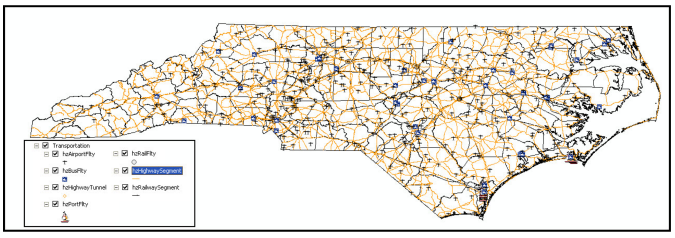


Fig. 5. Transportation network including transportation facilities.

5.2.3 Simulation Proxy

The proxy for the critical infrastructure component maintains a simplified connectivity network of critical infrastructure. In this graph, restaurants are connected to the nearest plant for impact approximation due to an event. When the proxy receives a signal of a new event (e.g., storm path change, new day for approximation), it computes which infrastructures are affected by the event. For example, when a user moves the hurricane simulation forward to a new day, our proxy checks which infrastructures are newly affected and produces an estimate and its corresponding visualization (e.g., condition color changes for restaurants affected by power plant disruptions).

5.2.4 Implementation Notes and Performance

As for all the other VASA simulation components, we use a web service that can accept requests from the VASA Workbench and send either complete or approximate simulation results ready to be presented in the user interface (see Figure 1). The critical infrastructure server itself has two components. One contains a searchable database of the pre-computed ensemble of simulation runs. The other accepts current storm path and other inputs from the weather server, converts them into courses of action, and computes a fresh set of cascading infrastructure disruption results. When a request is issued via the user interface, the Simulation Proxy determines the weather inputs to send to the ensemble database component that immediately selects the closest ensemble simulation for use in the visual analysis. This proxy is then replaced by the more accurate result based on the current weather simulation as soon as it is available. Therefore an emergency response manager can make initial decisions based on the proxy and then refine them, if desired, once the up-to-date result is available.

5.3 Supply Chain Component

Most food systems involve a number of firms from on-farm production of inputs through processing, distribution and retail sales. For the fast food system in VASA, three different firms have collaborated to provide the data on normal system performance: a vertically integrated poultry firm (hatchery to processed chicken), a warehouse and distribution firm, and a fast food restaurant firm. Each firm contributed data from their portion of the supply chain to enable modeling of product movement from farm to restaurant. The type of data provided includes geospatial information on the facilities involved (e.g., feed mills, hatcheries, poultry farms, poultry processing facilities, distribution centers and restaurants), normal transportation routes and scheduled times from each facility to the next facility in the system and details on actual shipment quantities on average (hatchery through processing) or actual shipment records for a limited time frame (distribution centers to restaurants). As an illustration of the amount of data that drives these systems, one week of data on product delivered from the two distribution centers to the nearly five hundred individual restaurants alone is over 120,000 individual records.

Hurricanes pose significant risks for normal supply chain operation from impassable roads, power outages and floodings disrupting facility operation and distribution of products throughout the system. Understanding which routes and locations are likely to be at risk from a storm would enable a firm to develop contingency plans in advance of a storm, thus reducing operational losses immediately after a storm. Given that daily sales at larger fast food restaurants can be $4,300-$7,400, losses can mount up quickly. If in the case of an impending storm or immediate aftermath, near real time rerouting could enable firms to most efficiently maintain their distribution systems for both maintaining product distribution and retrieving of food from restaurants without power to minimize spoilage losses.

5.3.1 Simulation Model

Food contamination can occur both intentionally or as a malicious act at any point in the supply chain and can result in significant public health consequences, from morbidity to mortality. While firms are required to have information one step forward and one step back in their supply chain, they often have difficulty gaining visibility beyond that. By gathering data from each step in the supply chain, it is possible to trace product from farm through to restaurant and from restaurant back to farm. Using data on actual lot sizes from the firms involved, two illustrative contamination scenarios were constructed to illustrate how differently seemingly similar contamination scenarios would transpire. This system also illustrated a common problem of “hidden nodes” in

the system, i.e., facilities that one firm in the system does not realize are part of its supply chain. One of the poultry slaughter and processing facilities ships raw poultry to a further processing facility that then ships the resulting product to the distribution centers. If there were a contamination at the "blind" facility, neither the distribution firm for the restaurant firm would initially know that it was part of their supply chain. A contamination scenario builder is now under development that would enable users to model a wide range of contamination events and see how they would propagate through the supply chain.

Our simulation model can generate food-borne illness data based on an approach similar to the Sydovat [21] system. There are two major components of the model for generating synthetic illness data: temporal and spatial data. A time series is contructed from its individual components (day-of-week, interannual, interseasonal, and remainder) similar to seasonal trend decomposition. To generate the time series of food-borne illnesses for a user-injected restaurant location, the user defines the mean daily count of illnesses along with seasonal and day of week components. If the historical food-borne illnesses data is available then seasonal and day of week components can be randomly selected from this historical data. Spatial locations for temporal data are generated based on the density distribition that approximate the population in that area. Additionally, users can customize the grid size and density distributions.

5.3.2 Simulation Proxy

Our simulation proxy for the supply chain component maintains a low-fidelity representation of the transport network. This is used together with the weather polygons to approximate when a distribution center and store must shut down. For food-poisoning data, this inherently contains spatially-distributed points of ill people simulated based on the simulation model (Section 5.3.1). To visualize the spatial distribution and the hotspots of the poisoned people, the proxy in this component uses a modified variable kernel density estimation technique with varying scales of the parameter of estimation based upon the distance from a patient location to the k_{th} nearest neighbor [36]. The model used for estimating the number of people poisoned is the same model utilized in Maciejewski et al. [1, 22], but we adjust parameters to consider different population densities in different regions.

5.3.3 Implementation Notes and Performance

The supply chain component is built in ArcGIS and Arc Network Modeler so that storm impacts can model solutions accounting for restaurants out of service (power, flooding) and impassable roadways.

5.4 Routing Component

The purpose of the routing component is to provide a mechanism for other VASA components to find appropriate routes from one facility to another given a dynamically changing world model, where roads may become impassable due to weather or other widespread emergencies.

5.4.1 Data Model

We obtained the addresses of two distribution centers and 505 fast-food outlets, as well as the route information that links the centers to the outlets. We geocoded the addresses using the Environmental Systems Research Institute (Esri) ArcGIS 10.2 Server with the Network Analyst extension, and StreetMap Premium for ArcGIS (TomTom North America data) Geodatabase. We then calculate N shortest path routes, where N is the number of routes specified in the input data, using Esri's Network Analyst Route tool and the StreetMap Premium road network. The road network has a long list of attributes used to determine the shortest route, including road class, speed limit, number of lanes, and weight restrictions.

5.4.2 Simulation Model

The input to the routing component is a GeoJSON polygon representing an area impacted by severe weather (such as a hurricane). The component ingests the GeoJSON object as a polygon barrier in the road network. Attributes of the road network are weighted to create a friction surface which iterates through routing options to determine the optimal route. The model does not currently include current traffic conditions or construction activity, but these factors could be added in the future. Each route minimizes the travel time between the distribution center and the first store or between stores. This set of routes represented the baseline scenario—how delivery trucks would travel under normal circumstances. Since delivery trucks can no longer reach outlets covered by the weather barrier, the routing service recomputes the routes with the barrier in place and returns new routes which avoid the outlets and roads covered by the barrier. If the barrier covers a distribution center, no deliveries will be made to outlets serviced by the center. The routes are output as a set of large GeoJSON objects and sent back to the caller.

5.4.3 Simulation Proxy

The main focus of the proxy in the routing component is on approximating the number of routes that will be replaced if a complete simulation result exists. The proxy investigates which nodes in routes are expected to be disabled when there is an event. Then, after the investigation, it builds a polygon by connecting outer-most nodes and visualizes the polygon. This gives awareness to a user that the routes in the polygon are likely to be changed after a complete simulation. A user can initiate the simulation by clicking the "run" button (Figure 9).

5.4.4 Implementation Notes and Performance

The goal was to use as much Commercial Off-The-Shelf (COTS) software as possible when implementing the routing model. The Esri suite of Geographical Information System (GIS) tools is widely used in a variety of industries and provides a robust set of tools and data. Specifically, we used ArcGIS Server 10.2 with the Network Analyst extension. The server provides web-based services through REST endpoints and provides a robust API accessed with HTTPS GET or POST requests. The VASA workbench initiates a request to the routing service by providing a GeoJSON representation of the affected area. The affected area polygon is input to Network Analyst Service to recalculate the route to traverse around the affected area. The response is two large GeoJSON objects containing a list of outlets no longer reachable, incremental travel time between stops, and the new route. Currently, the route processing requires 2-3 minutes to complete; this can be significantly improved when a production server is commissioned.

6 Examples

We showcase the utility of the VASA Workbench and our current simulation components using three examples: the impact of weather on macro-scale supply chains, foodborne illness contamination and spread, and a simplified cyber-attack on the power grid infrastructure.

6.1 Supply Chains in Hurricane Season

Our first example is the potential impact of hurricanes on North Carolina's critical infrastructure, especially our food distribution network, in North Carolina (NC). Our exploration begins by selecting appropriate historial hurricanes for examination using the calendar view as shown in Figure 2, where each hurricane name, duration, and selected summary attribute (e.g., maximum hurricane wind speed) are provided. While we investigate the paths of these historical hurricanes, we see that Irene in 2011 and Sandy in 2012 passed over NC. Because Sandy passed over only a small area in upper NC (Fig 2 (d), red hurricane polygon), we choose to focus on Irene for further investigation.

One interesting date is August 27, 2011 when Irene passed directly over eastern NC, an area with many power generation facilities, as shown in Figure 6 (top-right, purple circles). After we set up the wind tolerance value for these facilities to be 34 knot, our hurricane proxy instantly estimates which restaurants will be impacted based on the relationships between the units and the restaurants and colors the impacted restaurants red. Here, we also initiated a complete simulation for power outages and transportation network damage. Next, a polygon is shown representing an area where restaurants are disabled and which roads are blocked (bottom-left in Figure 6). To efficiently manage distribution, this impact requires the food provider to change its

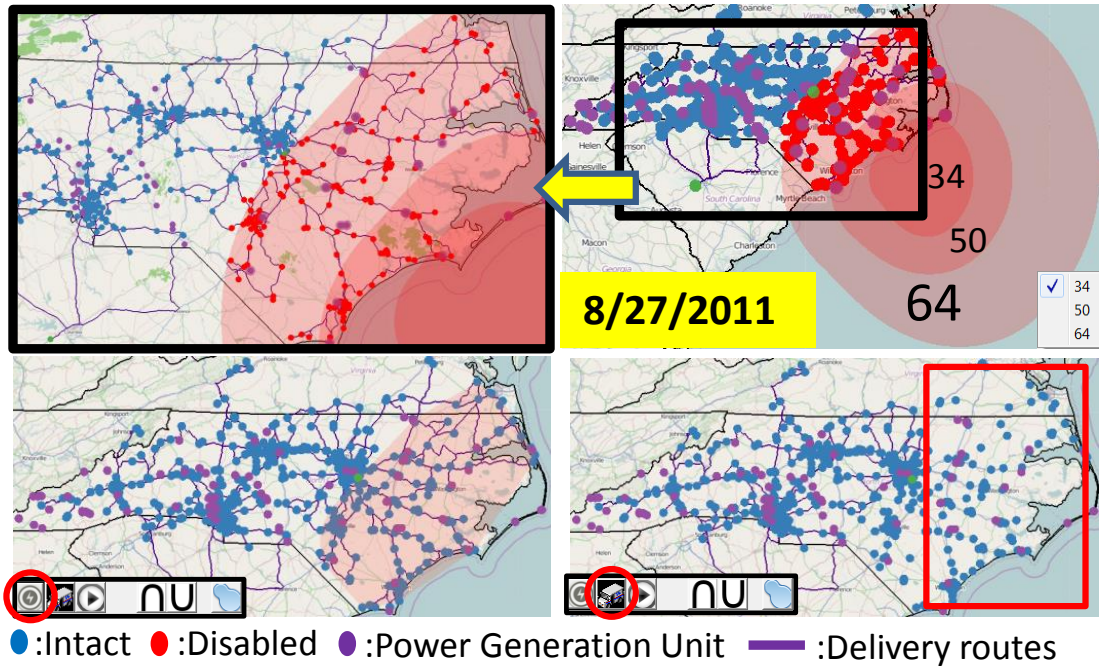


Fig. 6. In this simulation, power generation units were hit by up to 34 knot during Hurricane Irene on August 27, 2011. Our hurricane proxy instantly estimates the impacted restaurants (right-top, left-top). Note that one distribution center (green) is outside the hurricane. After a complete power-grid simulation run is finished (by clicking the circled lightning button), a polygon representing the power outage area is shown. Next, this polygon is sent for use in computing new food delivery paths. Note that food is not delivered to the power outage area (right-bottom, red box).

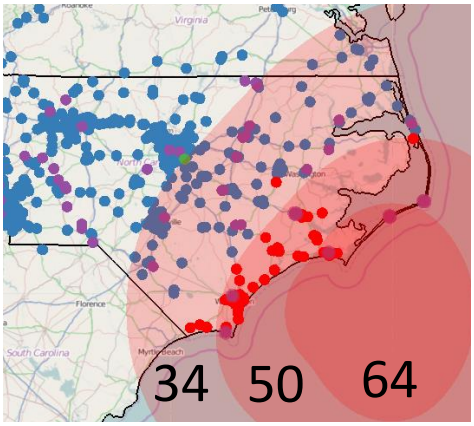


Fig. 7. If the power generation units could have resisted up to 50 knot wind, the number of impacted restaurants will be much smaller.

delivery schedule, and this new routing is computed based on the impacted restaurant polygon and road conditions (e.g., blocked by flooding). After a simulation to compute the new routes (by clicking the truck button in a red circle, right-bottom Figure 6), we see that the updated delivery paths do not include the affected restaurants. The economic loss caused by this event is estimated based on the model in Section 5.3 as being up to $1.13 million. Another possible what-if question is "How different would the result be if the power generation units can resist winds up to 50 knot?" Figure 7 shows the first step of the analysis where we see many fewer restaurants affected compared to Figure 6 top-right (units are resilient to 34 knots). In this case, the estimated losses are less than $333,000.

6.2 Fast Food Contamination

Food poisoning is an illness caused by eating contaminated food containing viruses, bacteria and germ-generated toxins. There are many possible causes of food contamination including storage at inappropriate temperatures [19], improper food handling, and cross-contamination during processing or packaging. As unfortunately experienced several times per year, tracing back the cause of the contamination is a very difficult lengthy process. In this example, we explore a hypothetical scenario demonstrating how VASA can be used to trace-back the root causes of an incident of foodborne illness.

To create the distribution of the ill population, we simulate the distribution of contaminated food to stores, then simulate the illnesses in the neighboring areas using the simulation model discussed in Section 5.3. This creates the common base scenario of reports of people who are ill, their date of illness and their location to create the food contamination scenario for the trace-back investigation.

For example purposes, we simulated these illnesses occurring during a three day span (September 1, 2011 to September 3, 2011) as shown in Figure 8. Since this is almost one week after Hurricane Irene, one may assume that power outages during the storm could be the possible reason behind the contamination. To confirm this hypothesis, we looked at the hot spots in Figure 8 and identified the stores closest to these hot spots. On cross comparison, we can identify the common products/lots in those stores, their distribution center, as well as their delivery mechanisms. As shown in Figure 8 bottom matrices, the rows represent 3 food processing centers and 4 types of food, and there is a column for each restaurant. Each cell is colored such that the darker the red color, the higher the amount of each product provided. Here, the restaurants in the affected area that are selected in the box in the top-left are highlighted with light green boxes. For stores S9 and S12, only one food processing center provided products, while other processing centers supplied most of the food throughout the network. Upon further inspection, one can determining that 3rd and 4th row product lots are common in most of the restaurants where individuals are. Some example routes are shown in Fig. 2 (e) where each route supplies 3-4 restaurants. A red bar means the supplied food and the green bar means the food consumed at a restaurant. Here, we see that a large amount of the third and fourth foods (blue circles in Fig. 2 (e)) are delivered and will all be consumed within a few days. Therefore, these two product lots are good candidates for further inspection in tracing back the contaminated food item.

6.3 Cyberattack on Critical Infrastructure

Part of the mission of the VASA project is to study the impact and mitigation of man-made attacks on societal infrastructure. Cybersecurity is becoming an increasingly important threat to modern society [11]

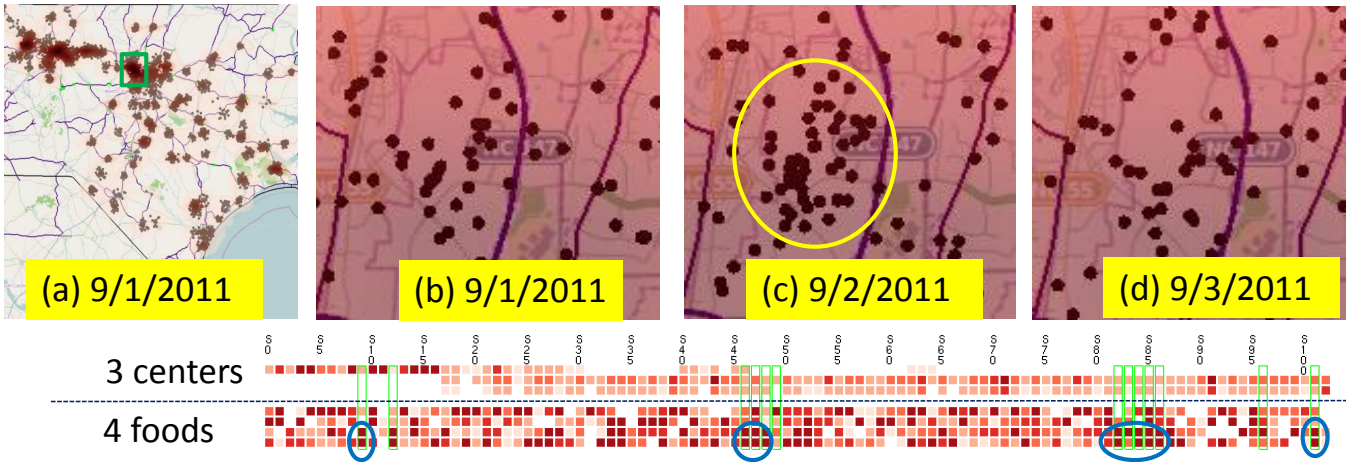


Fig. 8. Ill people caused by contaminated food is presented using a KDE hotspot visualization. In (a), the darker location has a larger number of poisoned people. Brown points mean ill people in the reported location. The locations highlighted by a green box in (a) is magnified in (b), (c) and (d) on different dates. As the timeline shows, the number of ill people increased until 9/2/2011, then started decreasing on 9/3/2011. The bottom matrices show which food processing centers (1–3) were involved and which foods (1–4) were delivered to which store in 8/30/2011, two days before the illness. Here, the restaurants in the light green boxes are the those selected by the thicker green box in (a). We see that a large quantity (darkest red pixels in blue circles) of two foods (third and fourth rows) are commonly provided to restaurants in the area.

Fig. 9. An example of the cyberattack simulation. (left) A user selects to disable power transmission by a cyberattack in the option menu and selects the region shown in the red rectangle. One power plant (purple dot) is included within this rectangle (shown in the dark red circle). (second-left) The infrastructure proxy instantly estimates the affected restaurants (red dots), and a full simulation is initiated. (second-right) Power outage regions are presented by the polygon, and new distribution routes are computed. (right) The new routes are shown as paths and do not include the affected restaurants but, unlike the hurricane scenario, all roads are available for food distribution. For comparison, see the path radiating from the polygon that was not allowed in the hurricane scenario.

and may have a significant effect on an increasingly connected society where power plants and substations are all controlled from afar.

While we do not yet provide a cyberattack module for VASA, many of the simulation components provide direct access to changing the state of particular infrastructure components through VASA. This enables us to simulate a cyberattack by, for example, shutting down a particular or several specific critical infrastructure component even if it is not affected by weather or other natural threats. Figure 9 shows a screenshot of an analyst studying precisely such a scenario where a power plant has been disabled by a cyberterrorist. Here the analyst simulates that the terrorist shuts down the power transmission grid for one of the two power plants in the town by drawing a red rectangle (left) around it. Then, the infrastructure proxy instantly estimates possible affected restaurants and a full simulation is initiated for more accuracy (second-left). The simulation result is presented by a polygon and new route computations can be initiated (second-right). The new routes with impacted restaurants are visualized. Note that this routing example is different from the hurricane case because roads are still passable: purple paths are still shown within the polygon (right).

7 Conclusion and Future Work

We have introduced the notion of visual analytics for simulation steering within the context of societal infrastructure. To our knowledge, ours is the first to study visual analytics for simulation from a *systems-of-systems* [12] perspective, where multiple heterogeneous—often physically distributed—systems are combined into a unified, more complex system in which the linkages between components provide a sum greater than its constituent parts. This notion transcends individual simulation models and instead chains together multiple high-fidelity simulations into large-scale asynchronous pipelines. The VASA system we presented as a practical example of such an approach is a distributed application framework consisting of a central Workbench controlled by an analyst and a set of loosely coupled simulation components implemented as distributed network services.

Big data simulation is a powerful new tool for data science, and while our work on applying visual analytics to this domain is conceptually complete, it really only scratches the surface of what is possible. Future work on the VASA system will involve integrating even more advanced and detailed simulation components, such as high-fidelity power grid models, gas pipelines, and power plants for energy infrastructure; bridges, tunnels, and causeways for transporation networks; and hospitals, police stations, and fire stations for societal infrastructure. In doing so, we envision designing additional novel visual representations and interactions for configuring these components as well as visualizing their proxy, intermediate, and final results.

Acknowledgments

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under award no. 2009-ST-061-CI0002.

REFERENCES

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 191–200, 2011.

[2] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, June 2008.

[3] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Verlag, 2006.

[4] N. V. Andrienko and G. L. Andrienko. Interactive visual tools to explore spatio-temporal variation. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, pages 417–420, 2004.

[5] N. V. Andrienko and G. L. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.

[6] L. Anselin. From SpaceStat to CyberGIS: Twenty years of spatial data analysis software. *International Regional Science Review*, 35(2):131–157, 2012.

[7] J. Biddiscombe, J. Soumagne, G. Oger, D. Guibert, and J.-G. Piccinali. Parallel computational steering and analysis for HPC applications using a ParaView interface and the HDF5 DSM virtual file driver. In *Proceedings of the Eurographics Conference on Parallel Graphics and Visualization*, pages 91–100, 2011.

[8] B. Broeksema, T. Baudel, A. G. Telea, and P. Crisafulli. Decision exploration lab: A visual analytics solution for decision management. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1972–1981, 2013.

[9] S. Buckley and C. An. Supply chain simulation. In *Supply Chain Management on Demand*, pages 17–35. Springer, 2005.

[10] L. Costa, O. Oliveira, G. Travieso, F. Rodrigues, P. Boas, L. Antiqueira, M. Viana, and L. Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 3(60):319–412, 2011.

[11] R. Deibert. Towards a cyber security strategy for global civil society? Technical report, The Canada Centre for Global Security Studies, 2011.

[12] D. DeLaurentis and R. K. Callaway. A system-of-systems perspective for public policy decisions. *Review of Policy Research*, 21(6):829–837, 2004.

[13] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara. Interactive focus+context analysis of large, time-dependent flow simulation data. *Simulation*, 82(12):851–865, 2006.

[14] D. Fisher, I. O. Popov, S. M. Drucker, and m. c. schraefel. Trust me, I'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1673–1682, 2012.

[15] G. Fuchs and H. Schumann. Visualizing abstract data on maps. In *Proceedings of the International Conference on Information Visualization*, pages 139–144, 2004.

[16] S. Hadlak, C. Tominski, H.-J. Schulz, and H. Schumann. Visualization of attributed hierarchical structures in a spatiotemporal context. *International Journal of Geographical Information Science*, 24(10):1497–1513, 2010.

[17] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 115–124, 2000.

[18] Q. Ho, P. H. Nguyen, T. Åström, and M. Jern. Implementation of a flow map demonstrator for analyzing commuting and migration flow statistics data. *Procedia - Social and Behavioral Sciences*, 21:157–166, 2011.

[19] B. C. Hobbs. *Food poisoning and food hygiene*. Edward Arnold and Co., London, United Kingdom, 1953.

[20] A. Kamran and S. U. Haq. Visualizations and analytics for supply chains. Technical report, IBM, February 2013.

[21] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. Cleveland, S. Grannis, and D. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *Computer Graphics and Applications, IEEE*, 29(3):18–28, May 2009.

[22] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics and Applications*, 29(3):18–28, 2009.

[23] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010.

[24] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, M. Wade, and D. S. Ebert. Understanding syndromic hotspots - a visual analytics approach. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technologoy*, pages 35–42, 2008.

[25] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert. Visual analytics law enforcement toolkit. In *Proceedings of the IEEE Conference on Technologies for Homeland Security*, pages 222–228, 2010.

[26] A. Malik, R. Maciejewski, B. Maule, and D. S. Ebert. A visual analytics process for maritime resource allocation and risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 221–230, 2011.

[27] K. Matkovic, D. Gracanin, M. Jelovic, A. Ammer, A. Lez, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1449–1457, 2010.

[28] K. Matkovic, D. Gracanin, M. Jelovic, and Y. Cao. Adaptive interactive multi-resolution computational steering for complex engineering systems. In *Proceedings of the EuroVis Workshop on Visual Analytics*, pages 45–48, 2011.

[29] J. D. Mulder, J. J. van Wijk, and R. van Liere. A survey of computational steering environments. *Future Generation Computer Systems*, 15(1):119–129, 1999.

[30] C. Ncube. On the engineering of systems of systems: key challenges for the requirements engineering community. In *Proceedings of the IEEE Workshop on Requirements Engineering for Systems, Services and Systems-of-Systems*, pages 70–73, 2011.

[31] R. Perez. Supply chain model, April 2011.

[32] Z. Pousman, J. T. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[33] H. Ribicic, J. Waser, R. Fuchs, G. Bloschl, and E. Gröller. Visual analysis and steering of flooding simulations. *IEEE Transactions on Visualization and Computer Graphics*, 19(6):1062–1075, 2013.

[34] G. Satell. Why our numbers are always wrong. *Digital Tonto*, October 2012.

[35] G. Satell. Why the future of innovation is simulation. *Forbes*, July 2013.

[36] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[37] S. Terzi and S. Cavalieri. Simulation in the supply chain context: a survey. *Computers in industry*, 53(1):3–16, 2004.

[38] J. J. Thomas and K. A. Cook. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[39] C. Tominski, J. Abello, and H. Schumann. Axes-based visualizations with radial layouts. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1242–1247, 2004.

[40] J. S. Vetter and K. Schwan. Progress: A toolkit for interactive program steering. Technical Report GIT-CC-95-16, Georgia Institute of Technology, 1995.

[41] J. Waser, R. Fuchs, H. Ribicic, B. Schindler, G. Bloschl, and E. Gröller. World lines. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1458–1467, 2010.

[42] J. Waser, H. Ribicic, R. Fuchs, C. Hirsch, B. Schindler, G. Bloschl, and E. Gröller. Nodes on ropes: A comprehensive data and control flow for steering ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1872–1881, 2011.

Vis Comput (2014) 30:493–505
DOI 10.1007/s00371-013-0874-5

ORIGINAL ARTICLE

Vehicle object retargeting from dynamic traffic videos for real-time visualisation

Simon Walton · Kai Berger · David Ebert · Min Chen

Published online: 11 September 2013

Abstract One form of video visualisation is to transform traffic videos from a street view to an aerial view, which facilitates a summary overview of multiple traffic video streams. This paper presents an efficient and effective solution to mitigate the undesirable distortion of the re-targeted vehicle objects in traffic video visualisation. This is achieved by a series of automated algorithmic steps, including vehicle segmentation, vehicle roof detection, and non-uniform image deformation by applying a second homography. This technique has been integrated into a video visualisation system that creates an aerial view of re-targeted video streams on top of a conventional aerial view. The results have shown that the technique offers the system a significant improvement in visual quality without undermining the requirement for real-time video visualisation.

Keywords Video visualisation · Retargeting · Traffic visualisation · Car roof detection · Optical flow and edge detection

1 Introduction

In visual computing, *re-targeting* is a family of techniques for transferring an object or attributes of an object from one scene to another. These techniques support a broad range of image synthesis: for instance, including image and video re-targeting for displaying large images on mobile phones (e.g., [1, 2]), for content adaptation in drama production process [3], scene relighting (e.g., [4]), motion re-targeting and expression re-targeting in computer animation [5, 6], and design and style re-targeting [7, 8].

This paper is concerned with the problem of vehicle object re-targeting in the context of traffic video visualisation. One particular form of video visualisation requires the synthesis of an aerial view (i.e., helicopter or satellite view) from a camera view [9]. Such a visualisation enables users to build up a mental overview of the information captured on multiple traffic video streams in an *intuitive* manner. One difficulty in creating such visualisation is the unknown deformation to be performed on the moving vehicle objects that are being re-targeted from the original traffic video to a new visualisation. As illustrated in Fig. 1, the white van captured on the video appears to be badly distorted in (b) after being re-targeted onto the new visualisation based on a aerial view. Although users can mentally correct such distortion relatively at ease during visualisation, it is still highly desirable to correct such unwanted distortion visually as continuing watching poor quality visualisation would burden users with unnecessary cognitive load and possibly stress. One may suggest that such a problem could be solved by re-constructing a 3D object of the vehicle from multiple camera views, and re-rendering it from the aerial view. However, in most systems of this form of video visualisation, users are usually restricted by limited availability of cameras. Hence, it is more practical to assume each vehicle may only be captured on a single camera at a time. In addition, it is often necessary to synthesise visualisation from dynamic traffic video streams in real time, and hence the computation cost must be kept as low as possible. In this paper, we present a novel technical solution to mitigate this problem in the context of traffic video visualisation. Building on a traffic video visualisation system [9], the solution makes use of

Electronic supplementary material The online version of this article (doi:10.1007/s00371-013-0874-5) contains supplementary material, which is available to authorised users.

S. Walton · K. Berger (✉) · D. Ebert · M. Chen
University of Oxford, Oxford, UK
e-mail: kai.berger@oerc.ox.ac.uk

(a) Input (b) Re-target (naïve) (c) Re-target (Ours)

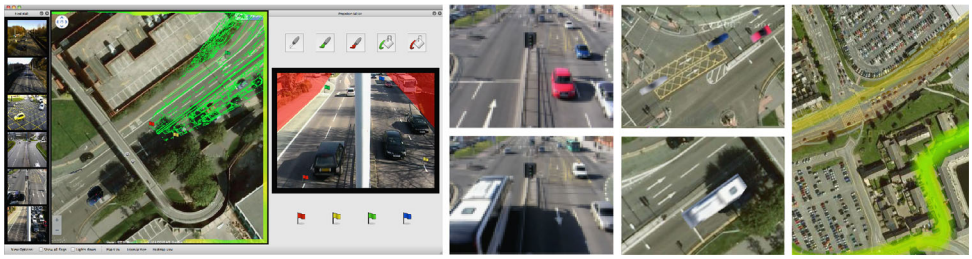
Fig. 1 In traffic video visualisation, object re-targeting from camera view footage to an aerial view of the same scene is necessary for offering an overview visualisation of traffic patterns. A naïve homography-based re-targeting leads to severe distortion effects in (**b**). We make use of optical flow to assist in segmenting the roofs of vehicles and in deriving a second homography to mitigate the distortions in (**c**)

a series of automated analytical steps to segment each vehicle from the background, estimate the top plane of a 3D bounding box of the vehicle, and apply a second homography to the vehicle when it is re-targeted onto the new visualisation.

The results show that not only can this technique mitigate distortions of almost all typical re-targeted vehicles in traffic video visualisation, but also can be performed in real time. In the remainder of this paper, we first provide a brief overview of the related works in Sect. 2, focusing on aspects of video visualisation, re-targeting in visual computing, and the relevant advances in video processing. In Sect. 3, we summarise a system that re-targets traffic video footages from camera view to a planar visualisation on top of a conventional aerial view such as a satellite image. In Sect. 5, we describe the effects and causes of distortion in video object re-targeting and we detail the technique for distortion mitigation. This is followed by Sect. 6 where we present and discuss a set of results. We offer our concluding remarks in Sect. 7.

2 Related work

The transformation of camera view video footage to cartographic aerial view can be classified as an image re-targeting operation on the input footage. Image re-targeting addresses the task to resize an image to a different size regardless of the resulting aspect ratio and is usually implemented by minimising a formulated energy functional [10]. In this context, Liu et al. [11] presented a method to intelligently crop video streams for mobile phone systems while allowing the user to interactively pan and resize the crop window. Carroll et al. [12] re-target images in order to change the perspective in the image artistically. Their energy functional is quadratic in the size of the re-targeted vertex positions. In another approach, Sacht et al. [13] re-target images from perspective views to cylindrical views. Video visualisation was first proposed by Daniel and Chen [14] as a means for summarizing a video segment into a single static image, and was later formalised by Botchen et al. [15]. Chen et al. [16] confirmed that using this kind of summarising, ordinary users can learn to detect and recognise *visual signatures* of events from video visualisation. Wang et al. [17] proposed to combine videos with a 3D environment model to support situational understanding. This can be considered as a simple form of re-targeting of an entire video stream. Remero et al. [18] investigated the visualisation of activities captured by aerial view in natural settings. Legg et al. used homographic projection to reconstruct relatively simple 3D scenes from a single camera view [19] and Parry et al. applied this technique to sports video visualisation [20]. Botchen et al. selected image frames in a video streams and re-targeted them into a summary visualisation, using VideoPerpetuoGram as the background context [21]. Höferlin made use of VideoPerpetuoGram in sports video visualisation [22]. Borgo et al. provided a comprehensive survey on video-based graphics and video visualisation [23]. There are many types of sensory device for collecting real-time road traffic information. Hoummady provided an excellent set of comments on the shortcomings of each type of sensory device in [24], and also proposed the use of video cameras as data sources for road traffic management. His proposal relies primarily on a computational device that is capable of analysing the scene and recognising 'vehicles, pedestrians, 2-wheel vehicles, etc.' automatically. The proposal does not include any detailed algorithmic solution to the traffic video footage analysis, which remains to be a very challenging technical problem in computer vision; nor does it involve the use of any visualisation techniques. There has been a huge amount of effort in developing techniques for scene analysis and object recognition in the context of traffic video footage processing. Zhu and Li [25] used a simple threshold-based background model for tracking moving vehicles, combined with smoothing filters. Vibha et al. [26] gave a vehicle-counting system, based on counting connected components of the background mask (also using a threshold-based background model). Using background models such as Gaussian mixture models to survey moving objects presents issues with the learning rate adopted that governs how quickly the model reacts to newly-background areas of the image. With a slower learning rate, moving objects tend to leave a trail of pixels that are struggling to learn about the background left in the objects' wake. Many specialised algorithms have also been developed for traffic video footage, such as identification of ground-plane homographies [27]. However, the state of the art frameworks rely heavily on 'two types of a priori information: (1) the contextual information of the camera's field



(a) (b) (c) (d)

Fig. 2 The graphical user interface [9], showing the user creating a homography using 'Flag & Cut' (**a**). Some traffic video footage (**b**) is re-targeted for visualisation in aerial view (**c**, distortion mitigation applied). Additionally, the current traffic flow (based on pixel occupancy) may be rendered as a heatmap atop the aerial view (**d**)

of view and (2) sets of predefined behaviour scenarios' [28]. As semantic specifications at different cameras vary significantly, it is difficult to port an automatic solution from a laboratory camera to a large number of real world cameras. Because it is problematic (or at least time-consuming) to specify such semantic information for each individual camera, using automatic computer vision to gather information from videos is yet to become a practical solution. This work focuses on the problem of mitigation of re-targeting errors. It is beyond the scope of this work to address the challenges in computer vision. We thus built on a traffic video visualisation system [9] that takes an approach that does not rely on automatic scene analysis and object recognition, and provides an easy-to-use user interface for entering necessary contextual information to generate dynamic visualisations without involving object and event recognition.

The most commonly-used convention for road traffic video visualisation is to colour-code lines or areas that represent roads on a map (e.g., [29]). Other traditional visual forms, such as time series plots and heatmaps (without using geographical maps) have also been used to visualise traffic information (e.g., [30]). Ang et al. [31] gave a visual analytical approach to traffic surveillance from multiple cameras with feature extraction to estimate vehicle trajectories.

3 System overview

We are building on a traffic video visualisation system [9] shown in Fig. 2 that addresses the following technical goals:

1. A video typically captures the perspective projection of a real-world scene, which is referred to as the camera view. The system has to re-target the video data onto a single aerial view, i.e., a 2D plane, with imagery information corresponding to the video.
2. The system has to remove the non-traffic (or *background*) information on the road in the input footage.
3. The object re-targeting has to be performed in real-time.
4. The system has to combine multiple traffic feeds from different camera views onto a single aerial view.

State-of-the art sensor based traffic monitoring systems (e.g., fibre-optic, piezo electric and inductive loops, infrared, laser and Bluetooth-based sensors, GPS, mobile phone data, and weight-in-motion) usually provide relatively accurate information about numbers of vehicles, but their deployment is usually restricted by the shortcomings of each type of sensor: for instance, weight-in-motion sensors and loop sensors require intrusive installation on the road, and are difficult to deploy in un-gated open space (e.g., some large car parks). Above-ground optical and electromagnetic sensors (excluding videos) have limited signal ranging, are erroneous in measuring complex traffic flows, and confine real-time traffic video visualisation to a limited set of visual designs—typically colour-coded lines or textual annotation on maps. The above-mentioned aspects were implemented in the functional order of a typical visualisation pipeline (Sect. 4): the first functional module focuses on data filtering and enrichment, transforming the raw pixel data to more meaningful information. The second module handles the necessary geometric transformation, mapping visual information from camera view to 2D planar space (i.e., pseudo aerial view). The third module renders the visual information from a pseudo-aerial view, and performs image-space composition with a new image background of a conventional aerial view. The three functional modules form a pipeline for generating a video-like layer rendering. In addition, the system provides an alternative pipeline for rendering an estimated traffic density information in aerial views, which we are not discussing further.

4 Video object re-targeting

Let $\mathscr{I} = \{I_1, I_2, \ldots, I_n\}$ an incoming video stream, where I_t be an image frame at time t. Our goal is to select a subset of pixels, $P_t \subset I_t$, from each frame, such that P_t are relevant to the objects of interest, in our case, the vehicle objects. P_t,

which may contain multiple vehicles, is then re-targeted in real-time to be displayed as foreground objects in a dynamic visualisation with an aerial-view background map. The layer where the foreground objects are displayed is referred to as a *live layer*. The visualisation framework for the live layer is described in further detail in Sect. 4.1.

A naive way to re-target P_t is to apply a homograhic transformation to each pixel $\mathbf{p} \in P_t$, resulting $\mathbf{p} \rightarrow \mathbf{p}'$. This process is described in Sect. 4.2. As this transformation can introduce significant distortion, which is unpleasant in viewing the dynamic visualisation. We introduce a second transformation $\mathbf{p}' \rightarrow \mathbf{p}''$ to mitigating such distortion. This will be described in Sect. 5.

Collectively, the two transformations re-target P_t to P_t'', which is displayed in the live layer. Thus, the integration of $P_1'', P_2'', \ldots, P_n''$ and the background map yields a dynamic visualisation, with animated vehicles on the roads in an aerial-view map.

4.1 Visualisation framework

This work has been built on a traffic video visualisation system [9] that aims to enable human operators to observe traffic patterns more effectively by transforming ground-level raw video data to synthetic aerial view videos. This can reduce cognitive load in building a mental model of traffic overview based on a number of camera views. Its focus lies on *data enrichment* with pixel-based analysis, i.e., extracting information from and augmenting pixels without relying on any algorithms for understanding the scene and its objects. This approach brings some advantages: (i) the pixel-level errors and uncertainty have less impact upon the overall perception than object-level errors and uncertainty; (ii) it does not require costly data collection and annotation for supervised training in machine learning (iii) pixel-based data filtering and enrichment requires less computational resources and is highly parallelisable. The existing system only classifies pixels in the video streams as either foreground or background by employing a background model. Any background model has inherent uncertainty, with the definition of 'certainty' tied to the context of the scene. For urban traffic video footage where traffic frequently comes to a halt (e.g., at a set of traffic lights), it is unacceptable for such traffic to be absorbed into the background. Therefore, the system accepts this uncertainty and introduces methods to ensure that temporarily-stationary vehicles on the layer rendering projection are still visible. This is implemented by attaching uncertainty data to a standard Gaussian Mixture Model (GMM) background model via histogram information. Using histogram information, the system estimates the predominant hue of the road area beneath a camera's projection $road_{rgb}$ and defines a heuristic that models the certainty of a pixel $\mathbf{p}$ being a *foreground* pixel:

$$certainty(\mathbf{p}) = 1 - \min\left(\|\mathbf{p}_{rgb} - road_{rgb}\|^2, 1\right) * tc(\mathbf{p})_{\mu}$$

where $tc(\mathbf{p})$ gives the topmost GMM cluster at $\mathbf{p}$, and its μ property denotes the amount of time that cluster has spent in its current sorted position. $\|\mathbf{p}_{rgb} - road_{rgb}\|$ gives the RGB colour space distance between the video colour at $\mathbf{p}$ and $road_{rgb}$. If a vehicle moves into $\mathbf{p}$ and stops, $\mathbf{p}$ will eventually become part of the background. Setting a smaller learning rate can alleviate this somewhat, but at the expense of slower convergence elsewhere. If $\mathbf{p}$'s topmost GMM cluster changes frequently, then this probably indicates movement in the area. When a cluster's μ increases, it becomes less certain as time goes on whether the cluster's pixel represents the road or whether it represents a stationary vehicle– thus, the uncertainty can be judged from μ.

4.2 Homography definition

Object re-targeting in video visualisation is different from that in most other visualisation techniques. Video data contains camera view spatial information, that is, the 3D real world geometry is represented as projected 2D geometry. Unlike many spatial data types in scientific visualisation (e.g., computed tomography data), the camera view representation in the camera view cannot be directly used by a renderer unless the rendering viewpoint coincides with that of the camera. In the employed visualisation system, the basic form of geometry in the final visualisation is predetermined by the requirement for a top-down view emulating a satellite viewpoint. While algorithms exist to automatically extract road features [32], e.g., for providing point correspondences for a homography estimation, the employed visualisation system utilises human intelligence instead to define how the video from each camera corresponds to its associated projection on the map. The three aspects of this definition are: camera view, aerial view, and vehicle.

The user first initiates the definition of a new camera to the system by locating an available camera feed on the feed wall and dragging it into the main aerial view, where the system places the camera on the map (represented by a small camera icon). Once a camera is dropped, the projection editor opens for the user to begin the process of defining four point correspondences of the homography H_{planar} for the camera, e.g., distinctive lane marks or signposts.

We use a point correspondence system to compute the homography. The corner points for the defining polygon are shown to the user as small flags of colours red, yellow, green, and blue. It is relatively trivial for the user to learn to find places on the video image that correspond with places on the satellite image given in the main map view—we have found that cues such as lane arrows, box junction hatches, signposts, lampposts, and other road markings provide useful flag positions. There exists the possibility of discrepancies in the video image obtained from the camera and the satellite imagery provided on the map due to out-of-date satellite

Fig. 3 A comparison of different rendering modes: (**a**) basic projection of the video onto the map; (**b**) the projection with the user mask enabled; (**c**) Gaussian Mixture Model & histogram matching method, distortions mitigated

imagery or obstructions. In such cases, the user relies more on visual feedback to converge to a suitable result.

Inaccuracies in the homography estimation be visible in the visualisation phase afterward, but the relative point locations remain sufficiently accurate after re-targeting for traffic monitoring by a human user. The point correspondences, however, may be replaced at any time (see Fig. 2). With the homography computed, the user can begin to define which portions of the video are active, and which are inactive. By 'inactive', we imply regions of the input video that provide no benefit when projected onto the map (pavements, sky, etc.). We provide three tools for this: (i) the scalpel for cutting polygonal regions away; (ii) the brush for brushing regions; and (iii) the bucket for the contiguous flood-fill of regions. All tools can function in both video-space and mapspace.

4.3 Visual mapping strategy

The employed traffic video visualisation system adapts a histogram matching method to *blend* the projections of the objects onto an existing satellite map. The source histogram is the histogram of the road of the video, and the target is the histogram of the road of the aerial view. It determines automatically which pixels of the video and which pixels of the aerial view are likely those of the road surface (obstructed by vehicles occasionally); see Fig. 3.

The rendering strategy utilises the information on background certainty by favouring pixels that have a high certainty of being foreground over pixels that have a lower certainty of being foreground: Given a pixel $\mathbf{p}$ in the camera

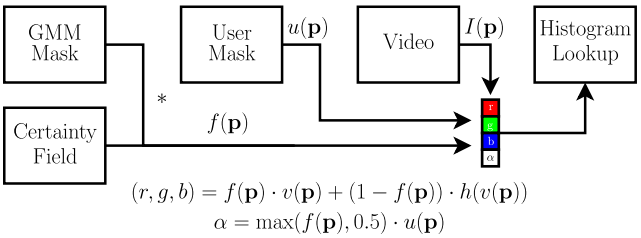


Fig. 4 A pixel $\mathbf{p}$'s final fragment colour is decided upon based on the binary user mask $u(\mathbf{p})$, the Gaussian Mixture Model binary mask multiplied by the $[0, 1]$ foreground mask $f(\mathbf{p})$, the video frame I, and the histogram lookup table $h(\cdot)$

view related to some point in aerial view, the system associates three attributes:

- A user mask $u(\mathbf{p})$ with binary value $\{0, 1\}$ where 0 corresponds to 'inactive' and 1 corresponds to 'active';
- A certainty field $c(\mathbf{p})$ with values $[0, 1] \in \mathbb{R}$ ranging from 0: 'uncertain' to 1: 'certain';
- A foreground mask $f(\mathbf{p})$ (obtained from the Gaussian Mixture Model) with binary value $\{0, 1\}$ where 0: 'background' and 1: 'foreground'.

The system then computes the final fragment colour and alpha values as in Fig. 4. The histogram function $h(\cdot)$ matches only a limited colour range to blend the video road colour into the map road colour. Note that the colour distance of the input colour to the predominant road colour is already taken into account in the certainty field. Occasionally we find that this strategy can partially fail where a car of similar colour to the road absorbs into the background, but we accept this uncertainty given the relatively low level of incidence.

Figure 5 gives an overview of the histogram matching computation. First, the mask is projected onto the map using homography H_{planar} to remove unneeded scene regions. Next, a clustering algorithm is applied to the pixel colour values of the underlying road surface in aerial view to ascertain the most frequent colour cluster that is found on the scene. A cluster's weight is biased according to the ratio of its RGB values:

$$\left(1 - \left(\left|I(\mathbf{p})_r - I(\mathbf{p})_g\right|\right)\right) \cdot \left(1 - \left(\left|I(\mathbf{p})_b - k \cdot I(\mathbf{p})_b\right|\right)\right)$$

where $k = 0.94$. The above equation gives a higher weight to pixels that have similar ratios of red to green with a relative ≈ 6 % drop in the blue component. k's 0.94 value has been found empirically by measuring the colour distributions of roads from the satellite imagery: the roads are generally grey but with a drop in blue due to the wavelength of the Sun's light. Once the clustering algorithm completes, a binary mask is created of those pixels on the map belonging to the winning cluster. Once this stage is completed, this mask is projected back onto the video using matrix H^{-1}_{planar}. From the GMM, a 'background' image is obtained by taking the

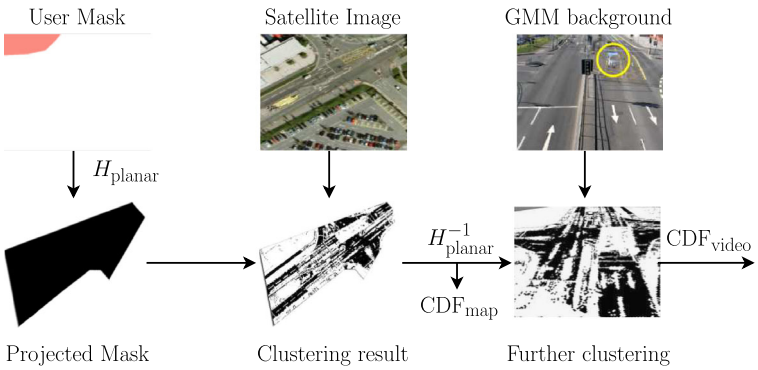


Fig. 5 Computing the histograms to be used for histogram matching. *Black pixels* represent the active area and *white pixels* represent the inactive area

current mean colour values of the top distribution for each pixel. Using this background image, a final clustering operation is performed on the video cluster mask using colour information from the GMM background. This final clustering step removes any erroneous information such as shadows being included in the CDF computation. Finally, a RGB histogram lookup table is computed using the source and target CDFs. In Fig. 5, the limitations of the algorithm are shown, where a blue car present at the beginning of the sequence (circled) has partially absorbed into the background.

To render each camera's projection, the system provides the histogram matching process with the map imagery beneath the projection. The camera's foreground mask $f(\mathbf{p})$, certainty field $c(\mathbf{p})$, current camera frame, and histogram LUT are uploaded to the GPU as textures. The map's projected boundaries are rendered to a bound framebuffer object, and the inverse homography H^{-1}_{planar} is used in a fragment shader to obtain the final video texture coordinate, discarding the fragment if this coordinate falls outside the [0, 1] boundary or if the user mask indicates that the sampled area is inactive. The blending operation shown in Fig. 4 is then performed obtain to obtain the final fragment colour. Once all camera views have been rendered to the framebuffer object, the aerial view texture is rendered to screen, and a morphological closure is performed on the final rendering (to reduce 'speckles' in the background model).

We have developed a visual mapping system that implements a conceptually straightforward pixel-based algorithm, mapping directly to an intuitive visualisation that shows the user a real-time estimate of the traffic's speed (the amount of foreground change per frame), density, and certainty level. Our algorithm can be viewed as a hybrid of a sensor-based and a video-based input. We treat each projected pixel in the scene as a switch, triggering whenever a car moves over it: Once the *density*, *change*, and *certainty* values have been accumulated, they are normalised by dividing by the total. For a given time window, a *rolling mean* is calculated for a collection of normalised values of the same type (e.g., *density*) obtained from the series of consecutive frames in the window. The rolling means for *density*, *change*, and *certainty* are then passed onto the timeline and heatmap visualisation as detailed below. The uncertainty in this system can be classified as *validity uncertainty*, that is, the uncertainty is based on deductive inferences rather than, for example, the quality of the data.

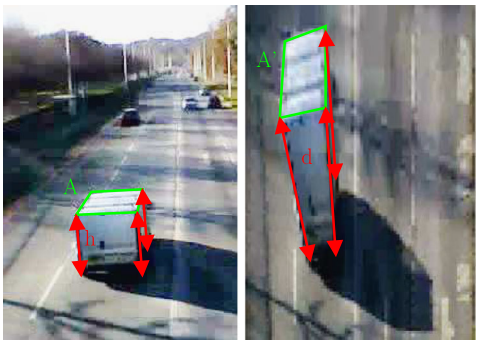


Fig. 6 The distortions d in the aerial view are related to the vehicles' heights h. The higher a vehicle the more it gets stretched after re-targeting. The area A of a vehicle's roof stays approximately constant (A')

5 Mitigating distortions of re-targeted objects

Height-related re-targeting distortions in the aerial view show severely for objects that move toward the horizon or that move perpendicular to the viewing axis of the camera; see Fig. 6. The vehicle is usually imaged with its roof and two sides, e.g., back and right side. While the imaged area A of the roof directly relates to the comprised area A' of the vehicle on street level, the sides only give sensible information about its height h. After applying the homography H_{planar} the imaged areas of the side mainly contribute of to the perceived distortion d while the roof appears to be displaced.

Ideally, only the roof, i.e., the top side of an approximated cuboid, should be re-targeted into the aerial view. It has to be noted that due to the fact that the roof is located at a height different than the ground plane, the roof has to be re-targeted using a different homography $H_{\text{planar,height}} \neq H_{\text{planar}}$. This

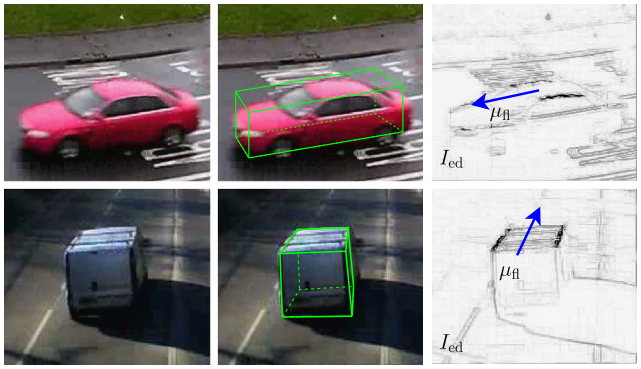


Fig. 7 A vehicle (*left column*) can be approximated geometrically as a cuboid (*middle column*) whose top plane correlates to the comprised area in a aerial view. Its side planes usually contribute to distortions. We detect the roof by angular thresholding in I_{ed} against the mean optical flow μ_{fl} (*right column*)

implicates that the user has to define two different homographies, i.e., 8 point correspondences instead of 4, to the system, Sect. 4, for each camera view footage under examination, Fig. 9. Note that a second homography is necessary to avoid large misplacements of the roof plane in the aerial view. In the following, we extract the vehicle roof in the input images by assuming (Fig. 7):

1. Vehicles can be approximated by cuboids
2. A vehicle roof is planar and rectangular
3. The main axis of its roof aligns to its velocity vector

Assume that a point p on the edge of the roof of a vehicle object V has a position in world space at times t and $t+1$ defined by $W_t(p)$ and $W_{t+1}(p)$, with W_* the rigid transformation of the vehicle object to the world space. Let now a camera projection matrix P for a static camera project the point p on the vehicle object V to pixel positions $\mathbf{i}_t$ and $\mathbf{i}_{t+1}$ in the camera image with $P(W_t(p))$ and $P(W_{t+1}(p))$. Assume further an adjacent, but visually distinguishable, point p' on the edge of the roof of V, which is projected to $\mathbf{i}'_t$ and $\mathbf{i}'_{t+1}$ with $P(W_t(p'))$ and $P(W_{t+1}(p'))$. The difference vector in the camera image between $\mathbf{i}_t$ and $\mathbf{i}_{t+1}$ can be described by

$$\mathbf{d} = \mathbf{i}_{t+1} - \mathbf{i}_t \tag{1}$$

Further, the projected edge between the adjacent points p, p' in the camera image can be described by

$$\mathbf{e}_t = \mathbf{i}_t - \mathbf{i}'_t \tag{2}$$

and

$$\mathbf{e}_{t+1} = \mathbf{i}_{t+1} - \mathbf{i}'_{t+1} \tag{3}$$

for both times t and $t+1$, respectively. There exists a reasonable velocity with which the vehicle moves, so that the point p' will be projected at $t+1$ close to the projected position of its adjacent point p at time t, thus $\mathbf{i}'_{t+1} \approx \mathbf{i}_t$ (this holds for the case when p' is located further to the rear of the vehicle than p). Inserting the inequality to Eq. (1) and Eq. (3), we get

$$\mathbf{d} \approx \mathbf{e}_{t+1} \tag{4}$$

We define the residual error of the approximation by the angular difference $atan2(\|\mathbf{d} - \mathbf{e}\|)$. Following the above reasoning, we calculate an optical flow $\mathbf{d}$ between two image frames at times t and $t+1$ and compare it with the edge image at time $t+1$ by computing the angular difference between the two images. Edge pixels with a small residual error are considered as belonging to the vehicles roof. Note that by calculating the angular difference only, we account for variation in velocity and by taking the absolute difference we account for the case that p' is not located further to the rear of the vehicle than p.

Based on the derived equations, we define the algorithm of our method, Fig. 8 as follows: At first segment, the vehicles in I_t by background subtraction, similar to Sect. 4, and calculates a dense optical flow I_{fl} [33] between downscaled versions (factor $\frac{1}{8}$) of I_t and I_{t-1}. We use a downscaled version to average out small texture-based deviations in the optical flow image. For each vehicle, i.e., component in I_{cc}, we calculate its mean flow vector μ_{fl} to arrive at a $I_{\text{fl,cc}}$. Then the edges in the image are compared to μ_{fl} for each component in I_{cc}. Edges, which roughly align to μ_{fl}, are the vehicle's roof or bottom boundary. These edges are filtered by angular thresholding to arrive at I_{thresh}. The final area of the roof is computed by performing a flood fill between the remaining edges for each component in I_{cc}. The area is re-targeted in to the aerial view with $H_{\text{planar,height}}$ to account for height differences. The missing image parts in the aerial view, which are generated by leaving out the vehicle's sides, are filled with information from the re-targeted background image $H_{\text{planar}} \star I_{\text{bg}}$. The long sides are then compared to the remaining velocity vector.

6 Results

We tested the proposed algorithm for reducing height-related re-targeting distortions with three different traffic scenarios:

- A motorway scene, with traffic flow mainly aligned to the viewing axis;
- A roundabout scene with traffic flow perpendicular to the viewing axis;
- A city bypass scene recorded from a steep angle.

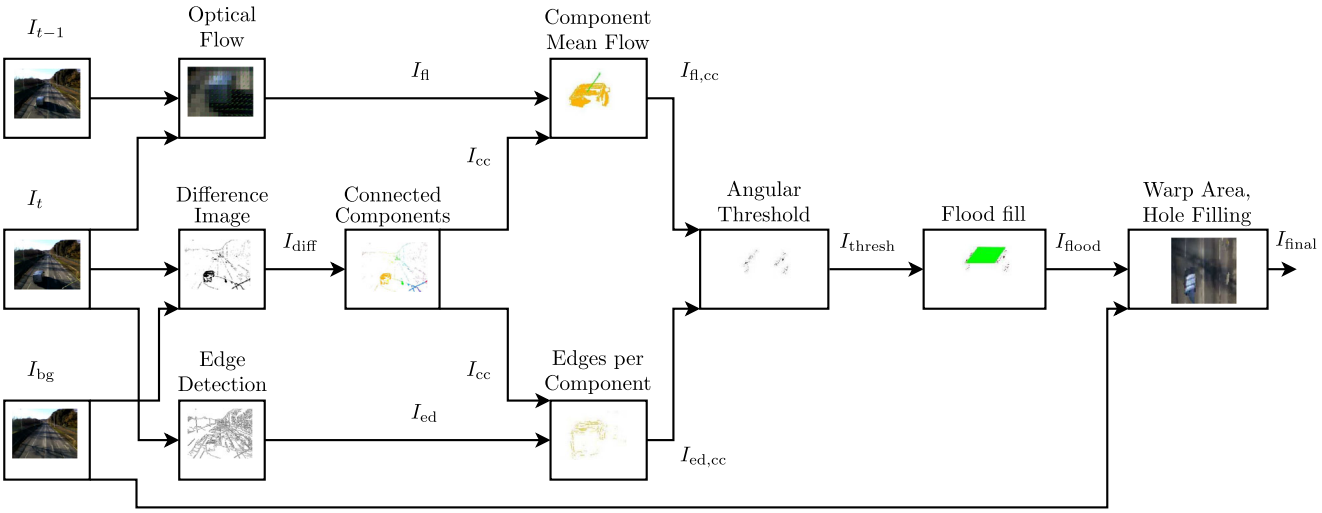


Fig. 8 The flow chart of our re-tageting algorithm as described in Sect. 5. It takes an average background image I_{bg} and two consecutive frames I_t and I_{t-1} (*first column*). Then the algorithm computes the flow field I_{fl} [33] between the frames (*second column*, *top row*) and the connected components I_{cc} in the current frame I_t (*third column*). As the mean flow vector of the vehicle in $I_{\text{fl,cc}}$ (*fourth column*, *top row*) should roughly align to the edges of a vehicle's roof in $I_{\text{ed,cc}}$ (*fourth column*, *bottom row*), an angular thresholding is performed (*fifth column*). The comprised area of each vehicle's roof is found by a floodfill in I_{thresh} (*sixth column*) and then re-targeted into the aerial view $H_{\text{planar,height}} \star I_{\text{flood}}$ (*seventh column*)

Table 1 We calculated the average angular error (AAE) of the computed mean optical flow μ_{fl} [33] to the estimated velocity vector for downsampled versions of the input sequence with different scaling factors. With decreasing input image size, the computed mean flow better approximates the velocity vector by averaging out. This is reasoned by the fact that a downsampled image averages out texture-based pixel deviations

Average angular error (°) vs. scale factor			
Scale factor	Motorway	Roundabout	Bypass
1	5.227°	4.070°	5.277°
1/2	3.171°	3.984°	4.234°
1/4	2.661°	3.899°	3.936°
1/8	1.407°	3.504°	2.149°

All traffic video footage was acquired using a set of digital camcorders recording video at 640 × 480 resolution, at six to twelve FPS. The average height of the bridges is around 20ft, and thus there was much intravehicle occlusion in each scene. It should be noted that this work currently focuses only on non-PTZ cameras.

At first we compared different downsampling factors on the input streams for its effect on the accuracy of the mean flow, Table 1. We computed the Average Angular Error (AAE) in degrees between the computed [33] mean optical flow vector and the vehicle's estimated velocity vector in the input stream. It can be seen that a downsampled image averages out small texture-based pixel deviations, and thus more robustly reveals the optical flow introduced by the vehicle's velocity. Best results could be found for a downsampling factor of $\frac{1}{8}$. At that resolution, the vehicles would only span a small amount of pixels in the image and the direction of their velocity vector would dominate the resulting optical flow. Afterward, we compared the amount of distortion on the input sequences in the aerial view for the naïve homography and the homography of our method, Table 2. The table lists five frames for each input sequence. The first row for each scene in Table 2 shows the input data before downsampling. Note, however, that the view is cropped partially to focus on the vehicles. In the second row for each scene in Table 2, the results for a naïve homography are shown in aerial view. In the third row for each scene in Table 2, the results of our method are shown in aerial view, i.e., a detected vehicle's roof is re-targeted with $H_{\text{planar,height}}$, which has been defined by the user before, Fig. 9. Missing regions in the resulting aerial view are filled with information from a background image I_{bg} and re-targeted with H_{planar}. The fourth row visualises the differences in the comprised area of the re-targeted vehicle: red for a naïve homography and green for our method. It can be seen that, while with the naïve re-targeting the comprised area varies with distance to the position of the camera, the comprised area remains more stable with our method (motorway). Also, the comprised area approximates the actual area of the vehicle imaged from a satellite position more accurately as only its roof, i.e., top plane, is re-targeted (bypass). Furthermore, the position of the re-targeted object appears to be more accurate than with naïve re-targeting (roundabout).

Table 2 Re-targeting distortion mitigation applied to three test scenes. The *first row* for each scene shows the input data before downsampling. The *second row* for each scene shows the results for a naïve homography in aerial view. The *third row* for each scene shows the results of our method in aerial view. In the fourth row for each scene, the comprised areas are compared for both methods in aerial view (*red* = naïve, *green* = our method). It can be seen that the comprised area in aerial view approximates the actual area more accurately with our method. The *first scene* shows a motorway with traffic flow mainly aligned to the viewing axis. The *second scene* shows a roundabout with traffic flow perpendicular to the viewing axis. The *third scene* shows a city bypass with multiple cars recorded from a steep angle. (*Continued*): Re-targeting distortion mitigation applied to three test scenes (*first row each*) and comparison (*fourth row each*, *red* = naïve, *green* = our method). It can be seen that the comprised area approximates the actual area more accurately with our method. The first scene shows a motorway, the *second* shows a roundabout, and the *third* shows a city bypass with multiple cars

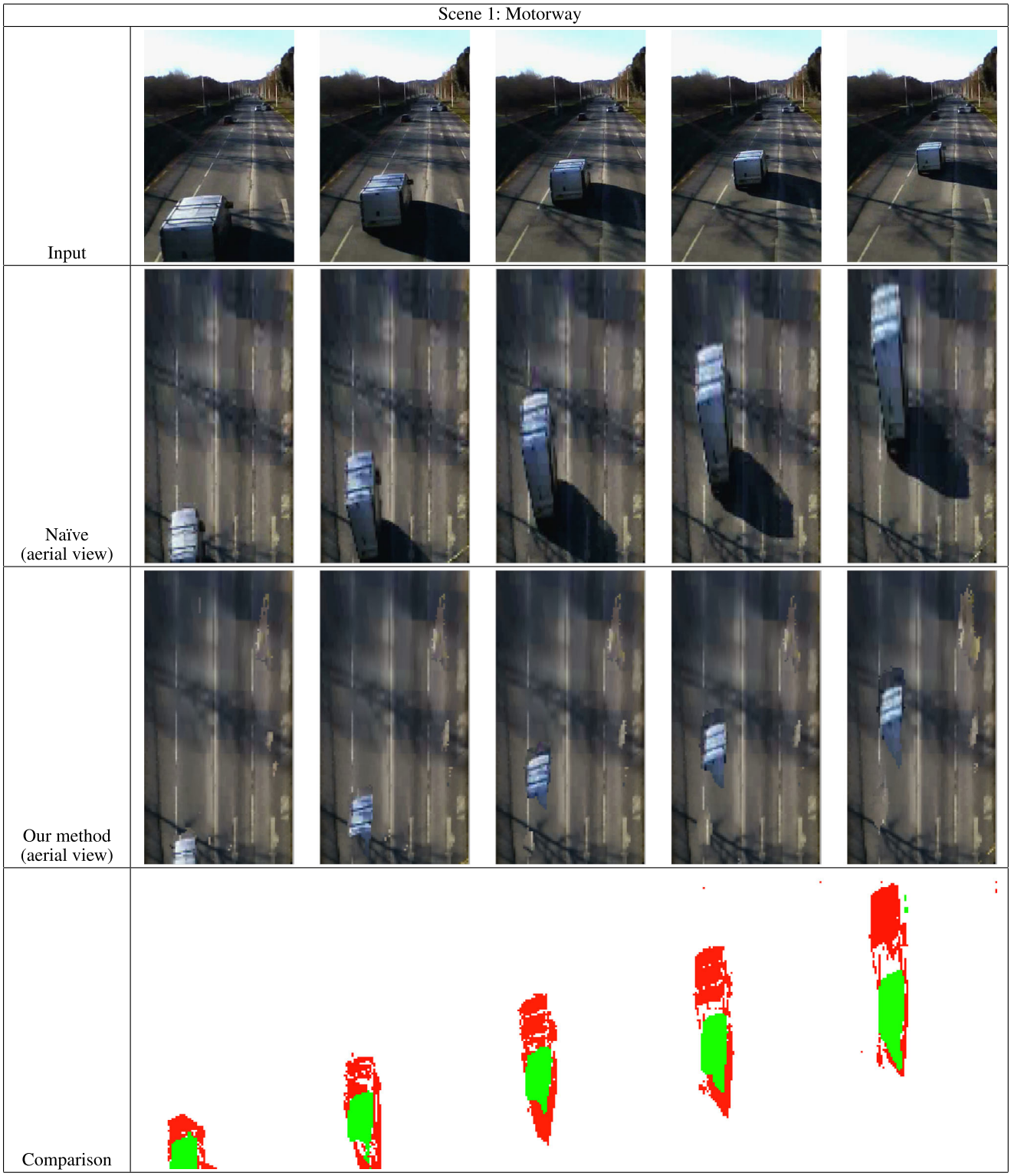


Table 2 (*Continued*)

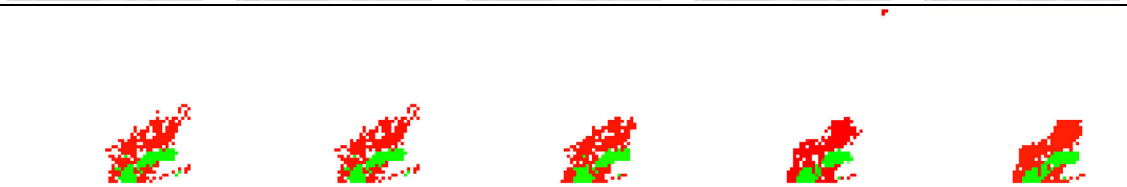
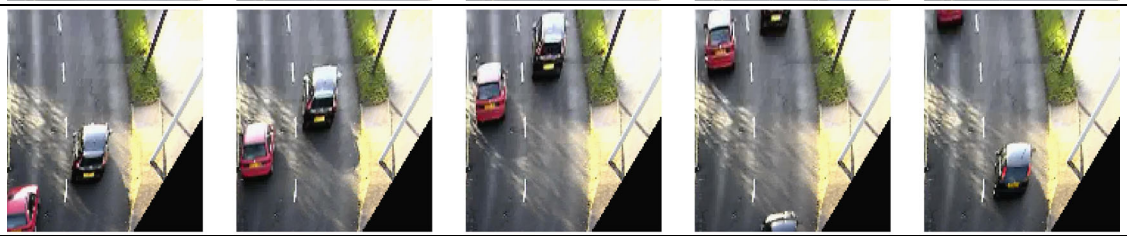
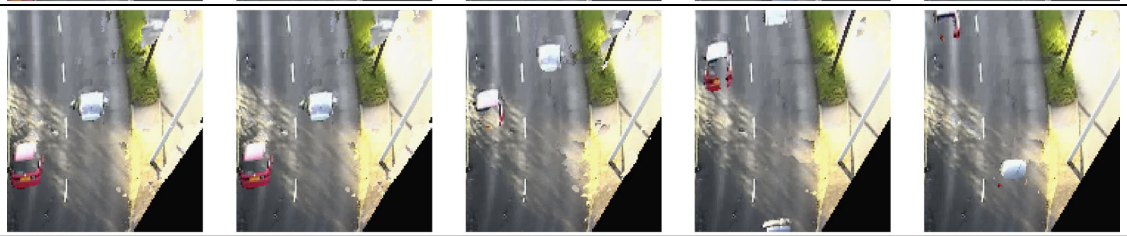
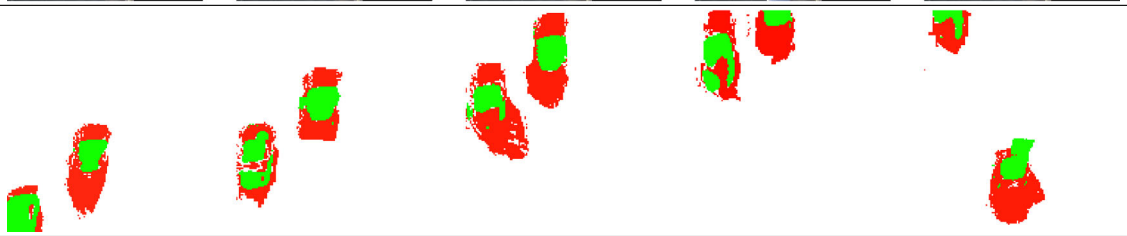
Scene 2: Roundabout	
Input	
Naïve (aerial view)	
Our method (aerial view)	
Comparison	
Scene 3: Bypass	
Input	
Naïve (aerial view)	
Our method (aerial view)	
Comparison	

Fig. 9 Homography estimation. Instead of using only one user-defined planar homography H_{planar} at ground level (*left*), we propose to use a second user-defined planar homography $H_{\text{planar,height}} \neq H_{\text{planar}}$ (*middle*) at a height different than the ground level (e.g., car level (*green*) or lorry level (*red*)). Note that both homographies, H_{planar} and $H_{\text{planar,height}}$, map to the same area in the aerial view (*right*)

7 Conclusions and discussion

We have presented a technique to mitigate distortions in object re-targeting in the context of a camera-based global-view traffic video visualisation system [9]. Such a system has the advantage over conventional monitoring systems in that the deployment is not as restricted by the shortcomings of the sensor. However, while such camera-based global-view traffic video visualisation systems directly offer an image of the global traffic situation to the user, they suffer from the fact that the re-targeted shape of the visualised objects, varies with distance to the position of the camera. After an overview over the rendering pipeline of the employed traffic video visualisation system, the distortions were identified and their causes were explained. Those distortions were addressed by applying a second homography to the vehicles' roofs, which have been detected with a simple real-time capable optical flow-based approach.

We compared our method to the naïve re-targeting method in three test scenes. It can be shown that by applying our method the re-targeted objects, i.e., the vehicles, were less distorted than with naïve re-targeting. There is much scope for future work in this area. The specification of the ground and height homography relies on the video and aerial view providing the same landmarks for registration by the user. In addition, we have concentrated only on aerial views and would like to investigate the interesting challenges that occur when they are replaced with an abstract map. Other challenges include intelligent removing of the existing vehicles on aerial view to arrive at more sophisticated background models.

Also, we would like to expand our work from vehicle movements to general object movement scenarios. Such a scenario to test the proposed approach could be intelligent pedestrian surveillance. Ellis and Ferryman, for example, collected a benchmark database [34] of pedestrian motions on a campus. A combined visualisation of the recorded scene in aerial view, i.e., a rendering of all re-targeted camera images promises a good overview and facilitates tracking of single persons. However, as the subjects' heights are usually greater than their comprised area on the ground, the re-targeted images are likely to show severe distortions. We think that our method can sufficiently help in mitigating such distortions.

References

1. Setlur, V., Lechner, T., Nienhaus, M., Gooch, B.: Retargeting images and video for preserving information saliency. IEEE Comput. Graph. Appl. **27**(5), 80–88 (2007)
2. Rubinstein Ariel Shamir, S.A.M.: Improved seam carving for video retargeting. ACM Trans. Graph. **27**(3), 16 (2008)
3. Rijsselbergen, D., Poppe, C., Verwaest, M., Mannens, E., Walle, R.: Semantic mastering: content adaptation in the creative drama production workflow. Multimed. Tools Appl. **59**(1), 307–340 (2012)
4. Ng, R., Ramamoorthi, R., Hanrahan, P.: Triple product wavelet integrals for all-frequency relighting. ACM Trans. Graph. **23**(3), 477–487 (2004)
5. Gleicher, M.: Retargetting motion to new characters. In: Proc. ACM SIGGRAPH, pp. 33–42 (1998)
6. Hecker, C., Raabe, B., Enslow, R.W., DeWeese, J., Maynard, J., van Prooijen, K.: Real-time motion retargeting to highly varied user-created morphologies. ACM Trans. Graph. **27**(3), 27:1–27:11 (2008)
7. Clarke, L., Chen, M., Mora, B.: Automatic generation of 3D caricatures based on artistic deformation styles. In: IEEE Transactions on Visualization and Computer Graphics, pp. 808–821 (2011)
8. Lin, J., Cohen-Or, D., Zhang, H., Liang, C., Sharf, A., Deussen, O., Chen, B.: Structure-preserving retargeting of irregular 3D architecture. ACM Trans. Graph. **30**(6), 183:1–183:10 (2011)
9. Walton, S., Chen, M., Ebert, D.: Livelayer: real-time traffic video visualisation on geographical maps
10. Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: Automatic image retargeting. In: Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia, pp. 59–68. ACM, New York (2005)
11. Liu, F., Gleicher, M.: Video retargeting: automating pan and scan. In: ACM Multimedia, pp. 241–250. ACM, New York (2006). doi:10.1145/1180639.1180702
12. Carroll, R., Agarwala, A., Agrawala, M.: Image warps for artistic perspective manipulation. ACM Trans. Graph. **29**(4), 127:1–127:9 (2010). doi:10.1145/1778765.1778864

13. Sacht, L., Velho, L., Nehab, D., Cicconet, M.: Scalable motion-aware panoramic videos. In: SIGGRAPH Asia 2011 Sketches, p. 37. ACM, New York (2011)
14. Daniel, G., Chen, M.: Video visualisation. In: Proc. IEEE Visualization, pp. 409–416 (2003)
15. Botchen, R.P., Bachthaler, S., Schick, F., Chen, M., Mori, G., Weiskopf, D., Ertl, T.: Action-based multifield video visualization. IEEE Trans. Vis. Comput. Graph. **14**(4), 885–899 (2008)
16. Chen, M., Botchen, R.P., Hashim, R.R., Weiskopf, D., Ertl, T., Thornton, I.M.: Visual signatures in video visualisation. IEEE Trans. Vis. Comput. Graph. **12(5)**, 1093–1100 (2006)
17. Wang, Y., Krum, D.M., Coelho, E.M., Bowman, D.A.: Contextualized videos: combining videos with environment models to support situational understanding. IEEE Trans. Vis. Comput. Graph. **13**(6), 1568–1575 (2007). doi:10.1109/TVCG.2007.70544
18. Remero, M., Summet, J., Stasko, J., Abowd, G.: Viz-a-vis: toward visualizing video through computer vision. IEEE Trans. Vis. Comput. Graph. **14**(6), 1261–1268 (2008). doi:10.1109/TVCG.2008.185
19. Legg, P., Parry, M., Chung, D., Jiang, R., Morris, A., Griffiths, I., Marshall, D., Chen, M.: Intelligent filtering by semantic importance for single-view 3D reconstruction from snooker video. In: ICIP, pp. 2433–2436 (2011)
20. Parry, M., Legg, P., Chung, D., Griffiths, I., Chen, M.: Hierarchical event selection for video storyboards with a case study on snooker video visualisation. IEEE Trans. Vis. Comput. Graph. **17**(12), 1747–1756 (2011)
21. Botchen, R., Bachthaler, S., Schick, F., Chen, M., Mori, G., Weiskopf, D., Ertl, T.: Action-based multi-field video visualisation. In: TVCG (2008)
22. Hoeferlin, M., Grundy, E., Borgo, R., Weiskopf, D., Chen, M., Griffiths, I., Griffiths, W.: Video visualization for snooker skill training. Comput. Graph. Forum **29**(3), 1053–1062 (2010)
23. Borgo, R., Chen, M., Daubney, B., Grundy, E., Jaenicke, H., Heidemann, G., Hoeferlin, B., Hoeferlin, M., Weiskopf, D., Xie, X.: A survey on video-based graphics and video visualisation. In: Eurographics 2011 STAR (2011)
24. Hoummady, B.: Method and device for managing road traffic using a video camera as data source. United States Patent No. US 6,366,219 B1 (2002)
25. Zhu, F., Li, L.: An optimized video-based traffic congestion monitoring system. In: Proc. 3rd International Conference on Knowledge Discovery and Data Mining, pp. 150–153 (2010). doi:10.1109/WKDD.2010.47
26. Vibha, L., Venkatesha, M., Prasanth, R.G., Suhas, N., Shenoy, P.D., Venugopal, K.R., Patnaik, L.M.: Moving vehicle identification using background registration technique for traffic surveillance. In: Proc. International MultiConference of Engineers and Computer Scientists, vol. I, pp. 19–21 (2008)
27. Arrospide, J., Salgado, L., Nieto, M., Mohedano, R.: Homography-based ground plane detection using a single on-board camera. IET Intell. Transp. Syst. **4**(2), 149–160 (2010). doi:10.1049/iet-its.2009.0073
28. Kumar, P., Ranganath, S., Weimin, H., Sengupta, K.: Framework for real-time behavior interpretation from traffic video. IEEE Trans. Intell. Transp. Syst. **6**(1), 43–53 (2005). doi:10.1109/TITS.2004.838219
29. Shekhar, S., Lu, C.T., Liu, R., Zhou, C.: CubeView: a system for traffic data visualisation. In: Proc. IEEE Intelligent Transportation Systems, pp. 674–678 (2002)
30. Lu, C.T., Boedihardjo, A.P., Zheng, J.A.IT.V.: Advanced interactive traffic visualisation system. In: AITVS: Proc. International Conference on Data Engineering, pp. 167–168 (2006). http://doi.ieeecomputersociety.org/10.1109/ICDE.2006.14
31. Ang, D., Shen, Y., Duraisamy, P.: Video analytics for multi-camera traffic surveillance. In: Proc. 2nd International Workshop on Computational Transportation Science, New York, NY, USA, pp. 25–30 (2009). doi:10.1145/1645373.1645378
32. He, Y., Wang, H., Zhang, B.: Color-based road detection in urban traffic scenes. IEEE Trans. Intell. Transp. Syst. **5**(4), 309–318 (2004)
33. Horn, B., Schunck, B.: Determining optical flow. Artif. Intell. **17**(1), 185–203 (1981)
34. Ellis, A., Ferryman, J.: Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In: AVSS 2010, pp. 135–142. IEEE, New York (2010)

Simon Walton received his B.Sc. in Computer Science from Swansea University in 2004 with honours. He went on and received his Ph.D. in Computer Science for his work on GPU-based Volume Deformation from Swansea University in 2007. He was a Research Assistant at Swansea University from 2009 until 2011, and is currently employed as a research associate at OeRC, University of Oxford.

Kai Berger received his Diploma in Computer Science from TU Braunschweig, Germany. He went on and received his Ph.D. in Engineering for his work on measuring and modelling light-matter interaction phenomena from TU Braunschweig in 2012. Currently, he is employed as a research associate at OeRC, University of Oxford.

David Ebert is the Silicon Valley Professor of Electrical and Computer Engineering at Purdue University, a University Faculty Scholar, a Fellow of the IEEE, and Director of the Visual Analytics for Command Control and Interoperability Center (VACCINE), the Visualisation Science team of the Department of Homeland Security's Command Control and Interoperability Center of Excellence. Dr. Ebert performs research in novel visualisation techniques, visual analytics, volume rendering, information visualisation, perceptually-based visualisation, illustrative visualisation, mobile graphics and visualisation, and procedural abstraction of complex, massive data. He has been very active in the visualisation community, teaching courses, presenting papers, co-chairing many conference program committees, serving on the ACM SIGGRAPH Executive Committee, serving as Editor in Chief of IEEE Transactions on Visualisation and Computer Graphics, serving as a member of the IEEE Computer Society's Publications Board, serving on the IEEE Com-

puter Society Board of Governors, and successfully managing a large program of external funding to develop more effective methods for visually communicating information.

Min Chen received the Ph.D. degree from University of Wales in 1991. He is currently a professor of scientific visualization at Oxford University and a fellow of Pembroke College. Before joining Oxford, he held research and faculty positions at Swansea University. His research interests inlude visualization, computer graphics and human-computer interaction. He has co-authored about 150 publications, including his recent contributions in areas of volume graphics, video visualization, face modelling, automated visualization and theory of visualization. His services to the research community inlude papers co-chair of IEEE Vis 2007 and 2008, Eurographics 2011, AEIC of IEEE TVCG, and co-director of Wales Research Institute of Visual Computing. He is fellow of BCS, EG and LSW.

EuroVis Workshop on Visual Analytics (2014)
M. Pohl and J. Roberts (Editors)

Visual Analytics for Risk-based Decision Making, Long-Term Planning, and Assessment Process

Silvia Oliveros-Torres[†1], Yang Yang[†1], Yun Jang[‡2], David Ebert[†1]

[1]Purdue University, USA, [2]Sejong University, South Korea,

Abstract

Risk-based decision making is a data-driven process used to gather data about outcomes, analyze different scenarios, and deliver informed decisions to mitigate risk. We describe the design and application of integrated visual analytics techniques and components to support risk-based decision making following a structured risk management process in the US Coast Guard domain. The components proposed perform the following interactive tasks: the identification of risk priority areas, the distribution of pre-computed risk values, and the analysis of coverage versus risk, all of which equip analysts with the tools to examine the different decision factors and assist course of action development in the long-term planning and assessment process.

1. Introduction

Risk-based decision making is a growing operational and business trend that currently lacks interactive tools to aid the decision makers. The term risk is defined as the "potential for an unwanted outcome resulting from an incident, event, or occurrence, as determined by its likelihood and the associated consequences" [Com10]. Therefore, risk-based decision making can be defined as a process that collects and organizes information about different possible outcomes in an ordered structure that helps analysts make informed choices [MMGW04]. Risk-based decision making provides a framework for making decisions and helps identify the greatest risk so the decision maker can prioritize efforts in order to minimize risk and support long-term planning.

However, performing risk analysis and long-term planning is a complex and challenging analytical task, in which the decision maker must set up the problem and determine inputs, outputs, and other factors that might influence the decisions. Research in other areas has shown that individuals often make sub-optimal decisions due to cognitive limitations [SLFE11] and information overload [EM08]. Moreover, the analyst could base his/her decisions on subjective, rather than objective, perception of the risk at hand.

Therefore, we have developed several visual analytics components that can facilitate and improve the process of risk-based decision making. These components, developed through a collaborative user-centered process with the U.S. Coast Guard, use graphical depictions to assist the cognitive process of quantifying and comparing lines of evidence [LCG*12]. Our interactive components facilitate thinking, thereby improving the analyst's understanding of the data and speeding the overall decision making process. The components include feedback and exploratory abilities to examine, filter, and modify certain parameters.

During development, we followed a procedure similar to Sedlmair et al.'s [SMM12] nine-stage framework for conducting design studies. The new components were added to the framework described by Malik et al. [MMME11] because the end users have an understanding and working knowledge of the system.

The new risk-based visual analytics components being applied to visualize and compare risk include the following:

- The use of interactive graphics and choropleth maps to visualize operational risk profiles.
- A method to visualize and identify areas of high risk and compare the changes in risk priority areas over time.
- A method to spatially evaluate and distribute precomputed risk values based on the underlying distribution of cases over time.

† e-mail: {solivero|yang260|ebertd}@purdue.edu
‡ e-mail: jangy@sejong.edu

2. Related Work

In this section, we review previous works that describe the use of visual analytics in communicating risk, some existing models for risk analysis, and different tools to address risk in the maritime security domain.

In risk communication, Lipkus and Hollands [LH99] demonstrated that static images displaying risk characteristics such as risk magnitude and cumulative risk communicate the risk values more effectively than a display of numbers. Savikhin et al. [SME08] demonstrate the benefits of applying visual analytics techniques to aid users in their economic decision making. In contrast, our components provide not only visualizations, but also integrated techniques to analyze the changes of risk values both spatially and temporally.

For risk analysis and modeling, Bonafede and Marmo [BM08] demonstrate that the use of graphs can reduce search times for solutions and for identification of data. They propose four sub-plots with bar graphs and parallel coordinates to compare clients. Feather et al. [FCKM06] describe a risk based decision process with a model that takes into account requirements, risks, and mitigation strategies using bar charts and treemaps. Both papers emphasize that no single visualization technique serves all purposes and instead it is better to use a mix of several. One limitation in their systems is the lack of support of spatiotemporal data. Migut and Worring [MW10] developed a framework that integrates interactive visual exploration with machine learning techniques to support the risk assessment and decision making process. Their visualizations include scatterplots and mosaic plots as tools to build classification models.

Willems et al. [WvdWvW09] presented a geographical visualization using density estimated heatmaps to display vessel movements and support coastal surveillance systems. Pelot et al. [PP08] created a grid colored map representing vessel traffic where they model and identify vulnerable areas. Marven et al. [MCK07] analyzed Search and Rescue operations for the Canadian Coast Guard, exploring the clustering of incident areas with two different models: a Spatial and Temporal Analysis of Crime (STAC) and kernel density estimation (KDE). Abi-Zeid et al. [AZF05] developed SARPlan, a geographic decision support system for planning search and rescue missions, originally developed for aeronautical incidents. Orosz et al. [OSB*10] developed PortSec for decision-making and planning of port resources to address security needs to outside threats and hypothetical scenarios.

3. Visual Analytics in the Risk Management Process

We used the risk management process originally specified in ISO 31000:2009 [ISO09] to provide the initial principles and generic guidelines for risk management. Based on this process, we developed specific goals that our new visual analytics components should achieve:

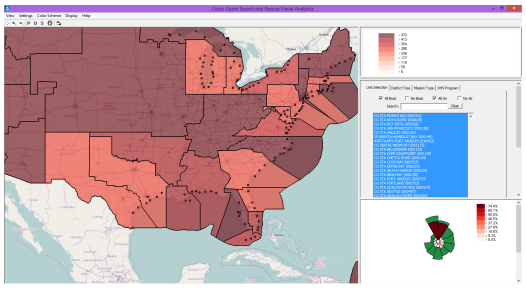


Figure 1: *View of the overall Visual Analytics System*

- Understand areas and missions driving the risk values.
- Identify risk priority areas and how they evolve over time.
- Visualize the geographical distribution of operations.
- Visualize the spatial distribution of the risk.
- Obtain details on demand about the operations.
- Provide a feedback loop if certain parameters change.

Malik et al. [MMME11] focused on the consequences of station closures, but the new additions to the system focus on Risk at the operational level. Such risk is assessed by the USCG Operational Risk Assessment Model (ORAM) [USC12]. Analysts at the Coast Guard Atlantic Area's Operations Analysis Division created this model to support mission planning and analysis of the Coast Guard's mission set. The model combines quantitative and qualitative theoretical frameworks to calculate and compare risk between the eleven Coast Guard statutory missions and geographical areas by providing the Risk Index Numbers (RIN) [USC12]. The RIN is a numerical value that characterizes and quantifies the qualities of risk. RIN values provided include both total risk and residual risk values as shown in Equation 1 [Com10].

$$\text{Total RIN} = \text{Residual RIN} + \text{Mitigated RIN} \qquad (1)$$

3.1. Operational Risk Profiles

The first step is to acquire an understanding on how the risk numbers behave for each district as well as how much risk was mitigated. Therefore, there are two main goals in visualizing the Operational Risk Profiles:

- Compare the RIN values between the districts for any given mission or combination of missions.
- Compare the RIN values between missions for any given district.

When performing total versus residual risk analysis, the ratio between the RIN values is more critical than the raw numbers; therefore, we choose a radial layout to focus on ratios and relative values since such layouts inhibit the analysts innate tendency to focus on these numerical details.

We went through several design iterations and presented different alternatives to our end users to gain feedback in terms of which design was the most effective in conveying the information and comparing the distribution of risk. A risk

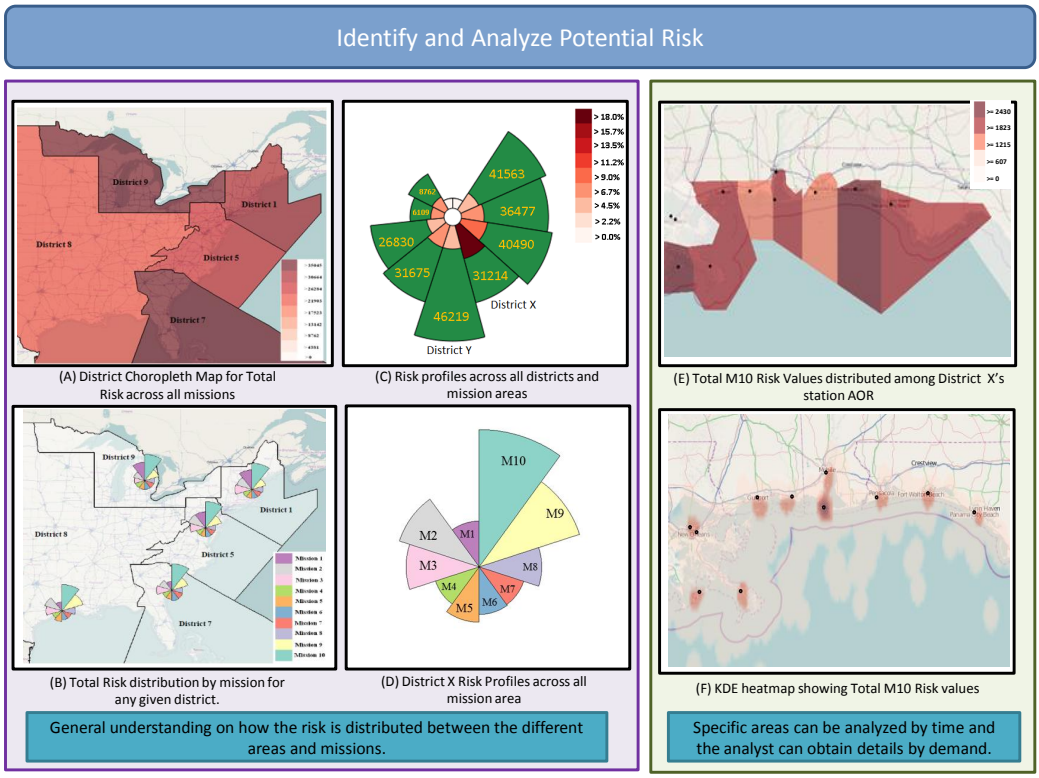


Figure 2: *General process for identifying and analyzing potential risk.*

pie graph was created with eleven fixed pie slices each representing a Coast Guard district as shown in Figure 2-C. The area of each outer pie slice is used to encode the comparison of total risk across districts, with larger pie slice corresponding to higher total risk. The area of inner pie slices represent the comparison of residual risk across districts. Each inner pie slice is also colored on a sequential red scale indicating the ratio of residual versus total risk for a given district. The choice of color (green indicates mitigated risk and red indicates residual risk) is consistent with the Coast Guard's Green-Amber-Red model. We allow interactive filtering by missions to analyze and compare the spatial distribution of risk across districts for any given mission or combination.

3.2. Risk Visualization Using Heatmaps

Next, we need to analyze risk priority areas and how they evolve over time. To quickly identify hotspots, a modified variable kernel density estimation technique (KDE) is employed on the map. Risk at the strategic level is not assigned to a specific unit or station, instead the analyst is able to observe areas with a high density of incidents independent of station location. The heatmap can display the RIN values for total, residual, and mitigated risk. The analyst can switch between the total risk and the residual risk to find hotspots where the risk has not been mitigated and examine the incident details in these zones. Analyzing the incident helps the analyst develop new strategies and courses of action to mitigate the risk.

3.3. Risk Distribution using Choropleth Maps

We utilize choropleth maps in two different ways to help visualize risk. The first option is to visualize any of the Risk values for any given mission or combination of missions by district (Figure 2-A), providing an effective way to present and share the information about risk levels within the U.S.

The second use of choropleth maps (Figure 2-E) highlights the risk distribution of the RIN values per district. During the process it is useful to visualize risk at the station level by using each individual station's Area of Responsibility (AOR). Certain mission's RIN numbers are computed at a district level rather than the station level. Therefore, in order to distribute the RIN values across the stations' AORs, we analyze the underlying incident distributions for a given time period. We use the incidents distribution as a basis to assign risk values across stations given the pre-computed total RIN values by district. The mathematical formula used to compute distributed RIN value for a particular station X that belongs to district Y is:

$$\text{station X RIN} = \frac{\text{Incidents in X}}{\text{Incidents } \forall \text{ stations in Y}} \times \text{district Y RIN} \quad (2)$$

The risk distribution choropleth map provides an easy way to visualize the variations in risk values for individual station's AOR and help identify stations that will potentially require allocation of more resources.

3.4. Visual Analytics System

The overall system provides multiple linked windows and advanced filtering techniques to perform spatio-temporal analysis on the risk data as shown in Figure 1. The system allows the user to visualize historical Coast Guard Data, such as the number and location of incidents that occurred during a certain period of time. It can analyze incidents occurring on specific date ranges to explore seasonal trends and it can filter incidents relevant to the analyst's hypothesis. The addition of the new components enables the Coast Guard analyst to perform risk-based analysis of the operation as well as long term planning by providing new visualization along with feedback loops that control resource allocation.

4. Case Study: Identify and Analyze Potential Risks

To illustrate the use of our system, we present an example use scenario using notional data. In decision making, several questions will drive the analyst in developing the planning strategy: What risks exist in the region and where they are distributed? Where are our resources allocated? What constraints exist in the system that will require a prioritization of resource use?

In a resource constrained environment, we want to use resources in the mission area that provides the greatest return on investment (large amount of total risk but very little residual risk). The first step in the risk management process is to identify potential risks; therefore the analyst begins by looking at the operational risk profile and the district risk choropleth map to observe the risk values at the district level across all mission areas.

Figure 2-C displays the total and residual risk and the ratio between them for all the districts across all mission areas. In this case, we can observe that although District Y has the largest total risk values, it mitigated most of the risk effectively. On the other hand, District X shows less total risk, but the amount of residual risk as well as residual to total risk ratio is the highest as encoded by the darkest red shade. District X can be seen as more problematic than District Y; thus, the analyst will focus more attention on analyzing this particular district. This visualization provides a starting point in understanding how risk is distributed among the different districts and focusing on districts with high risk concentration.

After identifying that District X has the greatest residual risk and the highest risk concentration, the next step is determining the key drivers of risk within a district. This leads the analyst to leverage other components of the risk visual analytics tool to specifically evaluate District X. For instance, the analyst can examine the distribution of risk across different missions in District X as shown in Figure 2-D to identify which mission type has the greatest risk in this district. The analyst can observe that most of the operational risk emerges from one of the missions, in this case M10.

New questions emerge at this stage: Are there several big events that drive the risk, or are there many small events with smaller consequences accumulated to affect the operation? So now we examine the spatial distribution of M10 risk within District X to analyze specific areas of high residual risk. Depending on the data quality regarding spatial location, the analyst has two options for drilling down into specific areas within District X. The first option is to use the risk heatmap described in Section 3.2 to locate risk priority areas, as seen in Figure 2-F. If the spatial location is not available, then we re-distribute the risk to station AORs as described in Section 3.3 and as seen in Figure 2-E.

5. Domain Expert Feedback

The prototype components went through an iterative design refinement process with the collaboration of four Coast Guard personnel: an operation research analyst, a former Coast Guard officer, one in-field officer, and a high level officer. Informal feedback is given below:

"These components aid the analyst in answering the questions that come from developing the planning strategy, often with a speed that was previously unattainable with the Coast Guard's usual brute force processing of thousands of lines of data to calculate summary statistics."

"This system provides a risk informed process for building a defensible planning baseline for the long-term planning process. Understanding the risk profiles provides analytic justification for resource use, and can aid in demonstrating effective application of resource use based on risk."

6. Conclusions

We have demonstrated how our interactive visual analytics components can facilitate the risk management process and evaluate courses of action. Within the maritime context, our interactive visual analytics environment utilizes KDE heatmaps to help identify risk priority areas, multiple designs to visualize risk profiles, a risk distribution choropleth map to visualize the spatial distribution of pre-computed risk values, and the coverage map overlaid with risk distribution for analysis of coverage capability/efficiency as well as potential need for resource reallocation or assets upgrade. Finally, we included a case study that examines the efficiency of Coast Guard operations and provides useful visual reference that can communicate recommendations based on risk management. The described risk-based decision making process serves as a blueprint for future systems dealing with risk values and resource planning.

Acknowledgment

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003. Jang's work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170).

References

[AZF05] ABI-ZEID I., FROST J. R.: Sarplan: A decision support system for canadian search and rescue operations. *European Journal of Operational Research 162*, 3 (2005), 630 – 653. Decision-Aid to Improve Organizational Performance. 2

[BM08] BONAFEDE C., MARMO R.: Operational Risk Visualization. *Science* (2008), 100–103. 2

[Com10] COMMITTEE R. S.: *DHS Risk Lexicon*. Homeland Security, 2010. 1, 2

[EM08] EPPLER M., MENGIS J.: The concept of information overload - a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004). In *Kommunikationsmanagement im Wandel*, Meckel M., Schmid B., (Eds.). Gabler, 2008, pp. 271–305. 1

[FCKM06] FEATHER M., CORNFORD S., KIPER J., MENZIES T.: Experiences using Visualization Techniques to Present Requirements, Risks to Them, and Options for Risk Mitigation. *2006 First International Workshop on Requirements Engineering Visualization (REV'06 - RE'06 Workshop)* (Aug. 2006), 10–10. 2

[ISO09] ISO 31000, Risk Management Principles and Guidelines, Geneva : International Standards Organisation, 2009. 2

[LCG*12] LINKOV I., CORMIER S., GOLD J., SATTERSTROM F. K., BRIDGES T.: Using our brains to develop better policy. *Risk Analysis 32*, 3 (2012), 374–380. 1

[LH99] LIPKUS I. M., HOLLANDS J. G.: The visual communication of risk. *Journal of the National Cancer Institute. Monographs 27701*, 25 (Jan. 1999), 149–63. 2

[MCK07] MARVEN C., CANESSA R., KELLER P.: Exploratory spatial data analysis to support maritime search and rescue planning. *Geomatics Solutions for Disaster Management* (2007), 271–288. 2

[MMGW04] MACESKER B., MYERS J., GUTHRIE V., WALKER D.: *Quick Reference Guide to Risk Based Decision Making (RBDM): A Step by Step Example of the RBDM Process in the Field*. EQE International, Inc., an ABS Group Company Knoxville, Tennessee, 2004. 1

[MMME11] MALIK A., MACIEJEWSKI R., MAULE B., EBERT D.: A visual analytics process for maritime resource allocation and risk assessment. In *Visual Analytics Science and Technology (VAST), IEEE Conference on* (oct. 2011), pp. 221 –230. 1, 2

[MW10] MIGUT M., WORRING M.: Visual exploration of classification models for risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (October 2010), pp. 11–18. 2

[OSB*10] OROSZ M., SOUTHWELL C., BARRETT A., CHEN J., IOANNOU P., ABADI A., MAYA I.: Portsec: A port security risk analysis and resource allocation system. In *Technologies for Homeland Security (HST), 2010 IEEE International Conference on* (Nov. 2010), pp. 264 –269. 2

[PP08] PELOT R., PLUMMER L.: Spatial analysis of traffic and risks in the coastal zone. *Journal of Coastal Conservation 11* (2008), 201–207. 2

[SLFE11] SAVIKHIN A., LAM H., FISHER B., EBERT D.: An experimental study of financial portfolio selection with visual analytics for decision support. In *System Sciences (HICSS), 44th Hawaii International Conference on* (2011), IEEE, pp. 1–10. 1

[SME08] SAVIKHIN A., MACIEJEWSKI R., EBERT D.: Applied visual analytics for economic decision-making. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)* (October 2008), pp. 107–114. 2

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 18*, 12 (2012), 2431–2440. 1

[USC12] USCG: *Joint, CG Atlantic Area and CG Pacific Area Operational Risk Assessment Model (ORAM), Executive Summary*. Unpublished Technical Document, 2012. 2

[WvdWvW09] WILLEMS N., VAN DE WETERING H., VAN WIJK J. J.: Visualization of vessel movements. *Comput. Graph. Forum 28*, 3 (2009), 959–966. 2

Public Behavior Response Analysis in Disaster Events Utilizing Visual Analytics of Microblog Data

Junghoon Chae[a], Dennis Thom[b], Yun Jang[c,*], SungYe Kim[d], Thomas Ertl[b], David S. Ebert[a]

[a]*Purdue University, USA*
[b]*University of Stuttgart, Germany*
[c]*Sejong University, South Korea*
[d]*Intel Corporation, USA*

Abstract

Analysis of public behavior plays an important role in crisis management, disaster response, and evacuation planning. Unfortunately, collecting relevant data can be costly and finding meaningful information for analysis is challenging. A growing number of Location-based Social Network services provides time-stamped, geo-located data that opens new opportunities and solutions to a wide range of challenges. Such spatiotemporal data has substantial potential to increase situational awareness of local events and improve both planning and investigation. However, the large volume of unstructured social media data hinders exploration and examination. To analyze such social media data, our system provides the analysts with an interactive visual spatiotemporal analysis and spatial decision support environment that assists in evacuation planning and disaster management. We demonstrate how to improve investigation by analyzing the extracted public behavior responses from social media before, during and after natural disasters, such as hurricanes and tornadoes.

Keywords:
Visual Analytics, Social Media Analysis, Disaster Management

1. Introduction

For emergency and disaster management, analysis of public behavior, such as how people prepare and respond to disasters, is important for evacuation planning. As social media has played a pervasive role in the way people think, act, and react to the world (more than 40 million Americans use social media Web sites multiple times a day [1]), social media is changing the way people communicate not only in their daily lives, but also during abnormal events, such as natural disasters. In emergency situations, people even seek social confirmation before acting in response to a situation, where they interact with others to confirm information and develop a better informed view of the risk [2]. A study commissioned by the American Red Cross, found that roughly half of the respondents would mention emergencies and events on their social media channels, and more than two-thirds agree that response agencies should regularly monitor postings on their websites [3]. Moreover, a growing number of people are using Location-based Social Network services, such as microblogs, where they create time-stamped, geo-located data and share this information about their immediate surroundings using smart phones with GPS. Such spatiotemporal data has great potential for enhancing situational awareness during crisis situations and providing insight into the evolving event, the public response, and potential courses of action.

For public behavior analysis in disasters, however, finding meaningful information from social media is challenging. It is almost impossible to perform a straightforward qualitative analysis of the data, since the volume of the data exceeds the boundaries of human evaluation capabilities and normal computing performance. Even though we could extract certain information from the data, it is not always easy to determine whether the analysis result of the extracted information is meaningful and helpful. Thus, there is a need for advanced tools to handle such big data and aid in examining the results in order to understand situations and glean investigative insights. Given the incomplete, complex, context-dependent information, a human in this analysis and decision-making loop is crucial. Therefore, a visual analytics approach offers great potential through interactive, scalable, and verifiable techniques, helping analysts to extract, isolate, and examine the results interactively. In this paper, we present an interactive visual analytics approach for spatiotemporal microblog data analysis to improve emergency management, disaster preparedness, and evacuation planning. We demonstrate the ability to identify spatiotemporal differences in patterns between emergency and normal situations, and analyze spatial relationships among spatial distributions of microblog users, locations of multiple types of infrastructure, and severe weather conditions. Furthermore, we show how both spatiotemporal microblog and disaster event data can help the analysts to understand and examine emergent situations, and evaluate courses of action.

This study is performed using Twitter messages called

*Corresponding author

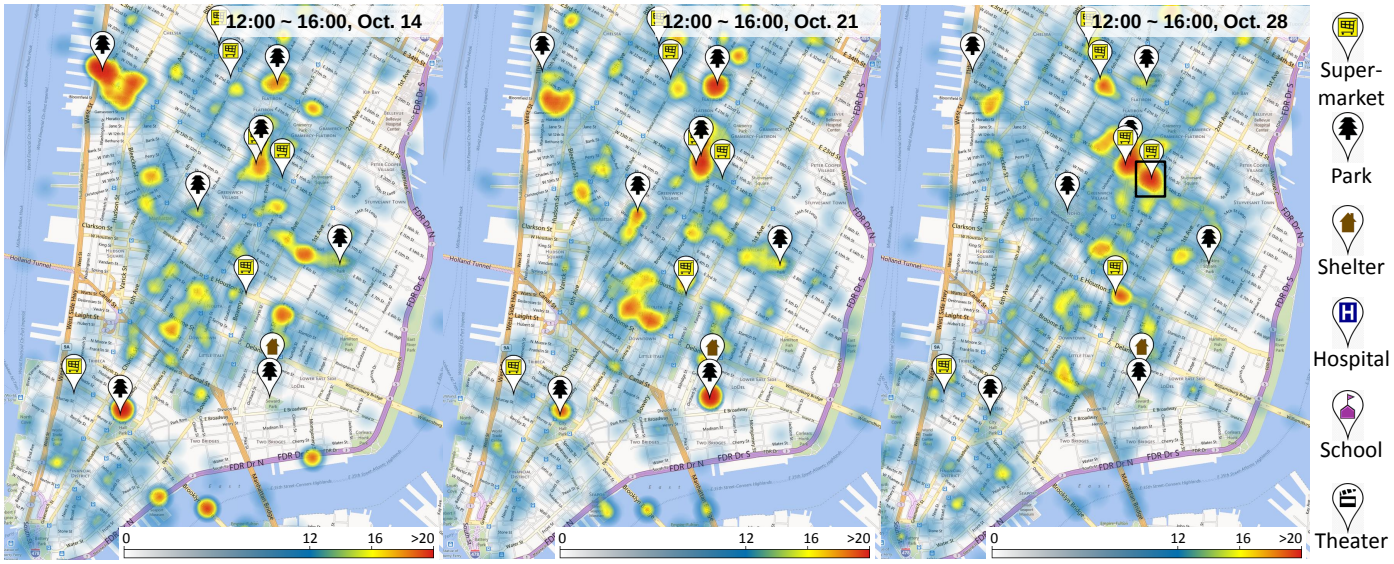


Figure 1: Spatial user-based Tweet distribution in the Manhattan area in New York City during four hours right after the evacuation order (from 12:00 PM to 4:00 PM on October 28th, 2012 (Right)). Previous distribution of Tweets on 14th (Left) and 21st (Center).

Tweets, as Twitter has been the most popular microblog service in the United States. In this paper, we extend our previous work [4] with additional features of our system and examine their capabilities with several expanded examples in Section 4.2. We also add a discussion section for comparisons and analysis of the case studies.

Our system evaluates visual analytics of spatiotemporal distribution of Tweets to identify public behavior patterns during natural disasters. The main features of our approach are as follows:

- **Spatial analysis and decision support:** The system provides effective analysis for exploring and examining the spatial distribution of Twitter users and supporting spatial decision-making using a large volume of geo-located Tweets and multiple types of supplementary information during specific time periods (i.e., disaster events).
- **Temporal pattern analysis:** Our visualization system enables the analysts to analyze the temporal distribution of the number of Twitter users posting Tweets in a given location and time.
- **Spatiotemporal visualization:** We provide a visualization that allows the analysts to simultaneously analyze both aspects: space and time in a single view.

We first review previous work in Section 2 and describe our interactive analysis system in Section 3. We present analysis results for two natural disaster events in Section 4 and discussion in Section 5. Finally conclusion and future directions are presented in Section 6.

2. Related Work

In recent years social media data has become a popular topic in a range of application domains. Several researchers have proposed and presented systems for social media analysis and important studies covering the use of social media during crisis events have been conducted. Most recent analysis environments for crisis-related social media exploration and visualization are from MacEachren et al. [5], Marcus et al. [6], and Thom et al. [7]. Their systems combine traditional spatial and geographic visualizations with means for automated location discovery, trend and outlier search, anomaly and event discovery, large scale text aggregation and highly interactive geovisual exploration. Approaches putting less focus on visualizations and more on fully automated data mining mechanisms have been proposed by Sakaki et al. [8] that use Kalman and Particle Filters to detect the location of earthquakes and typhoons based on Twitter. Various techniques for spatiotemporal data analysis and anomaly detection using visualization or machine learning techniques have been proposed by Andrienko et al. [9], Lee and Sumiya [10], and Pozdnoukhov and Kaiser [11]. Twitcident from Abel et al. [12] provides a web-based framework to search and filter crisis-related Tweets. Using the Netherlands emergency broadcast system, Twitcident automatically reacts on reported incidents and collects related information from Twitter based on semantic enrichment. In all these system the focus is primarily on individual messages and aggregated message volumes and how insight can be generated by understanding their content. In contrast, our system investigates a more user focused approach that tries to identify the whereabouts and movements of people in order to understand mass behavior.

Researchers have also examined the usage of Twitter during incidents and disasters. Terpstra et al. [13] investigate more than 90k Twitter messages that were sent during and after a storm hit the Belgium *Pukkelpop* musicfestival in 2011. They categorize Tweets into warnings about the severe weather conditions, rumors and self organization of relief measures. They show that valuable information for crisis response and decision support can be gathered from the messages. Vieweg et

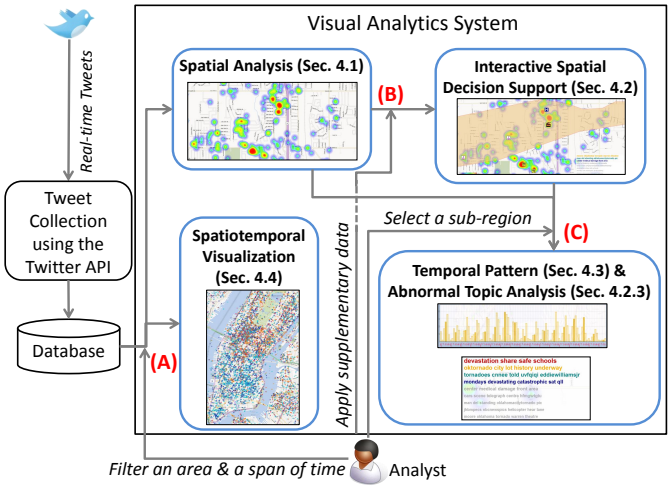


Figure 2: Overview of our interactive analysis scheme for public behavior analysis using social media data.

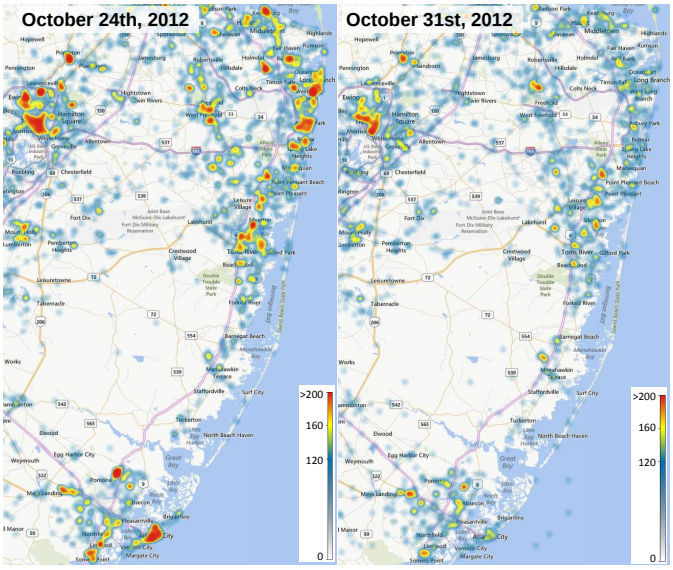


Figure 3: Twitter user distribution on the eastern coast area in New Jersey, after the hurricane passed over the area on October 31st (Right). Previous distribution on October 24th is shown on the Left.

al. [14] investigate the differences in reaction to different crisis events. For their study they investigate eyewitness reports in Twitter from people that were affected by Oklahoma Grassfires in April 2009 and Red River Floods in March and April 2009. Their research also demonstrates the high value that the extraction of meaningful comments from crisis-related communication can have to generate insights. Furthermore, Heverin et al. [15] demonstrate that Twitter can also be a useful source of information for smaller events as they investigate the reaction to a shooting of four police officers and the subsequent search for the suspect that took place in the Seattle-Tacoma area. Based on the collection and categorization of 6000 messages they are able to show that citizens use the service to communicate and seek information related to the incident.

In this paper we also present a case study on crisis-related information gathered from Twitter data. However, in contrast to the discussed studies that harvest information directly out of the content of the messages, our method is primarily based on observing movement patterns and identifying local hotspots in order to learn about the effects of the crisis and the performance of evacuation measures.

3. Problem Statement and Interactive Analysis Process Design

Analysis of public behavior, such as how people prepare and respond to disasters, plays an important role in crisis management, disaster response, and evacuation planning. Recently, social media becomes popular and people utilize it for communications not only in their daily lives, but also in abnormal disastrous situations. Thus, Location-based Social Networks services offer a new opportunity for enhancing situational awareness during disaster events. Unfortunately, collecting relevant data can be costly and finding meaningful information from the huge volume of social media data is very challenging. Therefore, there is a need for an advanced tool to analyze such massive ("big") streaming data and aid in examining the analysis results to better understand situations more efficiently.

Our proposed visual analytics approach provides multiple analysis methods: spatial analysis, spatial decision support, temporal pattern analysis, abnormal topic analysis, and interactive spatiotemporal visualization as shown in Figure 2. In our system, all methods are tightly integrated based on a user-centered design in order to enhance the ability to analyze huge social media data (Figure 2 (A, B, C)). Our Tweet collection component obtains real-time Tweets using the Twitter API—to collect about 2.2 million geotagged Tweets within the United States per day. In general for spatial analysis, the required accuracy of the geocoordinate depends upon the required level of location granularity. The data, however, is generated by very reliable GPS and software. We can be reasonably certain about the data accuracy as illustrated in [16]. For the temporal accuracy of Tweets, we use the time when each Tweet is created. Therefore, it is highly accurate if the time setting of the device posting a Tweet is correct. This large volume of data is stored in our database in order to maintain and track the history of the Twitter stream. Our system allows the analysts to query Tweets with a specific area and time span condition (Figure 2 (A)). The initially selected spatiotemporal context of Tweets can be represented by two different analytics components: spatial analysis and spatiotemporal visualization. Spatial analysis allows the analysts to examine the overall distribution of Twitter users and discover hotspots where relatively more Twitter users post Tweets. The analysts are able to add supplementary information (infrastructure locations, tornado paths) on top of current information representing outcomes in order to better understand events and increase situational awareness (Figure 2 (B)). Furthermore, the analysts can select a sub-region within the initial area, so that he can analyze the temporal patterns of the number of Twitter users and extract abnormal topics from the text messages in the selected region (Figure 2 (C)). In addition, our in-

teractive spatiotemporal visual analytics provides a single view representation for the analysis of both aspects: spatial and temporal characteristics of Tweets at the same time.

4. Spatiotemporal Analysis

In this work, we present a visual analytics approach to handle the vast amount of microblog data such as Twitter messages, provide interactive spatiotemporal analysis, and enable the use of multiple types of supplementary spatial infrastructure information for spatial decision support. Analysts select an initial spatiotemporal context of Tweets to be represented in the visualization to serve as a basis for analysis. They can also perform the interactive spatiotemporal queries that load the relevant datasets from a larger database.

4.1. Spatial Analysis

As mentioned in Section 1, social media embedding geolocation information into the data is extremely useful in analyzing location-based public behaviors. Such spatial analysis, therefore, is important in order to manage and prepare plans for disaster and emergency situations.

In late October in 2012, a massive hurricane, Sandy, devastated Northeastern United States [17]. Due to the severeness of the hurricane, on October 28th in 2012, the New York City Authorities ordered residents to leave some low-lying areas—the mandatory evacuation zones (red color) are shown in Figure 9 (Right). We investigate an area of Manhattan, since the area is the most populated and severely damaged. Through the map view in our system, analysts navigate to the Manhattan area in New York City and filter Tweets posted within the area. Initially we tried to reveal public movement flows during the disaster event, but the movement patterns were too complicated to find meaningful flows due to movement randomness and the visual clutter of the flows. Then, we examined the spatial distribution of the users for specific time frames. Based on our experiments, a geospatial heatmap was useful for an overview of the spatial distribution and for trend approximation. We utilize a divergent color scheme to generate the heatmap, where saturated colors are used for the data distribution to avoid any confusion from the color scheme from the desaturated colormap of the background map. Analysts can specify a threshold range to emphasize hotspots, where the upper bound is mapped to a red color and the lower bound to a yellow color. Additionally, the blue color is mapped by the analysts to the value of the overall distribution of Twitter users. In Figure 1, we show three heatmaps of spatial user-based Tweet distribution from 12:00 PM to 4:00 PM on October 14th (Left), 21st (Center), and 28th (Right). In this work, we use the number of Twitter users instead of the number of Tweets for the heatmap generations to properly reflect the flow of evacuation unbiased by personal Tweet activity or behavior of individual users, since some enthusiastic Twitter users generate a large number of Tweets at the same location during a short time period (more than 20 Tweets per hour). The heatmaps in Figure 1 (Left and Center) represent normal situations of Twitter user distribution in the Manhattan area, and the heatmap (Right) shows the situation right after the evacuation order that was announced at 10:30 AM on October 28th, 2012. This standard heatmap visualization allows analysts to explore the spatial pattern of Twitter users for any specified time period. In Section 4.2, we will provide further analysis for the spatial decision support.

Hurricane Sandy damaged not only New York City, but also the entire eastern coast area of New Jersey. Most cities in the area also announced evacuation orders on October 28th, 2012. The distribution of Twitter users in the area from Atlantic City to the upper eastern shore area for two different dates are shown in Figure 3. The heatmaps in Figure 3 (Left) represent the previous normal situation of Twitter user distribution on October 24th and the heatmap (Right) shows the post distribution after Sandy passed over the area on October 31st. As shown in the result, many hotspots are gone or diminished. This situation shows that the number of Twitter users had significantly decreased after the hurricane damaged the area. In fact, a huge number of homes were damaged or destroyed and a couple of million households lost power because of Hurricane Sandy [18]. In disaster management this type of visualization can support analysts estimating which areas were highly damaged and even which areas still need reconstruction.

4.2. Spatial Decision Support

In Section 4.1, we introduced our spatial analysis to explore the Twitter user distribution. In addition to the analysis, our system allows the analysts to utilize supplementary information in order to support understanding of the situations and decision-making in disaster management. The spatial characteristics together with heterogeneous information can assist in disaster management and migrating hazards where the problems have spatial components [19]. The supplementary information can be various types of infrastructures (i.e., school, park, supermarket, and shelter), as well as spatial information of disaster events (i.e., hurricane path and damage area of a tornado). In this section, we describe how our system supports spatial decision-making by correlating such spatial information with location-based microblog data.

4.2.1. Infrastructure Data

During a natural disaster event, such as Hurricane Sandy, analysts would assume that many people might want to go to the supermarket before staying or evacuating, but they would need supporting evidence before making appropriate decisions and plans. With our system support, the analysts can simply overlay the locations of large supermarkets on the heatmap of the Twitter user distribution. The infrastructure locations are indicated by standard symbols [20] as shown on the right side of Figure 1. A relatively large number of people immediately went to supermarkets nearby the evacuation area, instead of the emergency shelter as shown in Figure 1 (Right). However, October 28th was Sunday and many people generally would go for grocery shopping on Saturday or Sunday; therefore, the analysts might need to verify whether the heatmap shown in the figure is a normal periodic situation. The analysts can investigate new

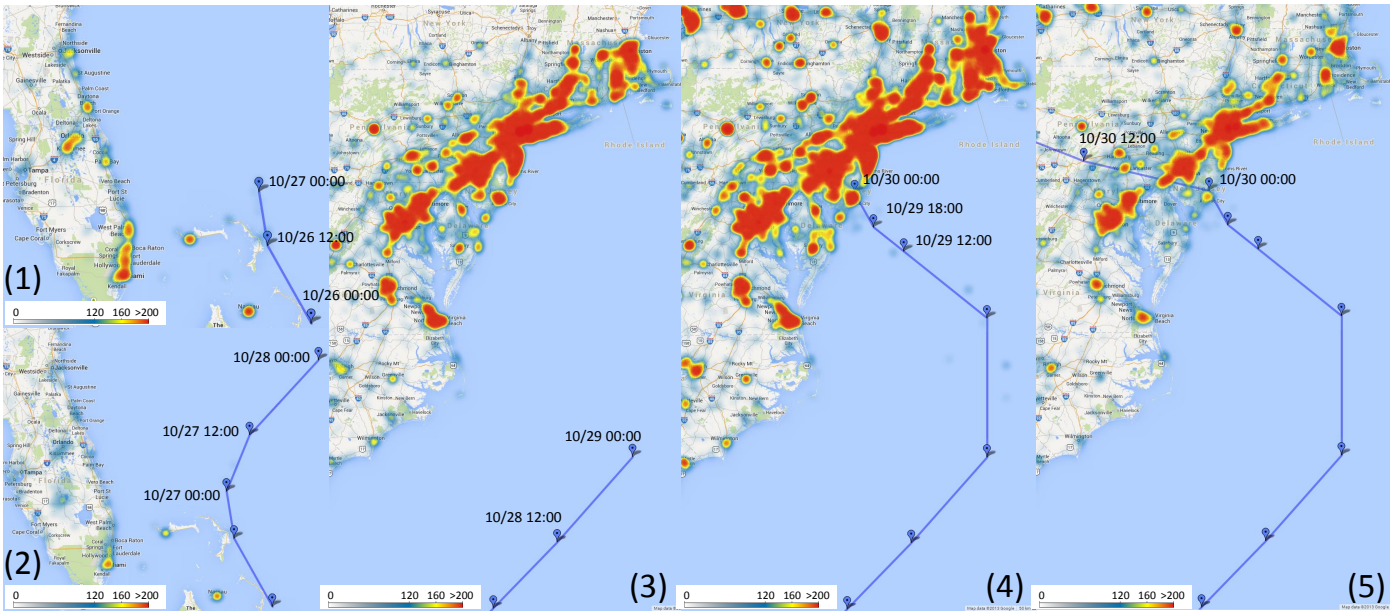


Figure 4: Distribution of Twitter users of each consecutive date (Oct. 26 ~ 30, 2012), who post hurricane related Tweets on the southeastern (1 and 2) and northeastern coast (3, 4, and 5) area of the United States. We can see the variance of Twitter user reactions along the track of the hurricane center locations.

Twitter user distributions for different time frames by simply manipulating the time context. In Figure 1 (Left and Center), we show two distributions for one and two weeks before the disaster period respectively. Here, we see that the hotspot locations are very different from the ones for October 28th shown in Figure 1 (Right). For further analysis, we can explore another popular Sunday location—large parks—by superimposing the locations on each heatmap. As shown in Figure 1 (Left and Center), many hotspots overlap with the park areas in normal situations. Therefore, we can conclude that the situation on October 28th is an unusual non-periodic pattern.

4.2.2. Disaster Event Data

In Section 4.2.1, we explained how the infrastructure data help the analysts to understand and examine the emergent situations. During severe weather conditions, people tend to be sensitive to the dynamic variance of the weather conditions. Relationship analysis, therefore, between the public responses and the spatiotemporal pattern of the severe weather is important. Our system overlays geographic information of disaster events, for example, center positions and tracks of a hurricane, and damaged areas by a tornado, in order to provide further analysis. Two case studies are presented as follows:

Track of Hurricane: Figure 4 (1) and (2) show the southeastern coast areas of the United States, whereas, Figure 4 (3), (4), and (5) show the northeastern coast areas. In the figures the distributions of Twitter users for each consecutive date, from October 26th to 30th, 2012, are presented using the heatmap visualizations. We use the number of Twitter users who posted Twitter messages containing one of the following keywords: *hurricane, storm*, and *sandy* in order to analyze Tweets that are highly related to Hurricane Sandy. Note that Hurricane Sandy reached the southeastern Florida coast on October 26th and passed, then, over the northeastern coast on October 30th, 2012 [17]. As shown in Figure 4, our system is able to overlay the track of the hurricane on the map. The blue pins and the blue lines represent the center locations of the hurricane and its path respectively.

Twitter users also actively respond to the severe weather conditions. In Figure 4, we indicate that the distribution pattern of Twitter users had dynamically varied along the track of the hurricane center locations. When Sandy moved to the southeastern coast on October 26th, there were bursts on eastern Florida's coast (Figure 4 (1)). Next day, the bursts disappeared, because Sandy moved towards the northeast away from the east coast of United States (Figure 4 (2)). Sandy kept moving towards a few hundred miles southeast of North Carolina on October 28th (Figure 4 (3)). In the next day, the hurricane's track bent towards the north and the hurricane made landfall at night in the northeast of Atlantic City (Figure 4 (4)). Throughout the days, Twitter users were actively reacting to Hurricane Sandy' arrival in a wide range of areas. After the landfall, the storm turned toward the northwest and was gradually weakened. The big outbreaks were diminished on October 30th as shown in Figure 4 (5). As shown in the figures, we can see how Twitter users reacted according to the spatiotemporal pattern of the severe weather conditions in the social media domain.

Damage Area from a Tornado: An extremely strong Tornado passed through the city of Moore in southern metropolitan Oklahoma City [21] in the afternoon on May 20th, 2013. The larger than one-mile-wide tornado damaged the city with a wind speed of more than 200 mph. Figure 5 shows the damaged part of the city. The tornado entered the area at about 3:16 PM and exited the area after about 10 minutes. We visualize the distribution of Twitter users on the map during 24 hours, from May 20th 4:00 PM to 21st 4:00 PM. We also over-

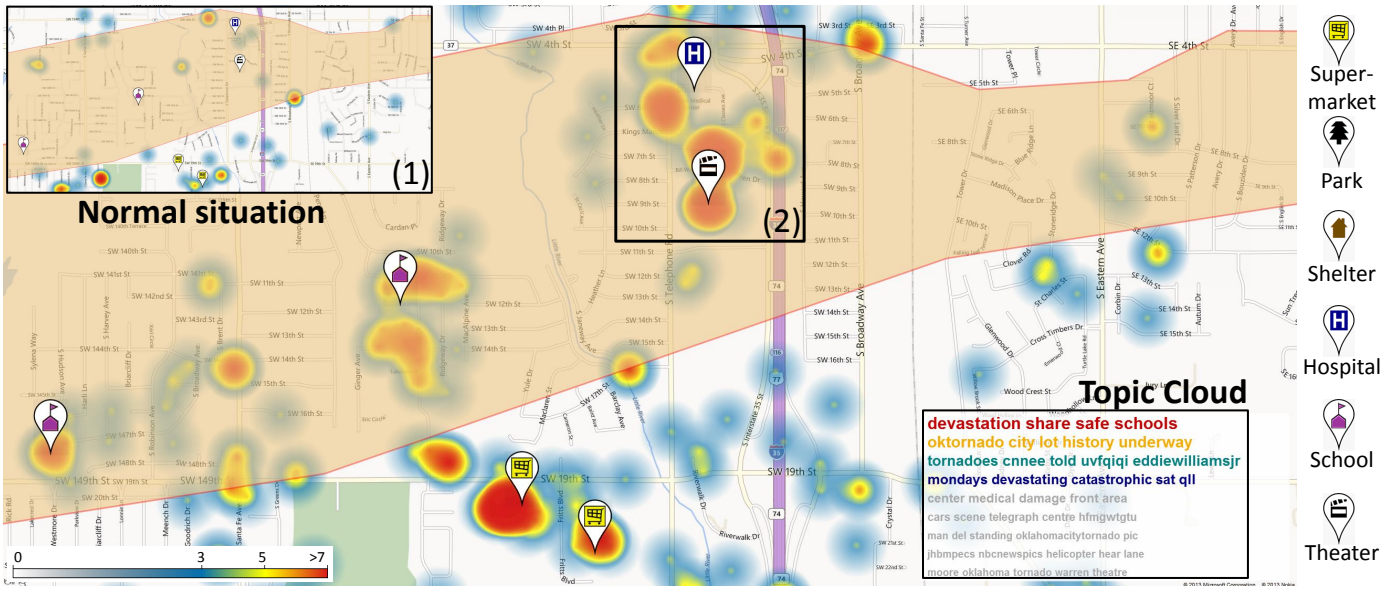


Figure 5: Spatial pattern of Twitter users during 24 hours in the city of Moore after damages from a strong tornado. Relatively many people moved to severely damaged areas after the disaster. This situation is much different from the previous normal situation (1). We selected a specific region (2) that includes severely damaged areas in order to extract topics (3) from Tweets within the selected area.

lay an approximate extent of tornado damage (transparent orange color) and locations of multiple infrastructures, such as schools, hospitals, and supermarkets, on the map view. Since the tornado suddenly happened and disappeared, we were not able to find significantly abnormal patterns before and during the event. After the disaster event, however, many Twitter users moved toward some specific areas: two elementary schools, a medical center, a theater, and two large supermarkets. The two elementary schools, the medical center, and the theater were located within the highly damaged area and they were severely destroyed. Also many people were hurt and died in these infrastructures. The increased number of Twitter users was probably due to the fact that many people went to these places in order to rescue the victims [22]. Moreover, people might have gone to supermarkets to obtain indispensable things. In Figure 5 (1), the heatmap shows a normal situation of Twitter user distribution in the same area. The distribution is very different from the situation after the tornado hit the area. This example demonstrates how our visual analytics system enables the analysts to analyze public responses using spatial disaster data and infrastructure data for disaster management.

4.2.3. Abnormal Topic Analysis

Our system also provides analysts with abnormal topic examination within the microblog data. Each Twitter message provides not only spatiotemporal properties, but also textual contents. The text messages are also important to understand and examine the emergent situations. Our system allows the analysts to extract major topics from many Tweets posted within a specific area using Latent Dirichlet Allocation (LDA) [23]. We also employ, then, a Seasonal-Trend Decomposition procedure based on Loess smoothing (STL) [24] to identify unusual topics within the selected area. For each extracted topic of the LDA topic modeling, our algorithm retrieves messages associated with the topic and then generates a time series consisting of daily message counts from their timestamps. The time series can be considered as the sum of three components: a trend component, a seasonal component, and a remainder. Under normal conditions, the remainder will be identically distributed Gaussian white noise, while a large value of the remainder indicates substantial variation in the time series. Thus, we can utilize the remainder values to implement control chart methods detecting anomalous outliers within the topic time series. We have chosen to utilize a seven day moving average of the remainder values to calculate the z-scores. Note that we use the z-score as the abnormality score in this work. If the z-score is higher than 2, events can be considered as abnormal within a 95% confidence interval. The details of these techniques are described in the previous work [25]. We select a sub area in Figure 5 (2) that includes severely damaged areas: the selected region (black rectangle) on the map. The extracted topics, which are ordered based on their abnormalities, are displayed as Topic Clouds at the bottom-right corner (Figure 5 (3)) on the map. The topic cloud is enlarged and shown in Figure 6. In this case study, most topics are related to the disaster event. However, the last topic—*moore, oklahoma, tornado, warren, theatre*, has a relatively low abnormality although they seem related to the disaster event, because tornadoes frequently occur in the area. Figure 7 shows an abnormality graph for the first topic in Figure 6. The abnormality score for the topic had significantly increased when the tornado hit the region on May 20th (Marked region). As shown in Figure 7, the abnormality score (6.75) is much higher than the average abnormality score(0.42); therefore, the analysis of the microblog data provides a statistically significant difference during this severe weather condition.

Figure 6: Topic cloud: Topics from Tweets within the selected area in Figure 5 (2) are ordered by their abnormality scores.

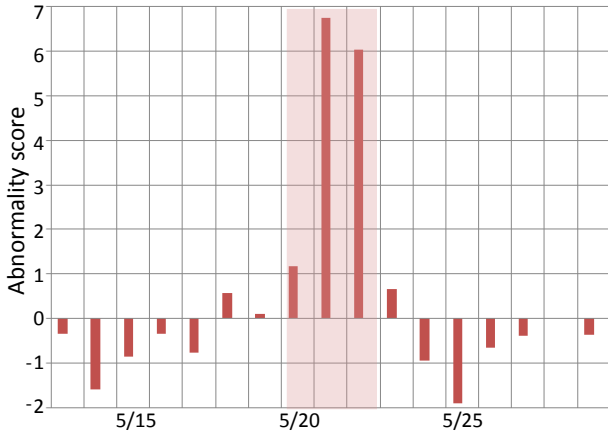


Figure 7: Abnormality of the first topic in Figure 6. The abnormality score of the topic had significantly increased when the tornado hit the region on May 20th (Marked region).

4.3. Temporal Pattern Analysis

In the previous sections, we presented the spatial analysis of social media and spatial decision support. In this section, we demonstrate analysis of the relationships between the temporal patterns of the number of Twitter users and certain public situational behaviors: how many people go where and how different is it from previous situations? Analysis of temporal trends and relationships between data values across space and time provides underlying insights and improves situational awareness [26, 27].

After selecting the initial spatiotemporal context of Tweets as a basis for the analysis, the analysts can explore the temporal patterns of the number of Twitter users who posted Tweets within the spatial boundary using the bar chart as shown in Figure 8. The values of each bar are the number of users in four hour intervals and represent data two weeks before and after the selected date. Once a mouse cursor hovers over one of the bars in the graph, every bar that corresponds to that time period, is highlighted in dark yellow color as shown in Figure 8. As previously mentioned, the heatmap in the figure shows the Twitter user density distribution from 12:00 PM to 4:00 PM on October 28th, right after the announcement of the evacuation order. We select a hotspot that includes one of the supermarket locations: the selected region (black rectangle) on the map in Figure 1 (Right). We can indicate that the number of Twitter users (red rectangle in Figure 8) in the corresponding time period is higher than for the same time period from other dates (October 14th, 21st and November 4th, 5th) by 35% more from the average. Moreover, there is another interesting finding—the number of people during each of the following time frame (4:00 ~ 8:00 PM) on the dates from the previous weeks are higher than the number of people in the selected time frame. This is because many shoppers were lining up at stores and emptied the shelves to prepare for Hurricane Sandy. Some actual Twitter messages posted in the area are following: *'The line at Trader Joes is unbelievable ...'* and *'There is amazing line here ...'*. Furthermore, since October 29th, the number of people has significantly decreased because most residents left the area before the arrival of the hurricane. The increase in the number of people after one week reflects that some people came back to the area.

4.4. Spatiotemporal Visualization

There is abundant research published on the topic of spatiotemporal data visualization. Still, exploration of time-referenced geographic data is still a challenging issue [28]. We introduce a modest visualization that enables analysts to analyze both aspects: space and time in a single view. Each Tweet is independent and contains multiple properties, such as location, time, the number of re-Tweet, etc. In this study, therefore, we utilize a glyph-based visualization to depict both location and time aspects of the independent data record using two visual features. As shown in Figure 9 (Left), each hexagon corresponding to a Tweet represents the spatial and temporal information where the center of each hexagon is the location of each Tweet and the color represents its posting time. In other words, space and time properties are encoded in a single visualization to harness the spatial analysis features of human visual perception [29]. In Figure 9 (Left), the hexagons with blue (12 PM ~ 6 PM) or green (6 PM ~ 12 AM)) color correspond to Tweets published on October 29th, 2012 and ones with orange or red color correspond to Tweets posted on the following day after the hurricane. New York City announced the evacuation of Zone A (red color) in Figure 9 (Right); residents in Zone A faced the highest risk of flooding, whereas, Zone B (yellow color) and Zone C (green color) are moderate and low respectively. In the visual representation, analysts can indicate overall spatiotemporal patterns of people and their movements during the disaster event—many people still remained at home one day after the mandatory evacuation order, but most people left home on the following day as the hurricane damaged the city.

5. Discussion and Evaluation

In this work we found out that the public responses to disaster events in social media streams are different according to the disaster event types. Hurricane Sandy had a long time duration—more than one week, and affected a wide range of areas. Therefore, there were many reactions in the potential damage area before the hurricane impacted the area. However, no or

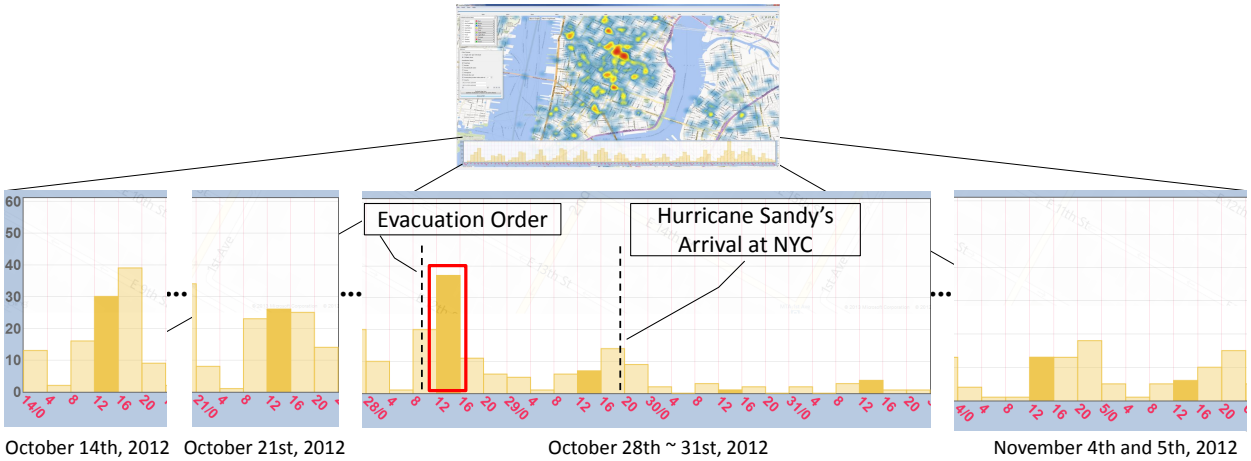


Figure 8: Temporal analysis for public behaviors during the disaster event, Sandy. Top shows our entire system view. The bar chart (Bottom) for the number of Twitter users within the selected region including a supermarket in Figure 1 (Right) in four hour intervals is shown. We see that many people went to the supermarket right after the evacuation order.

significantly less hotspots were found right after the hurricane passed over the area. This was because the hurricane severely affected the areas—communication facility damage and power outages occurred in the area. Moreover, we found out that unusual post-event situations in the Twitter user distribution continued for a certain time period from a couple of days to more than one week as shown in Figure 4 and 8. The analysts could estimate how long it took for the reconstructions in the areas.

Regarding the tornado case, we intended to find abnormal patterns in the Twitter user distribution before and during the disaster event but there was no unusual patterns in the area. In contrast to the hurricane, the tornado generally affected the areas relatively shortly, for example, a few minutes to an hour. The abrupt natural disaster did not strongly influence the social media stream before and even during the event. However, as shown in Figure 5, we were able to find many hotspots within the damaged areas after the tornado passed. In fact, the tornado damaged some small areas (i.e., a couple of miles wide), in contrast to the wide range of damaged areas for the hurricane case. This indicated that communication facilities were still available and many people were interested in the disaster, similar to the hurricane. Thus, our social media analysis could support the analysts to make plans and manage for the emergent situations according to the types of the disasters.

The above cases demonstrate how our system supports spatial decision making through evaluation of varying-density population area to determine changes in behavior, movement, and increase overall situational assessment. This increased spatial activity and behavioral understanding provides rapid situational assessment and provides insight into evolving situational needs to provide appropriate resource allocation and other courses of action (e.g., traffic rerouting, crowd control).

We requested informal feedback for the usability of our system from users within our universities, and received useful and positive comments and suggestions. They were interested in the findings of the abnormal situations during the disaster events in Section 4.1 and 4.3. They also noted that the use of the infrastructure symbols on the heatmaps improved the legibility of the Twitter user distributions in Figure 1 and they suggested a visualization for the deviations between multiple heatmaps in order to show the differences clearly, which we plan to develop in the future.

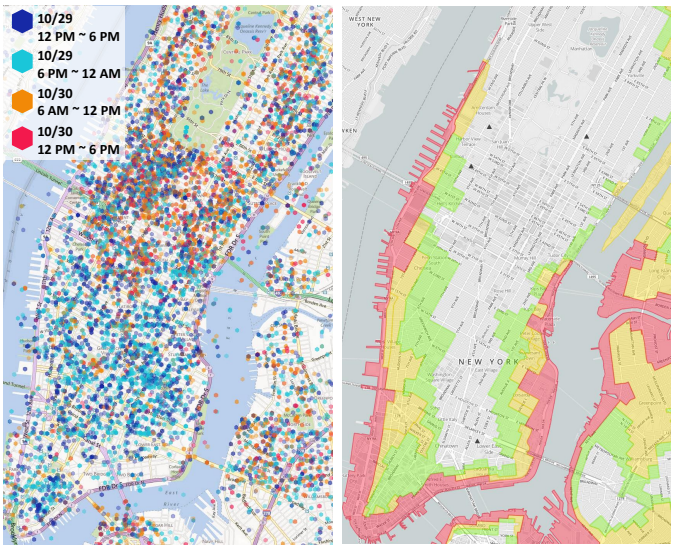


Figure 9: Visualization for spatiotemporal social media data (Left). A hexagon represents the spatial (position) and temporal (color) information of a Tweet. Hurricane evacuation map [30] (Right). Residents in Zone A (red) faced the highest risk of flooding, Zone B (yellow) and Zone C (green) are moderate and low respectively.

6. Conclusions and Future Work

In this work we presented a visual analytics system for public behavior analysis and response planning in disaster events using social media data. We proposed multiple visualizations of spatiotemporal analysis for disaster management and evacuation planning. For the spatial decision support, we demonstrated an analytical scheme by combining multiple spatial data

sources. Our temporal analysis enables analysts to verify and examine abnormal situations. Moreover, we demonstrated an integrated visualization that allows spatial and temporal aspects within a single view. We have still some limitations with these techniques including the potential occlusion issues in the spatiotemporal visualization. For future work, we will investigate the flow of public movement before and after disasters and the analysis for recovering from disasters and crises. We also plan to design the glyphs with varied sizes adapting to the zoom level in the spatiotemporal visualization. In addition, we will conduct a user evaluation for the usability and effectiveness of the geospatial visual support, and the impact of interactive spatiotemporal visual analytics using social media data on disaster management.

Acknowledgement

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0003. Jang's work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170). We would like to thank the reviewers for their valuable suggestions and comments, which helped to improve the presentation of this work.

References

[1] Webster T. The social habit - frequent social networkers. Retrieved March 3, 2013, `http://www.edisonresearch.com/home/archives/2010/06/the_social_habit_frequent_social_networkers_in_america.php`; 2010.

[2] Committee on Public Response to Alerts and Warnings Using Social Media: Current Knowledge and Research Gaps; Computer on Science and Technology Board; Division on Engineering and Physical Sciences; National Research Council . Public Response to Alerts and Warnings Using Social Media: Report of a Workshop on Current Knowledge and Research Gaps. The National Academies Press; 2013.

[3] American Red Cross . Social media in disasters and emergencies. Retrieved March 3, 2013, http://i.dell.com/sites/content/shared-content/campaigns/en/Documents/red-cross-survey-social-media-in-disasters-aug-2010.pdf; 2010.

[4] Chae J, Thom D, Jang Y, Kim SY, Ertl T, Ebert D. Visual analytics of microblog data for public behavior analysis in disaster events. In: EuroVis Workshop on Visual Analytics. 2013, p. 67–71.

[5] MacEachren A, Jaiswal A, Robinson A, Pezanowski S, Savelyev A, Mitra P, et al. Senseplace2: Geotwitter analytics support for situational awareness. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. 2011, p. 181 –90.

[6] Marcus A, Bernstein MS, Badar O, Karger DR, Madden S, Miller RC. Twitinfo: aggregating and visualizing microblogs for event exploration. In: Proceedings of the 2011 annual conference on Human factors in computing systems. ACM; 2011, p. 227–36.

[7] Thom D, Bosch H, Koch S, Woerner M, Ertl T. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In: IEEE Pacific Visualization Symposium (PacificVis). 2012, p. 41 –8.

[8] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. WWW '10; ACM; 2010, p. 851–60.

[9] Andrienko N, Andrienko G, Gatalsky P. Exploring changes in census time series with interactive dynamic maps and graphics. Computational statistics 2001;16(3):417–33.

[10] Lee R, Sumiya K. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. LBSN '10; ACM; 2010, p. 1–10.

[11] Pozdnoukhov A, Kaiser C. Space-time dynamics of topics in streaming text. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. LBSN '11; New York, NY, USA: ACM; 2011, p. 1–8.

[12] Abel F, Hauff C, Houben GJ, Stronkman R, Tao K. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In: Proceedings of the 23rd ACM conference on Hypertext and social media. ACM; 2012, p. 285–94.

[13] Terpstra T, de Vries A, Stronkman R, Paradies G. Towards a realtime twitter analysis during crises for operational crisis management. In: Proc. of the 9th Inter. ISCRAM Conf. 2012,.

[14] Vieweg S, Hughes AL, Starbird K, Palen L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the 28th international conference on Human factors in computing systems. ACM; 2010, p. 1079–88.

[15] Heverin T, Zach L. Microblogging for crisis communication: Examination of twitter use in response to a 2009 violent crisis in the seattle-tacoma, washington area. In: Proceedings of the 7th International ISCRAM Conference–Seattle; vol. 1. 2010,.

[16] Twitter . The geography of tweets. Retrieved August 24, 2013, `https://blog.twitter.com/2013/geography-tweets-3`; 2013.

[17] Wikipedia . Hurricane sany. Retrieved April 30, 2013, `http://en.wikipedia.org/wiki/Hurricane_Sandy`; 2012.

[18] Wikipedia . Effects of hurricane sandy in new jersey. Retrieved June 11, 2013, `http://en.wikipedia.org/wiki/Effects_of_Hurricane_Sandy_in_New_Jersey`; 2012.

[19] Andrienko G, Andrienko N, Jankowski P, Keim D, Kraak M, MacEachren A, et al. Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science 2007;21(8):839–57.

[20] Robinson A, Roth R, Blanford J, Pezanowski S, MacEachren A. Developing map symbol standards through an iterative collaboration process. Environment and Planning B: Planning and Design 2012;39(6):1034–48.

[21] Wikipedia . 2013 moore tornado. Retrieved June 12, 2013, `http://en.wikipedia.org/wiki/2013_Moore_tornado`; 2013.

[22] Twitchy . Convoy of hope: Glenn beck arrives in okla. with two truckloads of food, water and diapers. Retrieved August 26, 2013, `http://twitchy.com/2013/05/21/convoy-of-hope-glenn-beck-arrives-in-okla-with-two-truckloads-of-food-water-and-diapers/`; 2013.

[23] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

[24] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). Journal of Official Statistics 1990;6(1):3–73.

[25] Chae J, Thom D, Bosch H, Jang Y, Maciejewski R, Ebert D, et al. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: Visual Analytics Science and Technology, 2012 IEEE Conference on. Oct., p. 143–52.

[26] Maciejewski R, Hafen R, Rudolph S, Larew S, Mitchell M, Cleveland W, et al. Forecasting hotspots - a predictive analytics approach. Visualization and Computer Graphics, IEEE Transactions on April;17(4):440–53.

[27] Malik A, Maciejewski R, Hodgess E, Ebert D. Describing temporal correlation spatially in a visual analytics environment. In: System Sciences (HICSS), 2011 44th Hawaii International Conference on. Jan., p. 1–8.

[28] Andrienko N, Andrienko G, Gatalsky P. Exploratory spatio-temporal visualization: an analytical review. Journal of Visual Languages & Computing 2003;14(6):503 –41. Visual Data Mining.

[29] Treisman AM, Gelade G. A feature-integration theory of attention. Cognitive Psychology 1980;12(1):97 – 136.

[30] New York City . Hurricane evacuation map. Retrieved March 3, 2013, `http://gis.nyc.gov/oem/he/map.htm`; 2012.

VASA: Interactive Computational Steering of Large Asynchronous Simulation Pipelines for Societal Infrastructure

Sungahn Ko, Jieqiong Zhao, Jing Xia, *Student Member, IEEE*, Shehzad Afzal, Xiaoyu Wang, *Member, IEEE*, Greg Abram, Niklas Elmqvist, *Senior Member, IEEE*, Len Kne, David Van Riper, Kelly Gaither, Shaun Kennedy, William Tolone, William Ribarsky, David S. Ebert, *Fellow, IEEE*

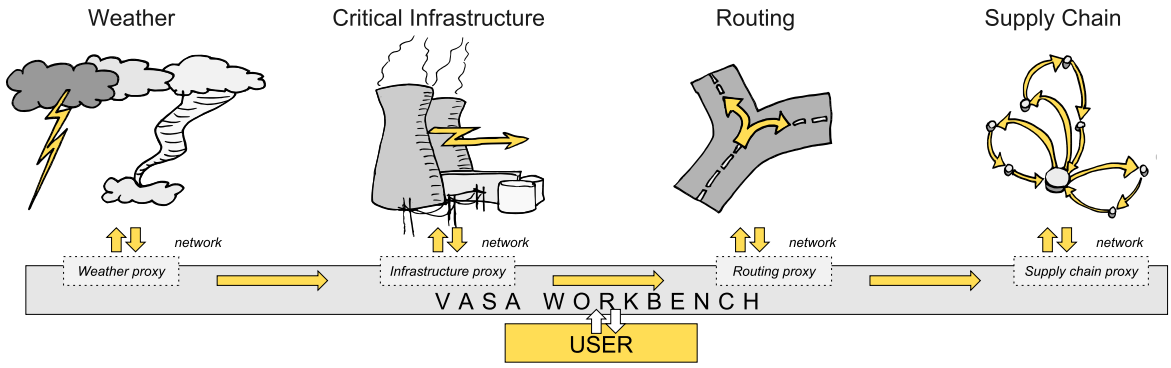


Fig. 1. Conceptual overview of the VASA system, including four simulation components for weather, critical infrastructure, road network routing, and supply chains, as well as the VASA Workbench binding them together.

Abstract—We present VASA, a visual analytics platform consisting of a desktop application, a component model, and a suite of distributed simulation components for modeling the impact of societal threats such as weather, food contamination, and traffic on critical infrastructure such as supply chains, road networks, and power grids. Each component encapsulates a high-fidelity simulation model that together form an asynchronous simulation pipeline: a system of systems of individual simulations with a common data and parameter exchange format. At the heart of VASA is the Workbench, a visual analytics application providing three distinct features: (1) low-fidelity approximations of the distributed simulation components using local simulation proxies to enable analysts to interactively configure a simulation run; (2) computational steering mechanisms to manage the execution of individual simulation components; and (3) spatiotemporal and interactive methods to explore the combined results of a simulation run. We showcase the utility of the platform using examples involving supply chains during a hurricane as well as food contamination in a fast food restaurant chain.

Index Terms—Computational steering, visual analytics, critical infrastructure, homeland security.

1 Introduction

Highways, interstates, and county roads; water mains, power grids, and telecom networks; offices, restaurants, and grocery stores; sewage, landfills, and garbage disposal. All of these are critical components of the societal infrastructure that help run our world. However, the complex and potentially fragile interrelationships connecting these components also mean that this critical infrastructure is vulnerable to both natural and man-made threats: twisters, hurricanes, and flash floods; traffic, road blocks, and pile-up collisions; disease, food poisoning, and major pandemics; crime, riots, and terrorist attacks. How can a modern society protect its critical infrastructure against such a diverse range of threats? How can we design for resilience and preparedness when perturbation in one seemingly minor aspect of our infrastructure may have vast and far-reaching impacts across society as a whole?

Simulation, where a real-world process is modeled and studied over time, has long been a standard tool for analysts and policymakers to answer such questions [10]. Using complex simulations of critical infrastructure components, expert users have been able to create "what-if" scenarios, calculate the impact of a threat depending on its severity, and find optimal mitigation measures to address them. In fact, analysts have gone so far as to name simulation as the "new innovation" [33]: instead of endeavoring to produce the perfect solution once and for all, this new school of thought is to create a whole range of possible solutions and determine the optimal one using modeling and simulation. For example, during the Obama reelection campaign, it was reported that Organizing for Action data analysts ran a total of 62,000 simulations to determine voter behavior based on data from social media, political advertisements, and polling [42]. Basically, the philosophy with big data analytics driven by simulation is not to get the answer perfectly right, but to be "less wrong over time" [32]. Put differently, while it would be inappropriate to state—as others have done [2]—that big data will ever overtake theory, it is clear that large-scale simulation is a new and powerful tool for making sense of the world we inhabit.

Applying simulation to the scope of entire critical infrastructures—such as transportation, supply chains, and power grids—as well as the factors impacting them—such as weather, traffic, and man-made threats—requires constructing large *asynchronous simulation*

- *Sungahn Ko, Jieqiong Zhao, Shehzad Afzal, Niklas Elmqvist, and David S. Ebert are with Purdue University in West Lafayette, IN, USA. E-mail: {ko, zhao413, safzal, elm, ebertd}@purdue.edu.*
- *Jing Xia is with Zhejiang University in Hangzhou, China. E-mail: xiajing@zjucadcg.cn.*
- *Xiaoyu Wang, William Tolone, and William Ribarsky are with University of North Carolina at Charlotte in Charlotte, NC, USA. E-mail: {xiaoyu.wang, ribarsky}@uncc.edu.*
- *David Van Riper, Len Kne and Shaun Kennedy are with University of Minnesota in Minneapolis, MN, USA. E-mail: {vanriper, lenkne, kenne108}@umn.edu.*
- *Greg Abram and Kelly Gaither are with University of Texas at Austin in Austin, TX, USA. E-mail: {gda, kelly}@tacc.utexas.edu*

Submitted to IEEE VAST 2014. Do not redistribute.

pipelines, where the output of one or more simulation models becomes the input for one or more other simulations arranged in a sequence with feedback. Such a *system-of-systems* [11, 28] (SoS) will enable leveraging existing high-fidelity simulation models without having to create new ones from scratch. However, this approach is still plagued by several major challenges that all arise from the complexity of chaining together multiple simulations in this way: (C1) *monolithic simulations* that are designed to be used in isolation, (C2) *complex configurations* for each model, (C3) *non-standard data exchange* for passing data between them, (C4) *long execution times* for each individual simulation that are not amenable to interactive visual analytics, and (C5) *uncertain and inaccurate* simulations compounded by their composition.

To address these challenges, we present **VASA** (Visual Analytics for Simulation-based Action), a visual analytics platform for interactive decision making and computational steering of these types of large-scale simulation pipelines based on a visual analytics approach. The VASA Workbench application itself is an interactive desktop application that binds together a configurable pipeline of distributed simulation components. It enables the analyst to visually integrate, explore, and compare the inter-related and cascading effects of systems of systems components and potential final alternative outcomes. This is achieved by visualizing both intermediate and final results from the simulation components using a main spatiotemporal view as well as multiple secondary views. The tool provides an interface for the analyst to navigate in time, including stepping backwards and forwards, playing back an event sequence, jumping to a particular point in time, adding events and threats to the timeline, and initiating mitigation measures. Moreover, it allows them to select between or combine different ensemble outputs from one simulation to be fed to other SoS components and explore consequences. Using this interface, an analyst could for example add a weather event (e.g., either an existing hurricane from a historical database, the union of several output paths, or simulation of a new one) to a particular time, and then step forward a week to see its impact on roads, power, and food distribution.

The simulation components provide the main functionality to the VASA platform. Each simulation component communicates with the Workbench using a representational state transfer (REST) API that standardizes the data and parameter exchange. The data flows and parameters passed in the pipeline can be configured in the Workbench application using a graphical interface. Furthermore, the Workbench also includes a local *simulation proxy* for each remote simulation component that provides real-time approximations of each simulation model to enable using them for interactive visual discourse. This feature also provides the computational steering functionality of the Workbench: after configuring a simulation run in an interactive fashion, the analyst can launch the (possibly lengthy) execution from the Workbench. The Workbench then provides tools to manage the simulation pipeline, for example to prematurely shut down a simulation component to accept a partial result, skip a run, or rerun a component with new parameters.

Our work on the VASA project has been driven by stakeholders interested in supply chain management of food systems, with an initial working example of a food production to restaurant system. For this reason, other than the VASA Workbench application and the protocols and interfaces making up the platform, we have also created VASA components for simulating weather (including storms, hurricanes, and flooding), the power grid, supply chains, transportation, and food poisoning. We describe these individual components and then present an example of how the VASA platform can be used to explore a what-if scenario involving a major hurricane sweeping North Carolina and knocking out a large portion of the road networks and power grid. We also illustrate how the tool can be used to simulate food contamination outbreaks and how this information can be used to track back the contaminated products to the original distribution centers.

2 Background

Visual analytics [36], can be a powerful mechanism to harness simulation for understanding the world. Below we review the literature in visual analytics for simulation and computational steering, as well as appropriate visual representations for such spatiotemporal data.

2.1 Simulation Models

The potential for applying visual analytics to simulation involves not only efficiently presenting the results of a simulation to the analyst, but also building and validating large-scale and complex simulation models. For example, Matkovic et al. [26] show that visual analytics can reduce the number of simulation runs by enabling users to concentrate on interesting aspects of the data. Maciejewski et al. [23] apply visual analytics techniques to support exploration of spatiotemporal models with kernel density estimation and cumulative summation. This approach has also been applied to epidemic modeling and decision impact evaluation [1]. Similarly, Andrienko et al. [5] propose a comprehensive visual analytics environment that includes interactive visual interfaces for spatial modeling libraries, including selection, adjustment, and evaluation. Our work is different from this prior art in that our approach combines multiple components in a simulation pipeline, where each stage in the pipeline provides visualization for analysis.

Supply chain management is another multi-decisional context where what-if analyses are often conducted to capture provenance and processes of supplies. Simulation is recognized as a great benefit to improve supply chain management, providing analysis and evaluation of operational decisions in the supply process in advance [35]. With the IBM Supply Chain Simulator (SCS) [9] and enterprise resource planning (ERP), IBM is able to visualize and optimize nodes as well as relations in the supply chain [20]. Perez also developed a supply chain model snapshot [29] with Tableau. However, existing visualizations of supply chains are mostly limited to either local supply nodes or a metric model rather than managing the overall supply process.

2.2 Computational Steering

Computational steering refers to providing user control over running computations, such as simulations. Mulder et al. [27] classify uses of computational steering as model exploration, algorithm experimentation, and performance optimization. Applications include computational fluid dynamics (CFD) [12], program and resource steering systems [38], and high performance computing (HPC) platforms [7].

For all of the above applications, the user interface is a crucial component that interprets user manipulation for configuring data, algorithms, and parameters. Controlling, configuring, and visualizing such computational steering mechanisms is an active research area. Waser et al. proposed World Lines [39], Nodes on Ropes [40], and Visdom [31] as well as an integrated steering environment [31] to help users manage *ensemble simulations*—multiple runs of the same or related simulation models with slightly perturbed inputs—of complex scenarios such as flooding. Endert et al. [13] show how to embed analysts in the analytics loop using computational steering. In the business domain, Broeksema et al. [8] propose the Decision Exploration Lab to help users explore decisions generated from combined textual and visual analysis of decision models rooted in artificial intelligence.

2.3 Spatiotemporal Data

Spatiotemporal visual analytics systems enable users to investigate data features over time using a visual display based on geographic maps [3]. In these systems, color, position, and glyphs display features of different regions by directly overlaying the data on the map.

Many approaches to visual analytics for spatiotemporal data exist [6]; a relevant sampling follows. Andrienko and Andrienko [4] use value flow maps to visualize variations in spatiotemporal datasets by drawing silhouette graphs on the map to represent the temporal aspect of a data variable. Hadlak et al. [16] visualize attributed hierarchical structures that change over time in a geospatial context. Fuchs and Schumann [15] integrate ThemeRiver [17] and TimeWheel [37] into a map to visualize spatiotemporal data. Finally, Ho et al. [18] present a geovisual analytics framework for large spatiotemporal and multivariate flow data analysis using bidirectional flow arrows coordinated and linked with choropleth map, histograms, and parallel coordinate plots.

Some approaches enable analysis of spatially-distributed incident data, which is of particular relevance here. Maciejewski et al. [22] propose a system for visualizing syndromic hotspots, while Malik et al. [24] develop a visualization toolkit utilizing KDE (Kernel Density

Estimation) to help police better analyze the geo-coded crime data. The latter system has also been extended [25] to historic response operations and assessment of potential risks in maritime environments.

3 Design Space: Steering System-of-System Simulations for Modeling Societal Infrastructure

Computational steering is defined as user intervention in an autonomous process to change its outcome. This approach is commonly utilized in visual analytics [36] when embedding a human analyst into the computation loop for the purpose of creating synergies between the analyst and computational methods. In our work, the autonomous processes we are studying are simulation models (often based on discrete event models) that are chained together into asynchronous simulation pipelines where the output of one or several simulations becomes the input to one or several other simulations. Such a simulation pipeline is also a *system-of-systems* [11, 28] (SoS): multiple heterogeneous systems that are combined into a unified, more complex system whose sum is greater than its constituent parts. Synthesizing all these components yields the concept of visual analytics for *steering system-of-system simulations*: the use of visual interfaces to guide composite simulation pipelines for supporting sense-making and decision-making. In this work, we apply this idea to modeling societal infrastructure, such as transportation, power, computer networks, and supply chains. Below we explore the design space of this concept, including problem domains, users, tasks, and challenges. We then derive preliminary guidelines for designing methods supporting the concept.

3.1 Domain, User, and Task Analysis

The concept of creating large-scale system-of-system simulation pipelines is applicable to a wide array of problem domains. Our particular domain is for business intelligence for supply chain logistics in the fast-food business, but we see multiple other potential applications:

- **Supply chain logistics:** Impact of large-scale weather events on the distribution of goods (particularly perishables, e.g., food).
- **Public safety:** Crime, riots, and terrorist attacks on critical infrastructure, such as on roads, bridges, or the power grid.
- **Food safety:** Incidence, spread, and causes of food contamination, often due to weather (power outage) or transport delays.
- **Cybersecurity:** Societal impact of cybersecurity attacks, such as on power stations, phone switches, and data centers.

The intended audience for computational steering of simulation models using visual analytics is what we call "casual experts:" users with deep expertise in a particular application domain, such as transportation, supply chain, or homeland security, but with limited knowledge of simulation, data analysis, and statistics. Their specific background depends on the problem domain; for example, they may be logistics analysts for supply chain applications, police officers for public safety, and homeland security officials for food safety and cybersecurity. Because of this "casual" approach—a term we borrow from Pousman et al.'s work on casual information visualization [30]—our intended users are motivated by solving concrete problems in their application domain, but are not necessarily interested in configuring complex simulation models and navigating massive simulation results.

Even if our primary audience is these casual experts, it is likely that the outcome of a simulation analysis will be disseminated to managers, stakeholders, or even the general public [36]. Thus, a secondary user group for consuming our analysis products is laypersons with an even more limited knowledge in mathematics, statistics, and data graphics.

In our particular application, we identified tasks for simulation steering by working with a group of analysts from a restaurant chain that has a very large number of restaurants across the U.S, as wel as a food supply chain involving farms, food processing centers, and food distribution centers. The two main concerns voiced by these analysts are better understanding and traceability of their supply chain and understanding resiliency/vulnerability of their food supply network, especially in relation to pertain to (C1) *severe weather* and (C2) *food poisoning*: understanding the impact of natural disasters (e.g., hurricanes) on their food supply chain, processing facilities, and restaurants, as well as determining the source and distribution of food contamination cases in relation to their restaurants. More specifically, our analyst audience wants to perform the following high-level tasks in relation to these two concerns:

- Increasing *preparedness* for potential scenarios;
- Improving the *resilience* of the restaurant chain; and
- Planning for *mitigation and response* to a situation.

A motivating example for our target analysts is to understand the impact of severe weather (e.g., hurricanes) on power plants and roads, which may directly or indirectly impact food processing centers, distribution centers, and restaurants. Direct impacts include power outages, flooding, and evacuation. Indirect impacts, on the other hand, occur due to direct impacts earlier in the supply chain, such as a farms, food processing or distribution centers going offline causing shortages and redistribution of products. Both types of impacts may cause closing of facilities, which in turn may lead to indirect impacts downstream in the supply chain. Detecting such closures allows the analysts to mitigate their impact, for example by rerouting deliveries from other distribution centers, or even transporting back frozen products from a restaurant lacking refrigeration due to an extended power outage. In a hurricane scenario, the primary task then becomes determining which facilities will be closed, which routes will be impassible, and the impacts and duration these will have throughout the supply chain. Similar effects can be caused by power failures caused by other events (e.g., severe summer demand, tornadoes, power grid cyberattacks). These failures can also impact food safety (C2) due to spoilage and conditions favorable for contamination. If this is not prevented, it leads to the second task named by our target analysts: the capability to model food contamination and backtrack to its source so that the contamination can be stopped. Similar to the hurricane example above, this also requires coordinating multiple interdependent simulation models. Unfortunately, our user group does not currently have tools for performing a series of simulations to explore these scenarios.

3.2 Challenges

Modeling the real world is a tremendously difficult and error-prone process. However, we leave concerns about the fidelity, accuracy, and quality of a simulation to researchers from the simulation field. Rather, in this subsection we concern ourselves with the challenges intrinsic to connecting multiple individual simulation models into large-scale pipelines. In the context of simulation steering for such pipelines, we identify the following main challenges from our analyst audience:

- C1 **Monolithic simulations:** While individual high-fidelity simulation models exist for all of the above components and threats, these models are monolithic and not designed to work together.
- C2 **Complex relationships:** Each high-fidelity simulation model consists of a plethora of parameters and controls that require expertise and training, which is exacerbated when several such models are combined into a single model.
- C3 **Non-standard data:** No standardized data exchange formats exist for passing the output of one simulation model, such as for weather, as input to another model, such as supply chain routing.
- C4 **Long execution times:** Most state-of-the-art, high-fidelity simulation models require a non-trivial execution time, often on the order of minutes, if not hours. Such time frames are not amenable for real-time updates and interactive exploration.
- C5 **Uncertainty and fidelity:** Chaining together multiple simulations into a pipeline may yield systematically increasing errors as uncertain output from one model is used as input to another. This is compounded by the fact that heterogeneous simulation models may have different levels of fidelity and accuracy.

3.3 Design Guidelines

Based on our review of the problem domain, users, and tasks above, as well as the challenges that these generate, we formulate the following tentative guidelines for designing visual analytics methods for steering system-of-system simulation pipelines:

G1 *Simulations as standardized network services:* Distributing simulation models as network services avoids the trouble of integrating a monolithic design with another system (C1) and automatically provides a data exchange format (C3). The simulations also become decoupled, which means they can be parallelized and/or distributed in the cloud to manage long execution times (C4).

G2 *Simulation proxies for interactive response:* Meaningful sensemaking in pursuit of one of the high-level tasks in Section 3.1 requires real-time response to all interactive queries. This means that long execution times (C4) of simulation models in the pipeline should be hidden from the user. We propose the concept of a *simulation proxy* as an approximation of a remote simulation service that is local and capable of providing real-time response at the cost of reduced (often significantly) accuracy.

G3 *Visual and configurable relationships:* The interactive visual interfaces routinely employed in visual analytics may help to simplify and expose the complex configurations necessary for many high-fidelity simulation models (C2), even for non-expert users.

G4 *Partial and interruptible computational steering:* Once an analyst has configured a simulation run using simulation proxies (G2) and visual mappings (G3), the full simulation pipeline must be invoked to calculate an accurate result. A full-fledged simulation run may take minutes, sometimes hours, to complete. The computational steering mechanisms provided by the software should provide methods for continually returning partial results [14] as well as interrupting a run halfway through.

G5 *Visual representations of both intermediate and final results:* To fully leverage the power of visual analytics, we suggest using interactive visual representations of simulation results. Such visualizations should be used for both intermediate data generated by a simulation component anywhere in the pipeline—which would support partial results and interrupting a run at any time—as well as for the final results. All visual representations should be designed with uncertainty in mind (C5), and providing intermediate visualizations should also help in exposing propagation of increasing error. Finally, it may also be useful to use visual representations for the approximations created by simulation proxies (G2), but these should be clearly indicated as such.

G6 *Spatiotemporal focus:* The primary data dimension of interest for results from simulation pipelines has both spatial and temporal attributes; for this reason, we will base the visual analytics interface on spatiotemporal visualization [3, 6]. Secondary visualizations may focus exclusively on time, space, or quantities.

4 VASA: Overview

VASA (Visual Analytics for Simulation-based Action) is a distributed component-based framework for steering system-of-system simulations for societal infrastructure. Figure 1 gives a conceptual model of the system architecture. At the center of the system is the VASA Workbench (Figure 2), a user-driven desktop tool for configuring, steering, and exploring simulation models, impacts, and courses of action. The workbench provides a visual analytics dashboard based on multiple coordinated views, an event configuration view, and a computational steering view. The workflow of the workbench revolves around initiating, controlling, analyzing, exploring, and handling events from the remote simulation components as well as the local simulation proxies.

Within the dashboard, events are displayed in a selectable calendar view (a) where each event's name, dates and a user-selected representative attribute (e.g., storm's maximum wind speed) are shown. The selected events from (a) are listed chronologically in the event viewer (b), where a user can select times for investigation. A toolbar (b-1) provides buttons for initiating simulations (e.g., cyberattack, storm simulations, distribution re-routing), selecting combinations of events (union, intersection, difference), selecting event visualization modes (polygons, contours), and triggering chronological playback.

Users can select a time within an event for comparison (right-clicking on a event's black rectangle), causing a red mark to be shown in the upper right corner of the associated event (b-2), and the corresponding impact to be highlighted in the main geospatial view (d-1, Sandy, red). This allows for comparing across different events and effects. We provide a legend window (c) for selected properties and the geographical view (d) renders the simulation results, including event evolution, routing paths, and impacts on critical infrastructures. Several of the dashboard widgets are plugged in from the simulation components. For example, a food delivery schedule to each store within a supply chain is provided in (e) where the x-axis presents corresponds to different restaurants while the y-axis represents different food processing centers or different types of foods. Here, the darker the red, the larger the quantity of the delivered food. The quantity information is provided in a tooltip that helps a user to estimate possible losses. This view enables traceback analysis (e.g., which type of food was contaminated from which processing centers, how much contaminated food was delivered to which store, etc) for food contamination incidents.

5 VASA: Components

The VASA suite currently provides four simulation components: weather, critical infrastructure, routing, and supply chains. Data from each component and proxy is processed and merged before being visualized in the Workbench. Each proxy not only processes and stores data for its own visualization but also communicates with other proxies upon request. For example, to visualize new delivery routes, the routing proxy asks the infrastructure proxy for impacted stores before approximating new routing information. In this way, the VASA system uses a loosely coupled state that is distributed across components and proxies. We review each of the VASA components below.

5.1 Weather Component

To provide analysts with a one-stop source for weather data, we implemented a server that asynchronously amasses data from on-line sources and presents it to clients through a RESTful web interface. This provides access to weather data—both historical, current, and modeled—through a singly authenticated VASA component. The service can be queried by the user or set into a push-mode to send new events to the VASA Workbench during severe weather season (e.g., hurricane, flood, tornado season, etc).

5.1.1 Simulation Model and Simulation Proxy

Beyond historical data, the VASA weather component currently collect both ADCIRC and NOAA weather models. The ADCIRC (ADvanced CIRCulation) model is a collaboration of several research centers off the East and Gulf coasts of the United States. Active during hurricane season, these models are run every four hours when storms are presents, producing ADCIRC-formatted datasets at fixed intervals forward from the start point. These results are made publicly available using THREDDS and OPeNDAP for cataloging, discovery and data access. Similarly, NOAA produces wind-speed probabilities along the tracks of many types of storms as contours at 34, 50, and 64-knot levels. When updated datasets appear on the respective dissemination sites, we import them onto the VASA weather server, which provides a simple API to access the data in convenient multi-resolution formats.

The proxy in this component has two roles. The first role is to prepare all event datasets from the remote event server. Therefore, the system first checks for new updates from the server. If there is a new update, it retrieves the data and caches it on the local workbench for faster loading. The second role is to visualize new status of an event on the date that a user selected and notify the status change of the event to other proxies. An example status change is a user changing the start date of a hurricane in the event viewer. When this happens, the proxy visualizes a new state of the hurricane on that date and notifies this change to other components, which initiates work by downstream proxies (e.g., estimating an area without power and impassable roads).

For visualizing weather data, the user can select the visual representation either as polygons or as contours as shown in Figure 2 (b-2, the last button). In polygonal mode, two probability models (blue with two different opacities) are projected as shown in the magnification view in Figure 2. Here, the smaller polygon represents a predicted path with high probability, and a larger one represents a predicted path with low probability. When a user selects a hurricane, the hurricane turns

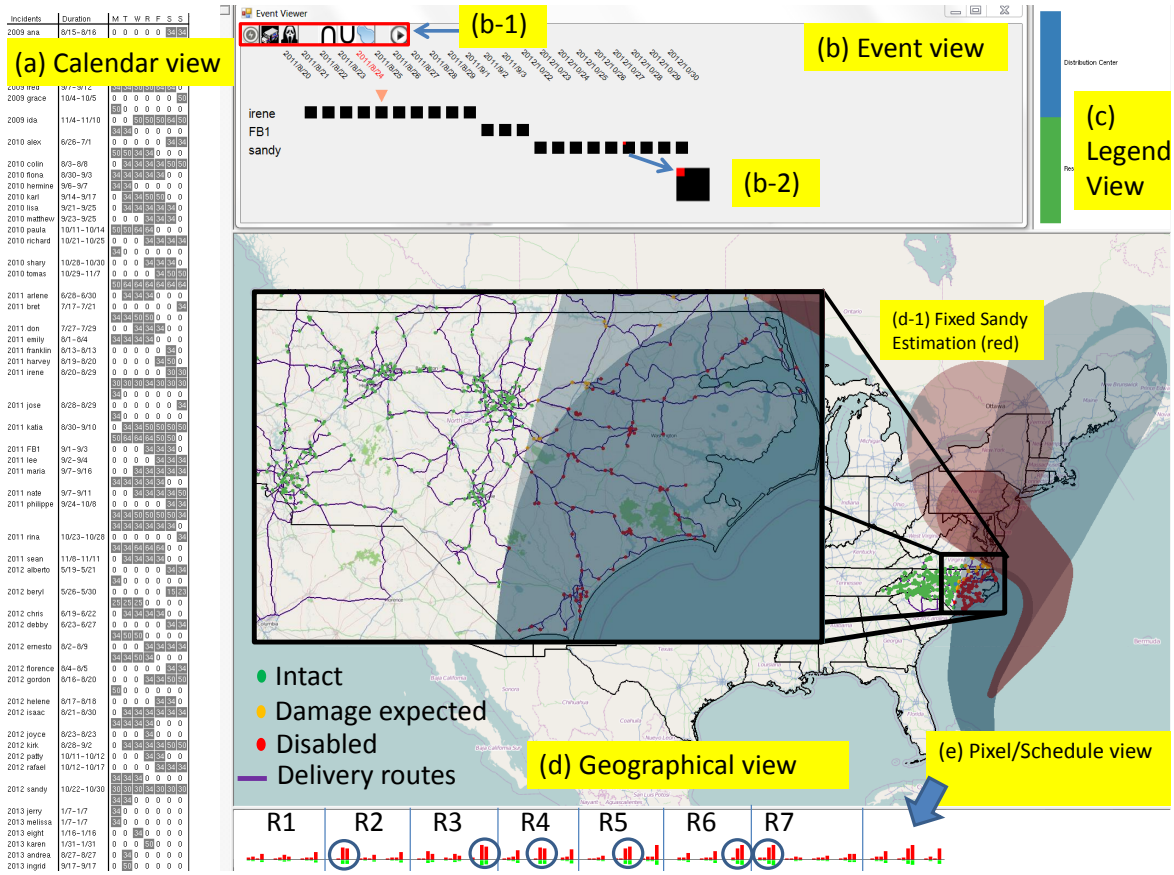


Fig. 2. Multiple coordinated views in the VASA Workbench. (a) Calendar view with available events (e.g., weather, food poisoning, cyberattack, etc). (b) Event timeline for configuring events. (b-1) Event buttons. (b-2) Fixed event. (c) Map legend. (d) Geographical map. (d-1) Fixed Sandy estimation (red). (e) Pixel/schedule view showing food deliveries. Each area divided by a blue line means a route that visits 3–4 restaurants, 3 times a week. This view also can be used for pixel-based visualization.

red for comparison to other hurricane paths. For example, in Figure 2 the paths of Hurricane Irene on August 24, 2011 (blue) and Hurricane Sandy on October 27, 2012 (red) are both rendered for comparison.

In contour mode, on the other hand, hurricanes are drawn using three different sizes of contours, each representing mean areas in different wind speeds (e.g., Hurricane Irene in our simulation model has 64 knot highest wind speed at the innermost contour, and 34 knot lowest wind speed at the outermost contour as shown in Figure 6). To utilize different wind speeds in simulation steering, a user can set up a threshold for infrastructures (e.g., a power generation unit is disabled if the wind hitting the plant has speed higher than 34 knot). In addition, a user can apply one of the contours for a time. For example, Figure 6 (top-right) presents which power generation units are affected when a contour with 34 knot hits the area. Here, red circles represent affected restaurants and purple circles represent power generation units supplying electricity to those restaurants.

5.1.2 Input and Output

The weather component often serves as a starting point for analysis by alerting severe weather conditions, and thus typically has no upstream component dependencies. Instead, simulation runs are often initiated by the analyst by adding weather events—current, modeled, or historical—to the timeline. Available weather events currently only include hurricanes, but are being expanded to other severe storm alerts, and are listed in a calendar view (Figure 2(a)).

5.1.3 Implementation Notes

The VASA weather component is implemented as a web service accessed using the common VASA RESTful API. All data objects are represented by URLs that encode procedures and parameters that, when issued, return JSON objects containing the results. This provides a very simple interface for use both by browser-based visualization UIs that use AJAX to issue requests asynchronously, and other platforms that provide access through language-specific interfaces.

5.2 Critical Infrastructure Component

Widespread emergencies such as hurricanes, flooding, or cyberattacks will often affect multiple societal infrastructures. High winds and flooding from a hurricane, for example, could knock out parts of the power grid, the effect of which would cascade to traffic signals, the communications network, the water system, and other infrastructures. The flooding might simultaneously make parts of the road network impassable. These breakdowns would affect critical facilities such as schools, hospitals, and government buildings. For longer-lived disasters, food distribution might break down due to power outage, route disruption, or other cascading effects. The purpose of VASA's critical infrastructure component is to simulate how such external emergencies, modeled in other components, will impact critical infrastructure.

5.2.1 Simulation Model and Proxy

To capture these complex, multifarious, and dynamic effects, the VASA critical infrastructure component takes into account the interrelationships between critical infrastructure systems. The simulation is built within the Vu environment [41] (Figure 3), which provides a rule-based framework for integrating multiple infrastructure submodels at a high level. This results in an interdependency ontology. Thus, for example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a school were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down. Likewise, telecommunication nodes would switch to backup power that might also shut down after its prescribed duration. There could also be outages due to power

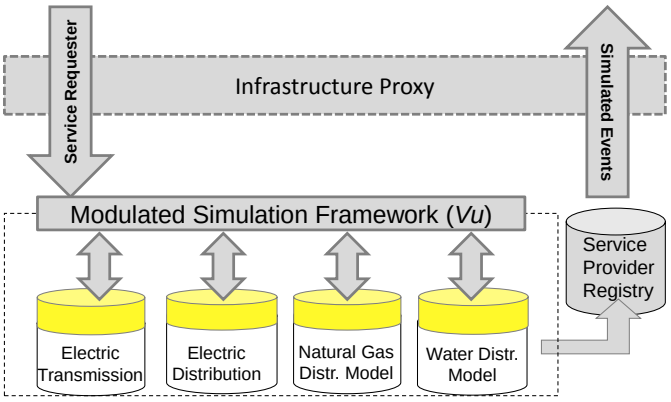


Fig. 3. Critical infrastructure server containing the pre-computed ensemble database and Vu environment with its simulation submodels.

load imbalances at other points in the grid.

The infrastructures we currently model include the electric grid; the communications network including TV stations, radio stations, cellular switch controls, and cell towers; transportation facilities including airports, bus terminals, rail lines and terminals, bridges, tunnels, and ports; the road network including main and secondary roads; natural gas pipelines and pumping stations; critical facilities including fire stations, police stations, schools, hospitals, emergency care facilities, manufacturing locations, government buildings, and hazmat facilities. Figure 4 shows the electric grid, which includes the complete transmission network down to substations for both North and South Carolina. Parts of the distribution network are also included, especially for critical installations. Figure 5 shows the transportation network for North Carolina, including roads, airports, rail lines, etc. For the purposes of VASA, we have also added store and distribution center locations for a large food chain in North and South Carolina. These facilities are linked to the power grid and road networks.

The proxy for the critical infrastructure component maintains a simplified connectivity network of critical infrastructure. In this graph, restaurants are connected to the nearest plant. When the proxy receives a signal of a new event (e.g., storm path change, new day for approximation), it computes which infrastructures are affected by the event. For example, when a user moves the hurricane simulation forward to a new day, our proxy checks which infrastructures are newly affected and produces an estimate and its corresponding visualization (e.g., color changes for restaurants affected by power disruptions).

5.2.2 Input and Output

The primary inputs of this component comes from the weather component represented as polygons of severe weather, such as wind speed, precipitation, and temperature data. Furthermore, the component also accepts direct manipulation of simulation parameters for particular facilities from the Workbench itself, such as the user manually shutting down a power substation. The outputs returned from the component is a list of facilities (e.g., restaurants, food processing centers and distribution centers) that are closed, and a duration of their closures.

Our prototype system currently uses data from the state of North Carolina, and the data collection and organization process involves locating and identifying components of the various infrastructures for the state. We use publicly available data sources, in some cases identifying infrastructure components by indirect means. For example, comprehensive information about the electrical grid is closely held by the utility companies. However, we have shown our results to utility company officials and received confirmation as to their high accuracy.

5.2.3 Implementation Notes

As for all the other VASA simulation components, we use a web service that can accept requests from the VASA Workbench and return simulation results ready to be presented in the user interface (see Figure 1). The critical infrastructure server itself has two components.

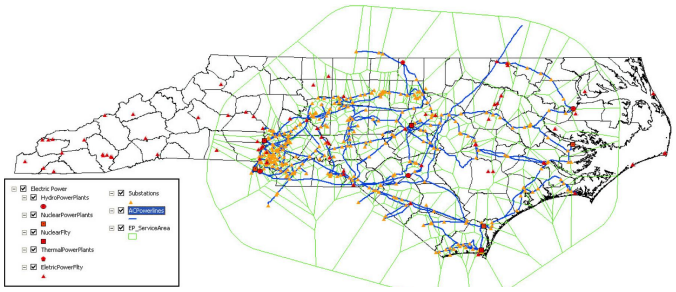


Fig. 4. Power transmission grid with parts of the distribution network.

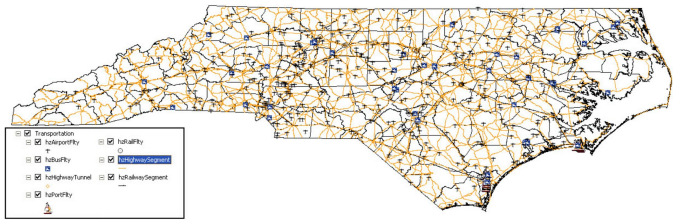


Fig. 5. Transportation network including transportation facilities.

One contains a searchable database of the pre-computed ensemble of simulation runs. The other accepts current storm path and other inputs from the weather component, converts them into courses of action, and computes a fresh set of cascading infrastructure disruption results via the Vu environment. When a request is issued via the user interface, the simulation proxy determines the weather inputs to send to the ensemble database component that immediately selects the closest ensemble simulation for use in the visual analysis. This proxy is then replaced by the more accurate result from the Vu model based on the current weather simulation as soon as it is available. Both the ensemble selection and the Vu model are depicted in Figure 3. Therefore, an emergency response manager can make initial decisions based on the proxy and then refine them once the up-to-date result is available.

5.3 Supply Chain Component

Most food systems involve a number of firms from on-farm production of inputs through processing, distribution and retail sales. For the fast food system in VASA, three different firms have collaborated to provide the data on normal system performance: a vertically integrated poultry firm (hatchery to processed chicken), a warehouse and distribution firm, and a fast food restaurant firm. Each firm contributed data from their portion of the supply chain to enable modeling of product movement from farm to restaurant. The type of data provided includes geospatial information on the facilities involved (e.g., feed mills, hatcheries, poultry farms, poultry processing facilities, distribution centers and restaurants), normal transportation routes and scheduled times from each facility to the next facility in the system and details on actual shipment quantities on average (hatchery through processing) or actual shipment records for a limited time frame (distribution centers to restaurants). As an illustration of the amount of data that drives these systems, one week of data on product delivered from the two distribution centers to the nearly five hundred individual restaurants alone constitutes more than 120,000 individual records.

Hurricanes pose significant risks for normal supply chain operation from impassable roads, power outages, and floodings disrupting facility operation and distribution of products in the system. Understanding which routes and locations are likely to be at risk from a storm would enable a firm to develop contingency plans in advance of a storm, thus reducing operational losses immediately after a storm. Given that daily sales at larger fast food restaurants can be $4,300-$7,400, losses can mount quickly. For an impending storm or immediate aftermath, rerouting could enable firms to most efficiently maintain their distribution systems for both maintaining product distribution and retrieving food from restaurants without power to minimize spoilage losses.

5.3.1 Simulation Model and Proxy

The primary objective of the supply chain component is to model distribution of product from food processing plant, through food distribution center, and to the restaurants. The routing of transports are handled in another component (Section 5.4); however, a primary concern of this component is to track product for the purpose of food safety.

Food contamination can occur both intentionally or as a malicious act at any point in the supply chain and can result in significant public health consequences, from morbidity to mortality. While firms are required to have information one step forward and one step back in their supply chain, they often have difficulty gaining visibility beyond that. By gathering data from each step in the supply chain, it is possible to trace product from farm through to restaurant and from restaurant back to farm. Using data on actual lot sizes from the firms involved, two illustrative contamination scenarios were constructed to illustrate how differently seemingly similar contamination scenarios would transpire. This system also illustrated a common problem of "hidden nodes" in the system, i.e., facilities that one firm in the system does not realize are part of its supply chain. One of the poultry slaughter and processing facilities ships raw poultry to a further processing facility that then ships the resulting product to the distribution centers. If there were a contamination at the "blind" facility, neither the distribution firm for the restaurant firm would initially know that it was part of their supply chain. A contamination scenario builder is now under development that would enable users to model a wide range of contamination events and see how they would propagate through the supply chain.

Our simulation model can generate food-borne illness data based on an approach similar to the Sydovat [21] system. There are two major components of the model for generating synthetic illness data: temporal and spatial data. A time series is contructed from its individual components (day-of-week, interannual, interseasonal, and remainder) similar to seasonal trend decomposition. To generate the time series of food-borne illnesses for a user-injected restaurant location, the user defines the mean daily count of illnesses along with seasonal and day of week components. If historical data is available, then seasonal and day of week components can be selected from this historical data. Spatial locations for temporal data are generated based on the population density distribition in that area. The analyst can also customize the grid size and density distributions.

Our simulation proxy for the supply chain component maintains a low-fidelity representation of the transport network. This is used together with the weather polygons to approximate when a distribution center and store must shut down. For food poisoning scenarios, this inherently contains spatially-distributed points of ill people simulated based on the simulation model (Section 5.3.1). To visualize the spatial distribution and the hotspots of the poisoned people, the proxy in this component uses a modified variable kernel density estimation technique with varying scales of the parameter of estimation based upon the distance from a patient location to the k_{th} nearest neighbor [34]. The model used for estimating the number of people poisoned is the same model utilized in Maciejewski et al. [1, 21], but we adjust parameters to consider different population densities in different regions.

5.3.2 Input and Output

This component accepts closures, including their durations, on supply chain facilities from the critical infrastructure component as well as severe weather polygons from the weather component. It then maintains and provides three types of information: (1) geo-information of all facilities of the supply chain, (2) delivery schedules, and (3) food products inventory in all locations (weight, size, and price).

5.3.3 Implementation Notes

The supply chain component is built in ArcGIS and Arc Network Modeler so that storm impacts can model solutions accounting for restaurants out of service (power, flooding) and impassable roadways.

5.4 Routing Component

The purpose of the routing component is to provide a mechanism for other VASA components to find appropriate routes from one facility to another given a dynamically changing world model, where roads may become impassable due to weather or other widespread emergencies.

5.4.1 Simulation Model and Proxy

The input to the routing component is a polygon representing an area impacted by severe weather (such as a hurricane). The component uses this input as a polygon barrier in the road network. Attributes of the road network are weighted to create a friction surface that iterates through routing options to determine the optimal route. The model does not currently include current traffic conditions or construction activity, but these factors could be added in the future. Each route minimizes the travel time between the distribution center and the first store or between stores. This set of routes represents the baseline scenario: how delivery trucks would travel under normal circumstances. Since delivery trucks can no longer reach outlets covered by the weather barrier, the routing service recomputes the routes with the barrier in place and returns new routes which avoid the outlets and roads covered by the barrier. If the barrier covers a distribution center, no deliveries will be made to outlets serviced by the center.

The main focus of the proxy in the routing component is on approximating the number of routes that will be replaced if a complete simulation result exists. The proxy investigates which nodes in routes are expected to be disabled when there is an event. Then, after the investigation, it builds a polygon by connecting outer-most nodes and visualizes the polygon. This gives awareness to a user that the routes in the polygon are likely to be changed after a complete simulation.

5.4.2 Input and Output

The severe weather data is ingested into the component as GeoJSON objects from the weather simulation component. Similarly, the calculated routes are output as a set of large GeoJSON objects and sent back to the caller (most often the supply chain component). One important input in this component is the impact area provided by the workbench that is presented by a polygon. Once this input is received, this component recalculate routes for the area in the polygon.

The geospatial database used by the component currently includes the addresses of two distribution centers and 505 fast-food outlets in our dataset, including the route information that links the centers to the outlets. This also includes the N shortest path routes, where N is the number of routes specified in the input data. The road network has a long list of attributes used to determine the shortest route, including road class, speed limit, number of lanes, and weight restrictions.

5.4.3 Implementation Notes

We implemented the routing component using ArcGIS Server 10.2 with the Network Analyst extension and the StreetMap Premium (TomTom North America data) road network. In general, the Esri suite of Geographical Information System (GIS) tools is widely used in a variety of industries and provides a robust set of tools and data. The server provides web-based services through REST endpoints and a robust API accessed with HTTPS GET or POST requests. The VASA workbench initiates a request to the routing service by providing a GeoJSON representation of the affected area. The affected area polygon is sent to Network Analyst Service to recalculate the route to traverse around the affected area. The response is two large GeoJSON objects containing a list of outlets no longer reachable, incremental travel time between stops, and the new route. Currently, the route processing requires 2-3 minutes to complete; this can be significantly improved in the future by commissioning a dedicated production server.

6 Examples

Here we demonstrate how the VASA system provides situational awareness using two examples: the impact of weather on macro-scale supply chains, and foodborne illness contamination and spread.

6.1 Supply Chains During Hurricane Season

Our first example is the potential impact of hurricanes on North Carolina's critical infrastructure, especially our food distribution network,

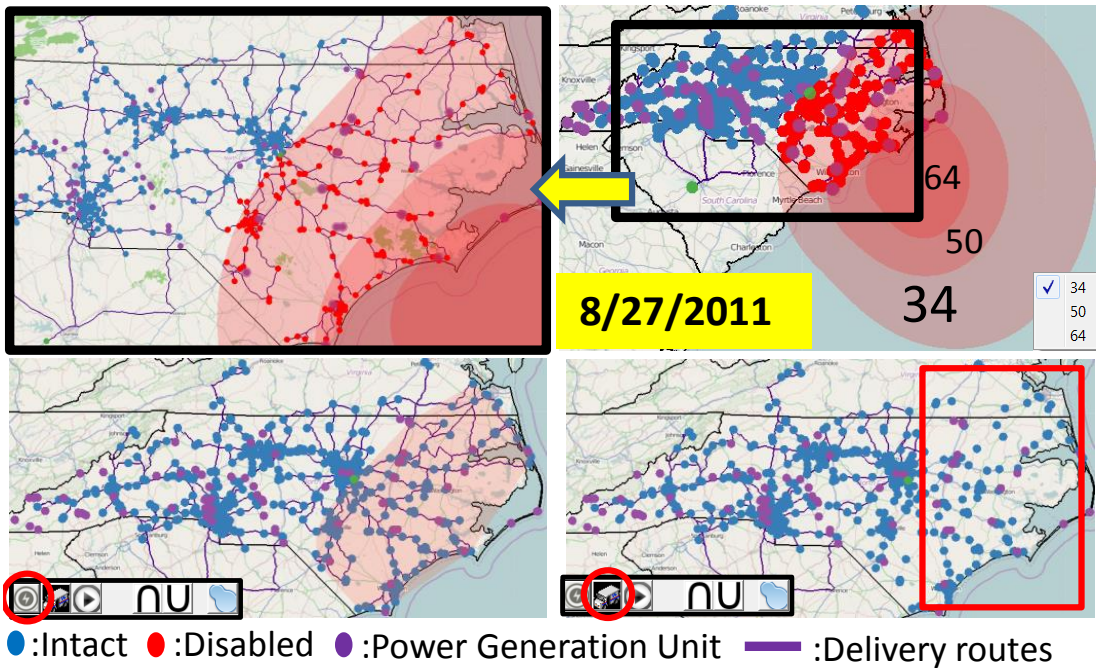


Fig. 6. In this simulation, power generation units were hit by up to 34 knot during Hurricane Irene on August 27, 2011. Our hurricane proxy instantly estimates the impacted restaurants (right-top, left-top). Note that one distribution center (green) is outside the hurricane. After a complete power-grid simulation run is finished (by clicking the circled lightning button), a polygon representing the power outage area is shown. Next, this polygon is sent for use in computing new food delivery paths. Note that food is not delivered to the power outage area (right-bottom, red box).

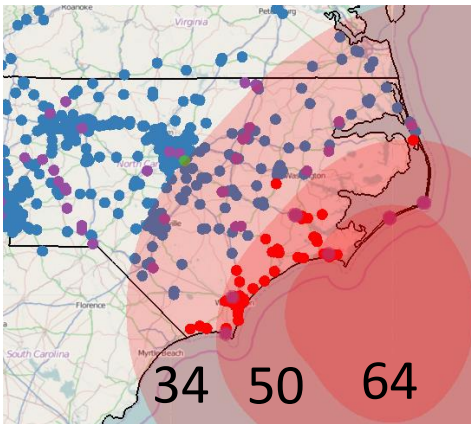


Fig. 7. If the power substations could have resisted up to 50 knot winds, the number of impacted restaurants would have been much smaller.

in North Carolina (NC). Our exploration begins by selecting appropriate historial hurricanes for examination using the calendar view as shown in Figure 2, where each hurricane name, duration, and selected summary attribute (e.g., maximum hurricane wind speed) are provided. While we investigate the paths of these historical hurricanes, we see that Irene in 2011 and Sandy in 2012 passed over NC. Because Sandy passed over only a small area in upper NC (Figure 2 (d), red polygon), we choose to focus on Irene for further investigation.

One interesting date is August 27, 2011 when Irene passed directly over eastern North Carolina, an area with many power generation facilities as shown in Figure 6 (top-right, purple circles). After we set up the wind tolerance value for these facilities to be 34 knots, our hurricane proxy instantly estimates which restaurants will be impacted based on the relationships between the units and the restaurants, coloring the impacted restaurants red. Here, we also initiated a complete simulation for power outages and transportation network damage. Next, a polygon is shown representing an area where restaurants are disabled and roads are blocked (bottom-left in Figure 6). To efficiently manage distribution, this impact requires the food provider to change its delivery schedule, and this new routing is computed based on the impacted restaurant polygon and road conditions (e.g., blocked by flooding). After a simulation to compute the new routes (by clicking the truck button in a red circle, right-bottom Figure 6), we see that the updated delivery paths do not include the affected restaurants. The economic loss caused by this event is estimated based on the model in Section 5.3 as being up to $1.13 million. Another possible what-if question is "How different would the result be if the power generation units could resist winds up to 50 knots?" Figure 7 shows the first step of the analysis where we see many fewer restaurants affected compared to Figure 6 top-right (units are resilient to 34 knots). In this case, the estimated losses are less than $333,000.

6.2 Fast Food Contamination

Food poisoning is an illness caused by eating contaminated food containing viruses, bacteria, and germ-generated toxins. There are many possible causes of food contamination including storage at inappropriate temperatures [19], improper food handling, and cross-contamination during processing or packaging. As unfortunately experienced several times per year, tracing back the cause of the contamination is a very difficult and lengthy process. In this example, we explore a hypothetical scenario demonstrating how VASA can be used to trace-back the root causes of an incident of foodborne illness.

To create the distribution of the ill population, we simulate the distribution of contaminated food to stores, then simulate the illnesses in the neighboring areas using the simulation model discussed in Section 5.3. This creates the common base scenario of reports of people who are ill, their date of illness, and their location to create the food contamination scenario for the trace-back investigation.

For example purposes, we simulated these illnesses occurring during a three day span (September 1, 2011 to September 3, 2011) as shown in Figure 8. Since this is almost one week after Hurricane Irene, one may assume that power outages during the storm could be the possible reason behind the contamination. To confirm this hypothesis, we looked at the hot spots in Figure 8 and identified the stores closest to these hot spots. On cross comparison, we can identify the common products/lots in those stores, their distribution center, as well as their delivery mechanisms. As shown in Figure 8 bottom matrices, the rows represent 3 food processing centers and 4 types of food,

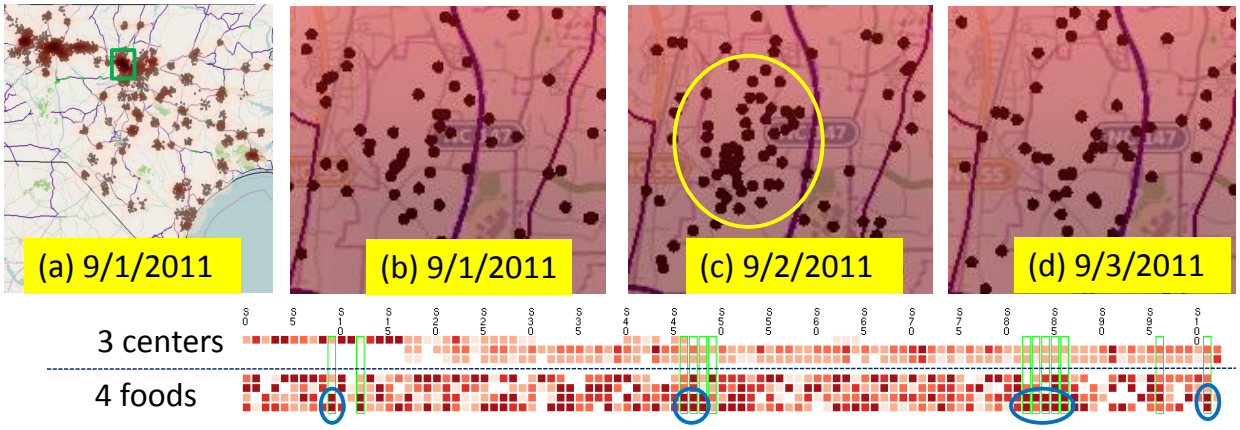


Fig. 8. Ill people caused by contaminated food is presented using a KDE hotspot visualization. In (a), the darker location has a larger number of poisoned people. Brown points mean ill people in the reported location. The locations highlighted by a green box in (a) is magnified in (b), (c) and (d) on different dates. As the timeline shows, the number of ill people increased until 9/2/2011, then started decreasing on 9/3/2011. The bottom matrices show which food processing centers (1–3) were involved and which foods (1–4) were delivered to which store in 8/30/2011, two days before the illness. Here, the restaurants in the light green boxes are the those selected by the thicker green box in (a). We see that a large quantity (darkest red pixels in blue circles) of two foods (third and fourth rows) are commonly provided to restaurants in the area.

and there is a column for each restaurant. Each cell is colored such that the darker the red color, the higher the amount of each product provided. Here, the restaurants in the affected area that are selected in the box in the top-left are highlighted with light green boxes. For stores S9 and S12, only one food processing center provided products, while other processing centers supplied most of the food throughout the network. Upon further inspection, one can determining that product lots in row 3 and 4 are common in most of the restaurants yielding ill individuals. Some example routes are shown in Figure 2(e), where each route supplies 3-4 restaurants. A red bar represents the supplied food and the green bar represents the food consumed at a restaurant. Here, we see that a large amount of the third and fourth foods (blue circles in Figure 2(e)) are delivered and will all be consumed within a few days. Therefore, these two product lots are good candidates for further inspection in tracing back the contaminated food item.

7 Discussion

We have received some initial feedback from various user groups as to the value of the VASA system. Our food supply chain experts helped develop the pipeline and tailor it for their workflow. We have also had very positive feedback from numerous regional and federal government officials on the value of the VASA workbench for use in command centers at the local and regional level for increased situational awareness and the ability for plan for both resiliency and response before and during an event. Feedback from regional Federal Emergency Management Agency (FEMA) personnel is that this system is novel in that it could enable unprecedented work within their organization: visual investigation on large multiple simulation runs and instance approximations under severe weather conditions. They noted that the system enables "The Whole Community" approach to meet the actual needs of residents, emergency managers, organizational and community leaders, government officials, and the general public when extreme weather impacts various societal infrastructures. They felt that the VASA tool would enable each community to make informed and timely decisions about how to manage throughout an extreme weather event. They also suggested extending our system to real-time weather data to respond to all warnings and alerts from the National Weather Service. We have also received similar positive feedback from non-governmental aid organizations.

While the VASA system is full-featured, it may be overkill for simple analyses that only require using a few simulation components. Furthermore, sometimes which simulations to use is not clear a priori, and analysts may have to explore the problem in-depth before they can make a decision. This is also one of the strengths of the VASA system: the VASA Workbench does not stipulate a specific simulation pipeline, but leaves this choice to the analyst. It also provides proxies to estimate simulations prior to a run, and visual and interactive representations of intermediate results. However, it is also true that for a limited simulation involving only a single simulation, using the entire VASA system may be excessive and introduce a lot of overhead.

A more general question is how the VASA approach to interactive computational steering will impact the overall analysis process. Since we have yet to conduct formal user studies with our target audience for the VASA project, it is too early to conclusively answer this question. However, our intuition is that the core benefit of VASA is to introduce interactive visual analytics to a domain that is fundamentally asynchronous and off-line. We speculate that this, in turn, will yield the same kind of rapid, iterative exploration of simulation scenarios that Fisher et al. [14] observed when introducing visualization of partial results to large-scale database computations. We think that this will contribute to analysts wasting less time on configuring their simulation runs and will yield more informed and well-designed results.

8 Conclusion and Future Work

We have introduced the notion of visual analytics for simulation steering within the context of societal infrastructure. To our knowledge, ours is the first to study visual analytics for simulation from a *systems-of-systems* [11] perspective, where multiple heterogeneous—often physically distributed—systems are combined into a unified, more complex system in which the linkages between components provide a sum greater than its constituent parts. This notion transcends individual simulation models and instead chains together multiple high-fidelity simulations into large-scale asynchronous pipelines. The VASA system we presented as a practical example of such an approach is a distributed application framework consisting of a central Workbench controlled by an analyst and a set of loosely coupled simulation components implemented as distributed network services.

Big data simulation is a powerful new tool for data science, and while our work on applying visual analytics to this domain is conceptually complete, it really only scratches the surface of what is possible. Future work on the VASA system will involve integrating even more advanced and detailed simulation components, such as high-fidelity power grid models, gas pipelines, and power plants for energy infrastructure; bridges, tunnels, and causeways for transportation networks; and hospitals, police stations, and fire stations for societal infrastructure. In doing so, we envision designing additional novel visual representations and interactions for configuring these components as well as visualizing their proxy, intermediate, and final results.

Acknowledgments

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under award no. 2009-ST-061-CI0002.

REFERENCES

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 191–200, 2011.

[2] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, June 2008.

[3] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Verlag, 2006.

[4] N. V. Andrienko and G. L. Andrienko. Interactive visual tools to explore spatio-temporal variation. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, pages 417–420, 2004.

[5] N. V. Andrienko and G. L. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.

[6] L. Anselin. From SpaceStat to CyberGIS: Twenty years of spatial data analysis software. *International Regional Science Review*, 35(2):131–157, 2012.

[7] J. Biddiscombe, J. Soumagne, G. Oger, D. Guibert, and J.-G. Piccinali. Parallel computational steering and analysis for HPC applications using a ParaView interface and the HDF5 DSM virtual file driver. In *Proceedings of the Eurographics Conference on Parallel Graphics and Visualization*, pages 91–100, 2011.

[8] B. Broeksema, T. Baudel, A. G. Telea, and P. Crisafulli. Decision exploration lab: A visual analytics solution for decision management. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1972–1981, 2013.

[9] S. Buckley and C. An. Supply chain simulation. In *Supply Chain Management on Demand*, pages 17–35. Springer, 2005.

[10] L. Costa, O. Oliveira, G. Travieso, F. Rodrigues, P. Boas, L. Antiqueira, M. Viana, and L. Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 3(60):319–412, 2011.

[11] D. DeLaurentis and R. K. Callaway. A system-of-systems perspective for public policy decisions. *Review of Policy Research*, 21(6):829–837, 2004.

[12] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara. Interactive focus+context analysis of large, time-dependent flow simulation data. *Simulation*, 82(12):851–865, 2006.

[13] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012.

[14] D. Fisher, I. O. Popov, S. M. Drucker, and M. C. Schraefel. Trust me, I'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1673–1682, 2012.

[15] G. Fuchs and H. Schumann. Visualizing abstract data on maps. In *Proceedings of the International Conference on Information Visualization*, pages 139–144, 2004.

[16] S. Hadlak, C. Tominski, H.-J. Schulz, and H. Schumann. Visualization of attributed hierarchical structures in a spatiotemporal context. *International Journal of Geographical Information Science*, 24(10):1497–1513, 2010.

[17] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 115–124, 2000.

[18] Q. Ho, P. H. Nguyen, T. Åström, and M. Jern. Implementation of a flow map demonstrator for analyzing commuting and migration flow statistics data. *Procedia - Social and Behavioral Sciences*, 21:157–166, 2011.

[19] B. C. Hobbs. *Food poisoning and food hygiene*. Edward Arnold and Co., London, United Kingdom, 1953.

[20] A. Kamran and S. U. Haq. Visualizations and analytics for supply chains. Technical report, IBM, February 2013.

[21] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. Cleveland, S. Grannis, and D. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics & Applications*, 29(3):18–28, May 2009.

[22] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics and Applications*, 29(3):18–28, 2009.

[23] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010.

[24] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert. Visual analytics law enforcement toolkit. In *Proceedings of the IEEE Conference on Technologies for Homeland Security*, pages 222–228, 2010.

[25] A. Malik, R. Maciejewski, B. Maule, and D. S. Ebert. A visual analytics process for maritime resource allocation and risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 221–230, 2011.

[26] K. Matkovic, D. Gracanin, M. Jelovic, A. Ammer, A. Lez, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1449–1457, 2010.

[27] J. D. Mulder, J. J. van Wijk, and R. van Liere. A survey of computational steering environments. *Future Generation Computer Systems*, 15(1):119–129, 1999.

[28] C. Ncube. On the engineering of systems of systems: key challenges for the requirements engineering community. In *Proceedings of the IEEE Workshop on Requirements Engineering for Systems, Services and Systems-of-Systems*, pages 70–73, 2011.

[29] R. Perez. Supply chain model, April 2011.

[30] Z. Pousman, J. T. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[31] H. Ribicic, J. Waser, R. Fuchs, G. Bloschl, and E. Gröller. Visual analysis and steering of flooding simulations. *IEEE Transactions on Visualization and Computer Graphics*, 19(6):1062–1075, 2013.

[32] G. Satell. Why our numbers are always wrong. *Digital Tonto*, October 2012.

[33] G. Satell. Why the future of innovation is simulation. *Forbes*, July 2013.

[34] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[35] S. Terzi and S. Cavalieri. Simulation in the supply chain context: a survey. *Computers in industry*, 53(1):3–16, 2004.

[36] J. J. Thomas and K. A. Cook. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[37] C. Tominski, J. Abello, and H. Schumann. Axes-based visualizations with radial layouts. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1242–1247, 2004.

[38] J. S. Vetter and K. Schwan. Progress: A toolkit for interactive program steering. Technical Report GIT-CC-95-16, Georgia Institute of Technology, 1995.

[39] J. Waser, R. Fuchs, H. Ribicic, B. Schindler, G. Bloschl, and E. Gröller. World lines. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1458–1467, 2010.

[40] J. Waser, H. Ribicic, R. Fuchs, C. Hirsch, B. Schindler, G. Bloschl, and E. Gröller. Nodes on ropes: A comprehensive data and control flow for steering ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1872–1881, 2011.

[41] X. W. William Tolone and W. Ribarsky. Making sense of the operational environment through interactive, exploratory visual analysis. In *Proceedings of the NATO/OTAN Symposium on Visual Analytics*, pages 9–1–9–13, 2014.

An Integrated In-Situ Approach to Impacts from Natural Disasters on Critical Infrastructures

Online ID 185

Abstract—Natural disasters can have a devastating effect on critical infrastructures, especially in case of cascading effects among multiple infrastructures such as the electric power grid, the communication network, and the road network. While there exist detailed models for individual types of infrastructures such as electric power grids, these do not encompass the various interconnections and interdependencies to other networks. Cascading effects are hard to discover and often the root cause of problems remain unclear. In order to enable real-time situational awareness for operational risk management one needs to be aware of the broader context of events. In this paper, we present a unique visual analytics control room system that integrates the separate visualizations of the different network infrastructures with social media analysis and mobile in-situ analysis to help to monitor the critical infrastructures, detecting cascading effects, performing root cause analyses, and managing the crisis response. Both the social media analysis and the mobile in-situ analysis are important components for an effective understanding of the crisis and an efficient crisis response. Our system provides a mechanism for conjoining the available information of different infrastructures and social media as well as mobile in-situ analysis in order to provide unified views and analytical tools for monitoring, planning, and decision support. A realistic use case scenario based on real critical infrastructures as well as our qualitative study with crisis managers shows the potential of our approach.

Index Terms—power simulation, visual analytics, critical infrastructure, mobile in-situ analytics, homeland security.

1 Introduction

Responding to the destructive impact of a volatile hurricane to a network of critical infrastructure is the central challenge for emergency responders. After witnessing the devastating destruction from Hurricane Sandy during 2012 and the flood disaster in Germany 2013 decision makers are on high-alert for threats to their critical infrastructures such as power lines, food networks, shelters, etc., potentially caused by impact from another natural catastrophes. Important backbones of our society are electrical power networks since the electricity supply has a strong impact on the fundamental societal structures such as life/health, environment, and economy. Especially, electric power systems are increasingly dependent on information and communication technology (ICT) systems as new monitoring, control, and protection functions, especially in the currently emerging Smart Grid installations. In order to deal with the increasing vulnerabilities of electric power systems, advanced ICTs, including network-based Supervisory Control and Data Acquisition (SCADA) systems or Wide Area Monitoring Systems (WAMS) have been deployed by the power industry.

Analyzing the vast amount of information from different domains is a complex analytical issue. The monitoring of the interconnections between power grids and digital networks requires the integration of several data sources. With an overview the crisis manager is able to understand and explore the crisis allowing her/him to project the future development and to make decisions. Situational awareness is important on all levels of crisis response that range from central command centers to site-commanders and boots-on-the-ground. All levels must be able to access the information of a crisis. Further, they need to communicate bottom-up or top-down since crisis managers typically rely on the aggregated information of the field and first responders lack context information.

Novel public communication platforms like social media services and other Web 2.0 sources have established a completely new information channel that can help to improve situational awareness for the decision makers. Citizens affected by critical events often report vital situation related information directly to messaging services like Twitter or Facebook. They use mobile and sometimes even GPS-enabled communication devices like smartphones or tablet computers. Gathering useful information pieces from the vast amounts of random unrelated chatter poses a completely new challenge for analysis and decision support systems.

Existing tools and systems do not support the integration of information over several critical infrastructures such as power grids and the ICT networks. The monitoring and understanding of the relationship of critical infrastructures and the coordinated management of their failures is therefore one of the biggest challenges in critical infrastructure protection and crisis response. In this paper, we present a system that supports all levels of command structures and enables situational awareness for crisis response. This system was developed within a nationwide interdisciplinary project [blinded for review] running for three years with an international research collaboration to a partner project [blinded for review].

Our contribution: We present a visual analytics system that: 1) supports all levels of crisis response with specialized equipment and visualizations for control rooms and mobile devices; 2) combines multiple critical infrastructures and social media by information abstraction; 3) enables interactive simulation and visualization of the subsequent development of a crisis; 4) enables interdisciplinary and distributed teams to understand and react on crisis situations.

2 Related Work

Visual analytics provides technologies that support the decision maker to gain a precise overview of the current situation by (1) automatically filtering irrelevant information, (2) presenting relevant data visually to optimize understanding, (3) optimizing the communication process among the stakeholders, and (4) supporting the analysts formulate and assess alternative solutions. Visual analytics fulfills the task to bridge the gap between several networks, especially in vision of the analysis of cascading effects. The combination of automated data processing and interactive visualization techniques serves as a means to cope with the complexity of the selected task. For instance, the advantage of such tools has been illustrated in an analysis of the 2005 outbreak of the avian flu by combining different analysis capabilities [17]. This scenario shows the power of an analytic setting that supports the analysis of complex, real-world scenarios. Also, first visual analytics tools in the area of crisis response were the result of SoKNOS [9]. This comprehensive environment requires integrated visual and traditional systematic analysis of massive data, including improved strategies for exploratory visual analysis, hypothesis testing and user-specific presentation of relevant information as a basis for actionable decision making. Furthermore, visual analytics tools for analyzing syndromic hotspots are presented by Maciejewski et al. [12] and allow the analyst

Submitted to IEEE VAST 2014. Do not redistribute.

to perform real-time hypothesis testing and evaluation.

Much prior research has focused on using simulation and predictive modeling to anticipate hurricane movement and suggest possible landfall and impact locations [23, 16, 19, 5]. The challenge of understanding these modeling efforts and their predictive capabilities from large collections of sensing and simulation datasets is ubiquitous. However, for such computation driven practice it is detached from the human-decision making process. It is important to gain enough insights to actionable knowledge, but the development of models and analysis of modeling results usually requires that models be run many times [4].

As part of the PSERC project, techniques are developed to visualize complex power systems and flows [15]. For the interactive analysis of network related data sources such as server logs or BGP protocol data, Fischer et al. developed a visual analytics expert system in [6]. A detailed overview of cyber security and privacy issues in a smart grid context is presented in [10]. The control room is the central part of a system, where all information comes together. Notifications and alarms are collected and transferred to a central node. These events are typically evaluated by rule-based systems, where the rules are defined by domain experts. Rooney et al. gives an introduction to Root Cause Analysis (RCA) in [18]. There are also systems that use simpler fuzzy logic rules as a vehicle that allows engineers to incorporate human reasoning in the control algorithm [13].

Social media, such as Twitter and Facebook, contain time-critical information that can enhance situational awareness. First approaches for building and improving decision making systems in this domain were introduced by Tomaszewski et al. [22]. MacEachren et al. [11] developed a visual analytics tool that allows for querying social media sources and depicting aggregated results on a geographical map. Thom et al. [20] present a novel cluster analysis approach to automatically detect spatiotemporal anomalies in Twitter messages.

Evacuation of large urban structures, such as campus buildings, arenas, or stadiums, is of prime interest to emergency responders and planners. Although there is a large body of work on evacuation algorithms and their application, most of these methods are impractical to use in real-world scenarios (non real-time, for instance) or have difficulty handling scenarios with dynamically changing conditions. Visual analytics involves effectively combining interactive visual displays with computational transformation, processing, and filtering of large data [21]. One focus of visual analytics is real-world problems involving situationally aware decision support. Andrienko et al.[1] described the sequence of tasks that are fundamental to exploratory visualization of spatio-temporal data. They incorporate these ideas into effectively demonstrating the movement of storks across a geographical area. Campbell and Weaver[3] investigate situational awareness during emergencies using two different tools: RimSim Response! (RSR) and RimSim Visualization (RSV). The work of Kim et al. [8] focused on the use of mobile devices for situationally aware emergency response and training, and thus, their approach is similar to our work. They demonstrated their system with an evacuation simulation of the Rhode Island club fire of 2003.

The discussed approaches and systems focus on a specific domain and do not combine various external data sources. Integrate sensor data (electricity, weather, supply), social media and in-situ analysis is a challenging task. Furthermore, most of the current systems are intended for domain expert users, although crisis management teams may consist of interdisciplinary members. Enabling interdisciplinary teams to analyze and understand interdependent data leads to efficient crisis response.

3 Design Framework: Target Users & Requirements

Large scale emergency response has a command structure as illustrated in Figure 1. A large police or fire response, for example, will have a site commander who deploys first responders. If the emergency is larger and more wide-spread, there will be a command center with oversight over multiple site commanders. A similar structure applies to breakdown of the electrical grid or other critical infrastructure. The deeper one goes into the command structure, the more mobile the responders are; they are focused on the locales, tasks, and decisions at hand and traditionally don't have contextual understanding or situational awareness. Typically, site commanders also don't have situational awareness (in terms of the deployment of their personnel and what they are doing or seeing, for example) nor do they have the context to make the most effective decisions. To make decisions, the first responder will want to know what is happening in the locale, what is about to hit where and when. The site commander will want to know similar things over a wider locale and must in addition organize and manage a group of responders. Aspects of all this should flow up to the command center for overall decision-making.

Networked mobile visual analytics can be an essential part of this contextual and situational awareness. When we have developed and deployed networked mobiles systems for emergency response in urban areas, worn or carried by first responders, police and emergency managers have immediately seen the power and usefulness of such systems. It is in this spirit that we have developed the networked mobile visual analytics interface as part of crisis response. It connects the responders with the command center so that everyone resides in the same context. However, even if they lose the network connection, they can carry enough of the context on their mobile devices and can take effective action. Further, any feedback and updates they provide will be transmitted to the central system. Detailed location, movement, and action updates can be placed in the appropriate spatio-temporal context so that commanders can see, in unprecedented detail and without ambiguity, what is going on (and responders can see, minute-by-minute where there fellow responders are and what they are doing). Specifically, our system is designed to accommodate the three essential personnel in a crisis respond scenario:

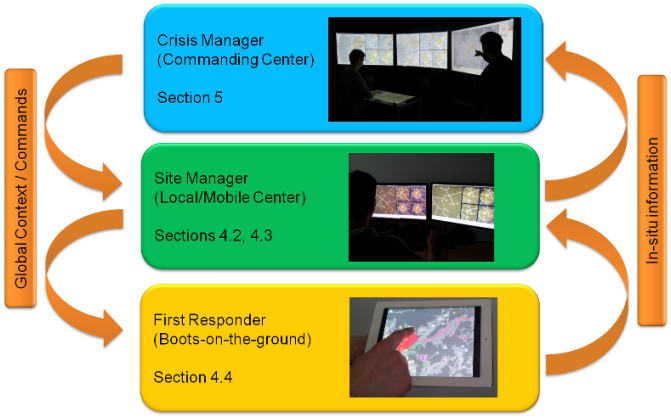


Fig. 1. Overview of the Users that our system is designed to facilitate.

Crisis Manger: Crisis Managers are a group of domain experts who oversees the entire emergence response process. This group is typically formed by interdisciplinary members from various analytical domains, such as grid operators from power companies, city-state officials, evacuation experts, and people from federal departments. Their objective is to understand and assess the severity of the crisis situation, select corresponding Site-Commanders with appropriate First Responders, and provide Just-In-Time decisions making based on inputs from social media and in-field communications. The natural of the heterogeneous data inputs for the Crisis Managers determined that they will benefit from a control room setup, where they will be provided with a combined overview of the crisis situation. As detailed in section 5, our setup supports distributed as well as collaborative analysis, provides overviews of the development of the ongoing crisis, and it further enables the Crisis Managers to interactively deploy and arrange response effort, and receive the communications from site commanders and real-time situation updates.

Site Commanders: With the advanced mobile technology (e.g., mobile emergency response vehicles), site managers are the critical link between Crisis-Manager at control center and First Responders in

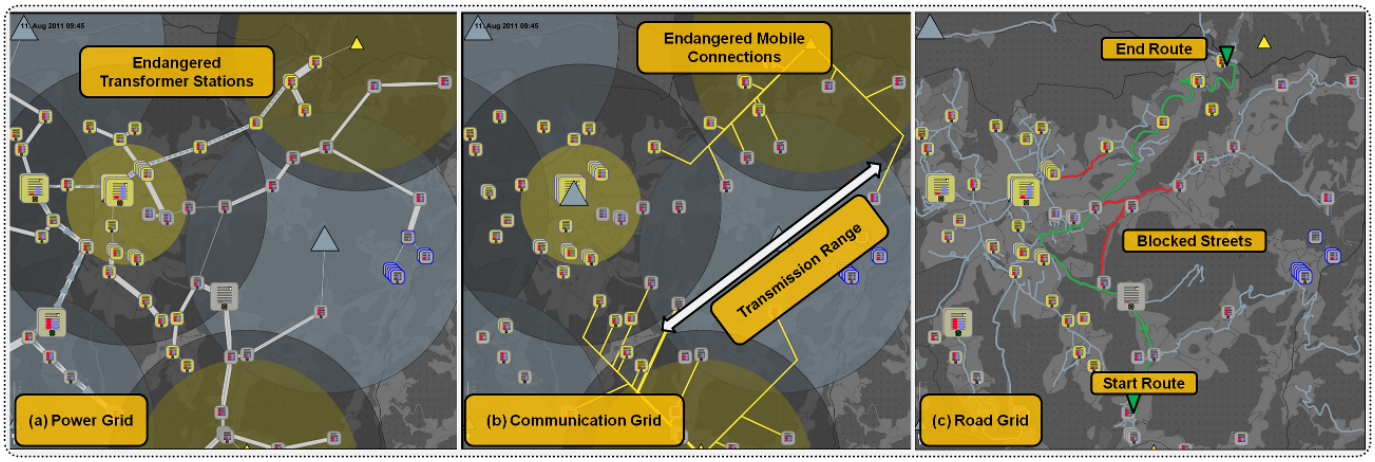


Fig. 2. Overview of the power, mobile, and road grid. Transformer stations (rectangles) are connected via power lines and are also connected to the communication infrastructure (triangles), which transfers the information to the central control room. The transmission range of the mobile stations is visualized as concentric circles. While gray indicates normal operation mode, the yellow elements on the screen reveal a severe situation. High deviations in voltage cascaded from the energy grid into the mobile grid due to failures of the power supply. Now, the operator must intervene immediately. With malfunctions in the mobile grid, the crisis response commands from the control room won't reach the field, which then would result in a black out. Further, roads are blocked and hinder the response team to reach the target area, which is visible in the command center after the first responders update the street status. The dynamic routing adapts to these constraints and calculates a different route.

the field. Based on our previous collaboration with local emergency responders, Site Commander (e.g., Police Chief) are often stationed near the crisis center where first responders were deployed to conduct on-site instructions and in-situ communications. At the mean time, they are relying on the mobility of the technologies to maintain an open communication channel with control center for further situation assessment and updates. Site Commanders act differently from Crisis Managers in a way that they have a more focused missions in a specific area that is assigned to them (e.g., a specific substation or a street blocks) and are in charge of provide real-time response as the crisis unfolds in the field.

First Responders: First Responders are the group that fights the diseases right in the center of where crisis occurs. These teams consist of various professionals, such as policeman, fire fighter, and power grid responders. These interdisciplinary group mainly conduct response effort in the field with instructions from Site Commanders. Their extreme needs of mobility determined that we need to provide them with a mobile-device based visual analytics system. Key functions in this system, as detailed in section 4.4, includes instructions that informs them about the areas that they need to focus their attentions, interactive methods to depict areas to prioritize their tasks, and finally communication methods to provide updates and situation reports to their Site Managers. All this information need to be shared through wireless networks that directly feedback to the Site Commanders and further to Crisis Managers.

4 System Components

4.1 Simulation of Critical Infrastructures

Large scale natural disaster, a cyber-attack, or other wide spread crisis may affect multiple infrastructures. To capture these complex, multifarious, and dynamic effects, we utilize a simulation model [blinded for review] that takes into account the interrelationships among critical infrastructures. The simulation is built within a rule-based framework for integrating multiple infrastructure components at a high level. The interlaced critical infrastructures are captures in a set of networks with each node having a set of properties according to its category and the edges providing a dependency rule according to the category and state of the two connected nodes. This results in a dependency/interdependency ontology (e.g., as illustrated in Figure 2(a) mobile transmitters are connected to transformer stations). Thus, for example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a school were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down. Likewise, telecommunication nodes would switch to backup power that might also shut down after its prescribed duration. Thus, the simulation also takes cascading events into account.

In some natural catastrophes some roads may be affected and rescue or response teams, especially with heavy gear, are not able to pass them. This has to be considered in the evacuation and logistic management at all three levels: E.g., the crisis managers must plan the logistics of gear and troops; the site-commander sends in first responders on different routes to the crisis and first responders will update the status of streets if they are not passable. Our system supports dynamic routing with the state-of-the-art algorithms such as Dijkstra's algorithm that consider to the current status of streets.

4.2 Visualization of Interdependent Infrastructures

A smart grid (energy network) typically consists of power lines at different voltage levels connected by transformer stations. These stations distribute the power over regions and supply streets, households, and industrial facilities. In our scenario, a mobile communication grid transfers information and control commands from the central control room to the electrical grid. The mobile transmitters itself are power-supplied by common transformer stations and thus, the infrastructures are tightly interconnected.

4.2.1 Information Abstraction & Visual Encoding

The complex and vast amount of information of each infrastructure is abstracted and reduced. This enables the decision maker to understand the full crisis in its context and to detect potential cascading effects. Every infrastructure is abstracted to an undirected graph. Its nodes are represented by symbols, such as rectangles for transformer stations and triangles for mobile transmission stations. The graph edges represent the domain dependent connection between infrastructure elements, such as power lines and mobile communication connections (see Figure 3(b) and Figure 2).

The status of each element is estimated by a state-evaluator model that is defined for each infrastructure. These models concern the actual information of the field, such as utilization and durability (see Figure 3(a)). We use a prediction module for our power grid that predicts the consumption and production at each transformer station according to weather forecasts and past data based on Monte Carlo simula-

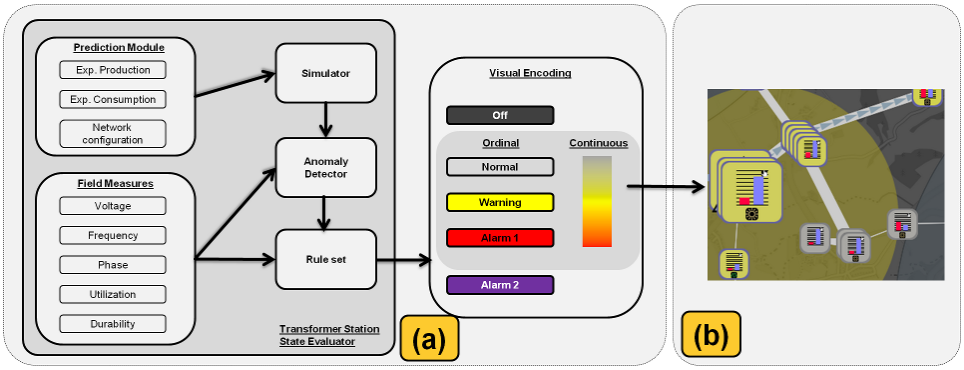


Fig. 3. (a) Station state evaluators consider the incoming measurements and the comparison to the expected behavior, which reveals anomalies. A set of rules maps this input to color that expresses the status of the element. (b) Domain details are added to the symbols such as power consumption and production (red and blue bars), as well as producer types (photovoltaic or biogas) for transformer stations.

tion. This information is sent to the simulation server that simulates the subsequent development. This "expected" behavior is compared to the actual measurements. Thus, anomalies are detected, which may reveal damaged or harmed devices. Further, the subsequent development can be visualized as small multiples (see Figure 4) in addition to the monitoring views (see Figure 2).

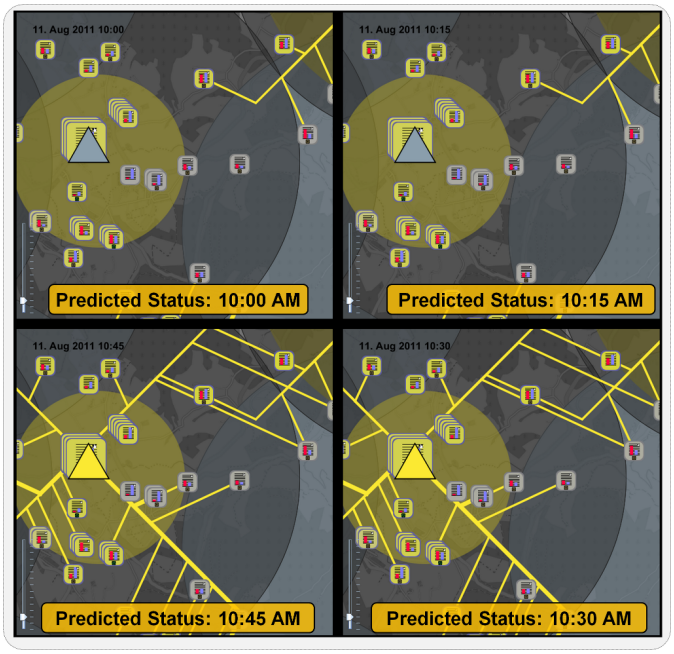


Fig. 4. Subsequent Development: If there is any emerging problem in the future, the prediction view will show the future status of the network by small multiples, which highlights the remaining time and the elements that are affected.

A set of rules maps the input of the field and anomaly detector to color. Saturated and intense yellow, red, and violet represent warnings and alarms. Less saturated colors stand for less serious events, such as gray for normal (uninteresting) status. Some rules provide continues values in addition to ordinal signals. For these rules we use a continuous color scale that varies over saturation and lightness from gray to yellow and over hue from yellow to red. Thus, severe events are perceptually highlighted on the dark background whereas less important events do have less visual impact [24]. The size of elements represents the topological importance of infrastructure elements. We consider central elements (and their dependencies) more important, since their failures are more acute than failures of border elements. Thus, the size of important elements is increased, which also highlights dangers or failures of central elements. We also add domain details into the symbols such as the current power production and consumption as well as the producer type for transformer stations (see Figure 3(b)).

4.2.2 Zoom, Focus & Details on Demand

Two major problems arise when graphs are visualized: the over plotting of nodes and intersection of edges. The over plotting of nodes can be compensated by stacking the overlapping nodes and visualizing them at their average location. They are sorted by their current status. The domain details (e.g., power consumption and producer type) are aggregated and visualized in the foreground element. For the intersection of mobile communications, we omit the painting of connections that are working properly and use edge bundling in order to avoid intersections. These aggregation techniques enhance the readability of the visualization, however, at the cost of hiding or divert some information. The user is therefore able to zoom into areas of interest. If enough space is available, the system will visualize the elements in their normal layout and will provide additional information (see Figure 5).

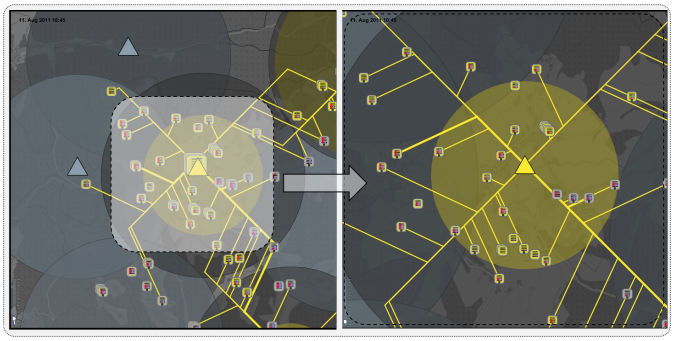


Fig. 5. Semantic zooming reveals more details on each zoom level as soon as enough space is available.

The user can further interact with the field via control panels, e.g., disabling powerlines or deactivating producers at a transformer station (see Figure 6). The user is further enabled to adjust the expected production and consumption for simulating alternatives and thus, can create several alternatives for decision making.

4.3 Accessing Human Sensor Information

With the rise of community driven content services, such as Twitter and Facebook, a new information channel for situation awareness has been established in the Web. In contrast to more traditional data sources, like structured sensor data or detailed reports from emergency responders, these new information channels pose novel requirements

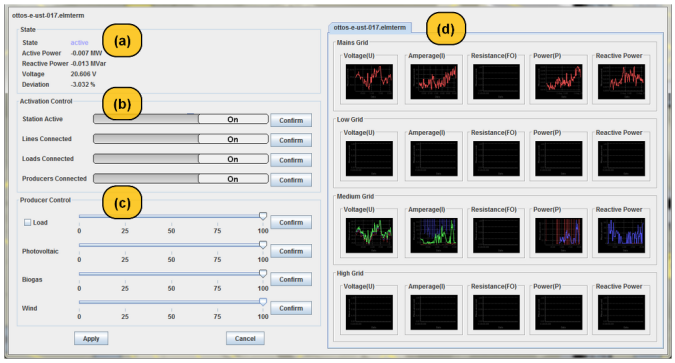


Fig. 6. Control panels visualize the actual measurement data (a). (b) Controls enable the operator to switch elements on and off. With (c) the operator can simulate or control the amount of energy consumption and production at a transformer station. (d) Charts visualize the development over the past hour / day.

for data filtering, ranking and aggregation. The relevant information has to be separated from general chatter and organized according to different topic categories. Large amounts of repetitive reports have to be integrated into a consistent and scalable situation overview. In our approach we propose novel methods to address these challenges in order to incorporate social media services as external community driven sensors within the command center environment.

4.3.1 Overview and exploration based on automated event identification

The complexity of events and the velocity of streaming data often hinders straightforward situation awareness during critical situations. Means for automated detection and display of possibly relevant clues can be a key factor in successfully mastering crisis management. In case of social media data, it is particularly important to detect possible first-hand accounts (e.g. eyewitness information) of on-going situation between large quantities of irrelevant information and to provide visual representations of the discussed topics and observations.

As in [20], we rely on the presumption that messages addressing local events are often of related content and structure and that they are furthermore located in a spatial and temporal neighborhood. This ultimately leads to spatiotemporal clusters of messages reporting on the same situation related topics and keywords. Based on a cluster analysis approach, adapted to the specifics of real-time data, we automatically detect such spatiotemporal anomalies in the continuous data stream. Once a timeframe and geographic region is interactively selected by the analyst, the system generates a map of detected anomalies within that region and timeframe by finding frequent keywords in the message clusters and place them as labels at the corresponding cluster locations on the map. In order to avoid overlapping labels and at the same time show the analyst as much information as possible, we apply a collision avoidance technique that allows overlapping labels to move small distances from their designated locations. Ultimately, the label is not shown on the selected zoom level, if a certain maximum distance for that zoom level has been exceeded.

Our technique provides a broad overview of all events that occur in a given geographic region and, more importantly, an indication of keywords and topics that might be a good starting point for further investigations (see Figure 8). This is particularly helpful if the analyst does not know in advance what to search for or to initially inform him of an unknown ongoing situation. By zooming into the map, our layout technique automatically provides more labels for the given area, as more screen space becomes available for the given region. The analyst thus receives more detailed indication of possible sub-events connected to a larger event and can use this as a basis to extend his investigation with traditional textual search, content analysis and focus and context visualizations.

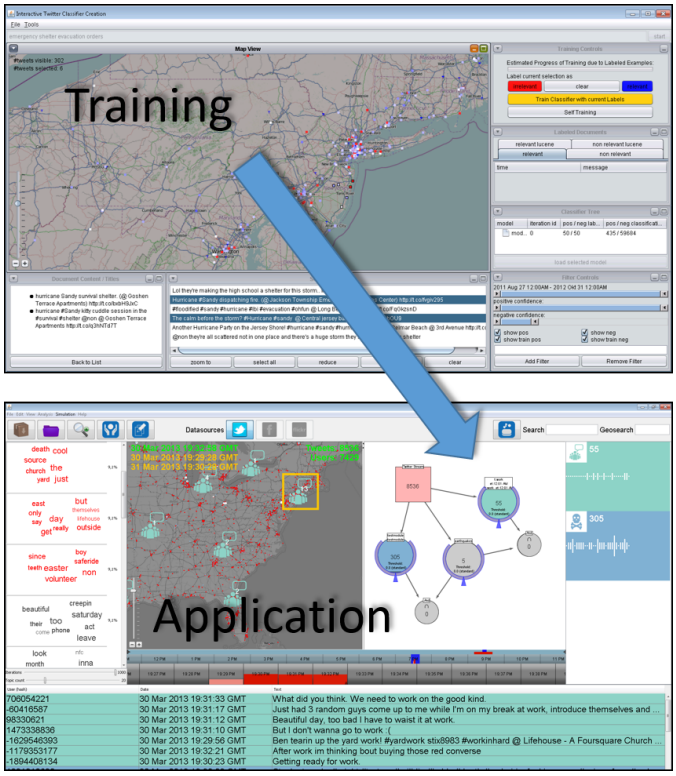


Fig. 7. SVM classifiers are trained based on existing social media data to enable detection of messages related to specific information needs. A visual tool supports the analyst by exploring the content and metadata of historic messages in order to find good training samples. The SVM classifier is trained by selecting and labeling sets of positive and negative samples. The outcome and performance of the classifier and the progress of the training are continuously visualized on the map and associated statistics. Bottom: Classification and filtering of crisis related messages. Based on his specific information needs the analyst can load and combine modules from a library of the pre-trained classifiers using set operations. The occurrence of all new messages is shown in real-time. The analyst can associate the modules in the filter combination with specific labels, colors and symbols that are then used to display messages detected by that module

4.3.2 Detection of highly relevant information items based on user-steered classification

Besides the need to be informed about unknown or unexpected events, analysts usually also have a distinct domain and area of responsibility and are thus able to define information types that are clearly relevant to their tasks. For example, police officers will always be interested in information about the use of firearms or other acts of violence in their precinct. However, plain keyword-based approaches to find messages fitting to the given information need are often not powerful enough, as the complexity and specifics of language use in social media data can often not be properly reflected.

Especially in real-time analysis scenarios analysts need means to quickly build highly customized filters based on their information needs, their knowledge structure and the specifics of the situation. In our approach we propose a two-step process where a library of Support Vector Machine (SVM) classifiers customized to the specific information needs is trained first, which can then be adjusted and combined with each other and with more simple keyword-, spatial, temporal-, spam- and other filters based on interactive visual set operations (see Figure 7). This idea has already been introduced in [2].

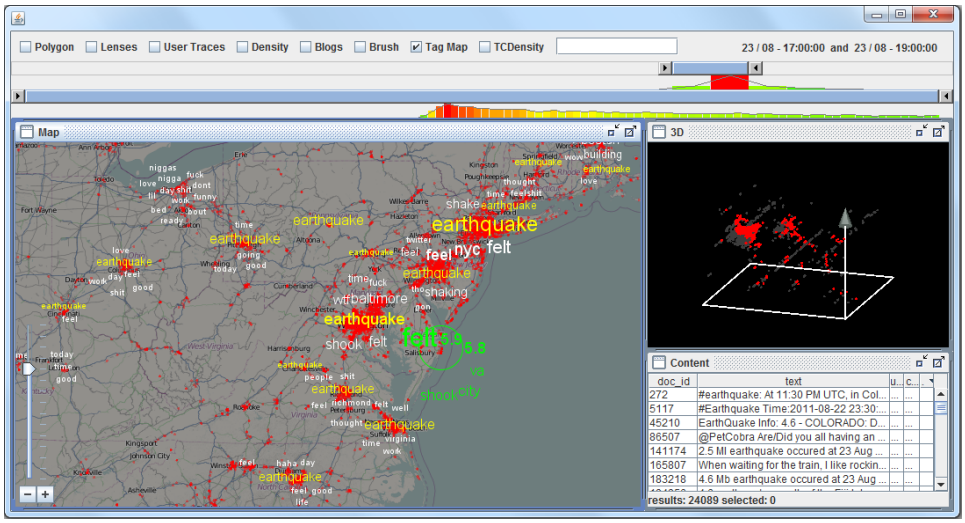


Fig. 8. Overview visualization of crisis related topics based on automated anomaly identification. The image shows the social media component with activated anomaly visualization during a large earthquake that happened in August 2011 near Washington DC. The observation of earthquake related events lead to several larger ”earthquake”-clusters in many cities along the east coast. Zooming into the map shows more tags in the respective area and also reveals smaller sub-events that were connected to the larger event. For example, several buildings in Washington were evacuated which resulted in the prominent topics ”building” and ”evacuate” in the area. Also, the Washington monument experienced damage from the vibrations, which was also observed and reported by several eyewitnesses, leading to the topics ”monument” and ”damaged” being shown on the map.

4.3.3 Classifier Training

Based on historic data of previous, well understood events, an analyst can explore social media messages to label positive and negative examples for a given event type. This is supported by a range of exploration and analysis tools. First, the analyst can use an integrated textual search engine in order to establish an initial exploration context of social media messages related to the event type. E.g., the analyst can use keywords like “power”, “outage”, “breakdown”, “damage”, etc. within a recorded dataset where a known larger crisis like a storm or earthquake has happened. Next, the analyst can use an exploration lens, a movable focus and context tool, which shows the most prominent terms in the region underneath it, in order to find hotspots of the given event. Once the analyst has identified a sufficiently large set of example messages related (positive) and unrelated (negative) to the event type, the analyst can label them as such to iteratively progress the semi-automated training process. The training examples are especially useful if they are near the SVM-classifiers decision border, i.e. they have a high probability of being relevant to the topic in terms of keywords, and just the specific combination of terms renders them related or unrelated (e.g. “This morning the power went down.” vs. “This morning I have no power to get out of bed”). In contrast to traditional active learning approaches, the progress of the training is constantly presented to the analyst by color coding of all messages and by visualizing the decision boundary in a spatial representation. By iteratively exploring and investigating the message set based on these visualizations the analyst can understand the effects of his training and assess the current performance of the classifier.

4.3.4 Real-Time Monitoring

With repeated classifier training analysts can create a comprehensive library of annotated classifier modules for different event and message types relevant to their domain. In a real-time analysis scenario they would usually load and configure a range of classifiers and filters at the beginning of the monitoring period. The classifier modules are arranged in a graph visualization and can be combined with each other and with ad-hoc defined keyword, spatiotemporal- and general purpose spam-filters using Boolean operations (AND, OR, Sym. Diff., etc.). Furthermore they can be modified by increasing the precision or recall of the detection using interactive controls and the most relevant classifiers can be tagged with custom selected colors and symbols in order to highlight corresponding messages if they have been detected. If an unusual situation unfolds, the analyst can also apply the filter- and classifier management to reflect and validate quantitative hypothesis using message volume indicators provided for each module.

4.4 High Mobility Visualizations & Visual Communication

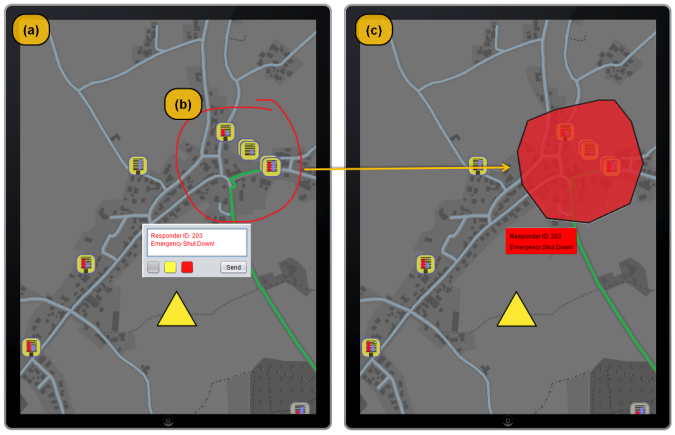


Fig. 9. Overview Mobile Concept for First Responders.

Mobility for the First Responder is a crucial aspect in their field of work. The absolute necessity of providing mobile system to campus police and first responders was observed during previous evacuation exercises [7]. First Responders are no longer satisfied with office-based work stations, but are demanding interactive mobile analysis techniques to carry out in the field. This doesn’t strictly refer to mobile devices, but includes movable equipments in general.

Therefore, we designed a network visual analytics system that utilizes the advancement of mobile devices (e.g., iPad). Our mobile interface aims to provide an interactive environment where the First Responders would be able to receive detailed information in addition to commands from Site Commanders and Crisis Managers, examine the crisis scenario around them, conduct search and research with clear routine information, and finally provide feedback information to the

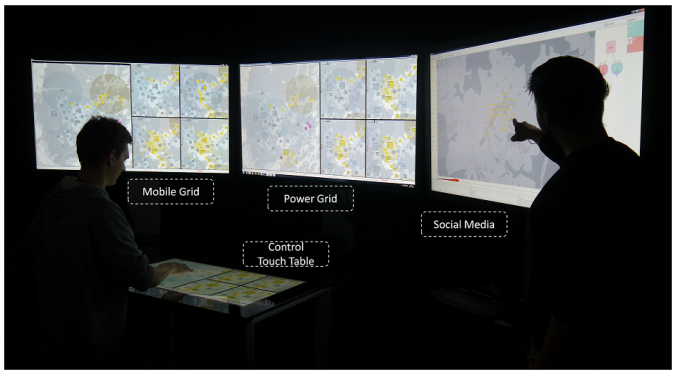


Fig. 10. Our control room setup consists of three high resolution displays for the visualizations and a touch table for the steering of clients. At any time it is possible to add further clients.

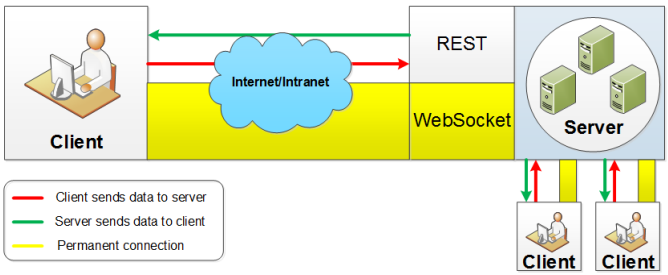


Fig. 11. Client-server architecture for the synchronization of different views and simulations results on multiple clients.

managers. They have access to the visualizations and information of the command center, which can be focused to their particular location and interest. Section 5 discusses details of our implemented architecture to support these functions.

To help users quickly select and focus on a geospatial region, we developed probing gestures, as shown in Figure 9 (B). This is an extremely important analysis feature because it allows the user to drive the analysis and focus on what is important to their needs on the go. A First Responder can, with their touch enabled glove, directly draw onto the map with his or her finger by drawing a bounding area around a region or mark specific points. The system samples the gestures and computes the convex hull or straight line with linear regression, if demanded. Thus, rapid and noisy drawings are smoothed (see Figure 9(c)). They can further annotate in the selections with real-time updates, as shown in Figure 9 (c), and share the information back to their commanders and other responders, through a fast wireless or satellite connection.

5 Control Room

The different components are combined in a control room setup that synchronizes all views and clients on demand and allows the integration of mobile devices.

5.1 Concepts

Crisis managers, often consist of several experts from different domains. A common way to analyze crisis scenarios is the subsequent analysis of incidents. Typically infrastructures build large graphs and therefore, it is not possible to limit the analysis to a single screen. Hence, we setup a control room (see Figure 9) that supports a distributed and collaborative analysis among several experts. Our setup consists of three high-resolution displays and one touch table.

We see four advantages in this setup: First, single experts use this setup to explore and explain incidents with the aid of different views and visualizations on the crisis scenario as illustrated in Figure 2; these are displayed on the three high resolution displays. Second, the setup can be used to illustrate alternative solutions for decision making: Multiple alternatives and their subsequent development can be visualized simultaneously, which supports experts to draw decisions. Third, if several experts synchronously use this setup, the work can be partitioned on the three displays as well as on the touch table. Every expert receives his own interaction device, for instance a cordless air mouse, which is applied to his own workspace. In case an expert needs to exchange information or enhance visualizations, we offer the possibility to synchronize the clients. Fourth, this setup easily enables a possible combination of the previously named social media component and critical infrastructures.

5.2 Architecture

In order to enable this vision of a distributed and collaborative environment, there is need for a supportive system architecture. In a collaborative working environment, multiple views and information have to be synchronized. We therefore set up a client-server architecture that supports synchronization across different clients. Our central server manages all connections to the clients and distributes information. Every interaction that needs to be synchronized is first sent to the server. However, the clients do not necessarily need to be synchronized and can also work independently if requested. In addition, the server also handles all connections to the external simulation servers and the local data sources (power grid, mobile grid, weather, and geography). The system requirements include hardware and system independence. In this distributed environment every client running a JAVA VM is able to connect to the server.

Protocols: The communication is established by web sockets and REST interfaces. The web sockets are permanent connections via TCP protocol, whereas the REST-services are based on explicit HTTP requests. Each client contains multiple sockets that will inform the client if some parameters, views or synchronization are changed. In this case the clients uses the REST interfaces to request explicit updates from the server. If a user requests different configurations or sends information, the client uses the REST interfaces to update the server, which then will update other clients if necessary. This keeps all connected clients synchronized and minimizes the traffic to and from the server, since only packages that are currently needed in the particular view of the client are transmitted.

Mobile Devices: Although overall function goals are straightforward, creating a way to visualize large amounts of infrastructure events while also providing interactive analytics is still a complex challenge. Largely this challenge is due to the number of infrastructure elements we need to consider versus the inherent hardware limitations of mobile devices. To accommodate the essential functions, our mobile design is heavily built around network optimization and scalability. We utilize a low-level graphics engine specifically built for rendering the graphs and provide interactive selection and drawing. It enables the device to render vast data sets very quickly. The graphics engine uses a scene graph style data structure for storing elements. It also utilizes graphics techniques like double buffering and constructing texture atlases on the fly for improved rendering. In addition, we utilize the multicore architecture built in the modern mobile devices to maximize its computing power. We utilize parallelism for all our REST requests to create asynchronous data pulls. We also load all data from the CPU to the GPU through a background processes. We purposely leave the user interface to its own separate process in order to not lock that thread.

While mobile OS incorporates mapping services, they are not specifically designed for displaying large data sets or any complex 2D drawing. Our customized tile map system, as shown in Figure 9 (A), bears the similarity with commercial mapping system, but also provide additional functionalities that fits our visual analytics needs. Since the map tiles are stored in the cloud and the visualization components are implemented as layers that can be stacked onto the map, our mobile interface remain lightweight. This approach also enables a smooth

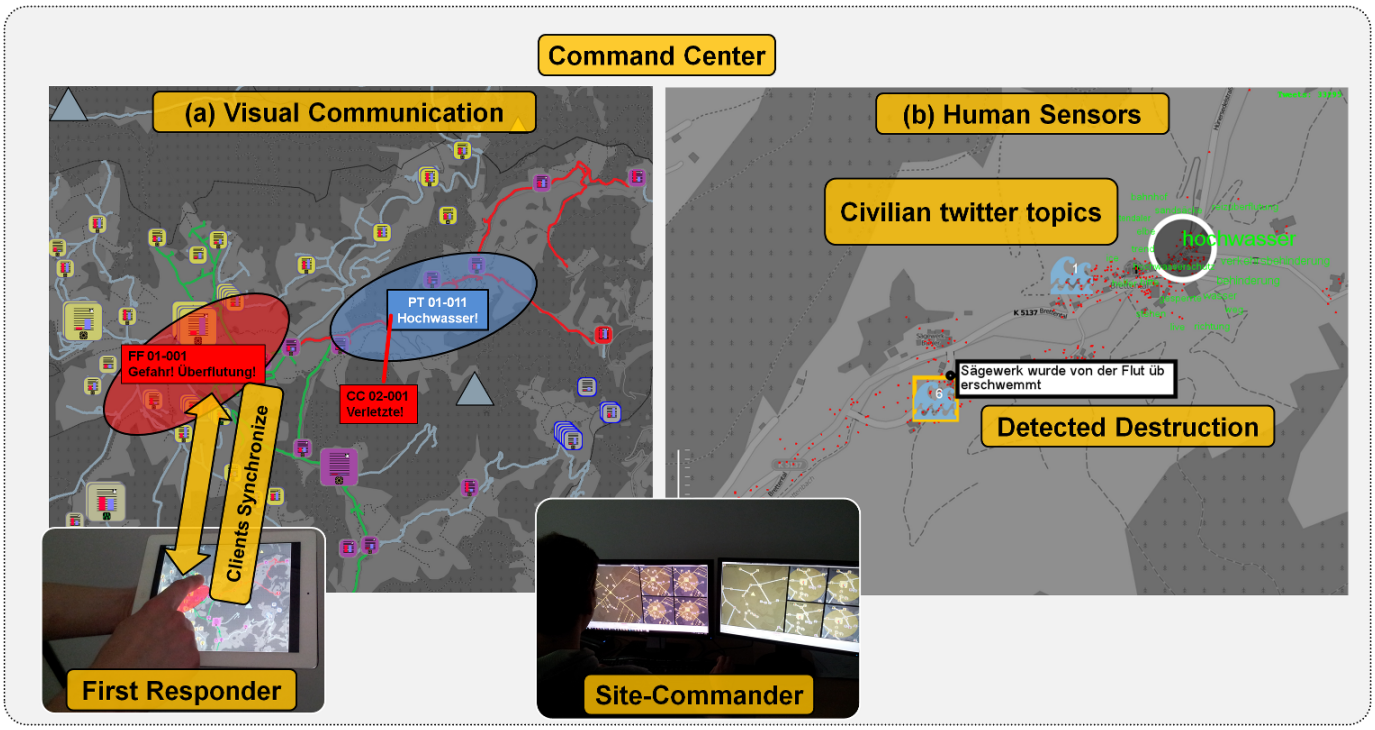


Fig. 12. (a) The north-eastern part is flooded, which results in partial blackout (violet stations) and endangered power grid. The first responders update the in-situ information about flooded areas and casualties ("Gefahr! Überflutung!", "Verletzte"). (b) The social media analysis early reveals messages about high water levels ("hochwasser") and detects destruction events in real time.

transition of updates from Control Room to First Responders in near real time.

Hardware issues: We further support devices that are too weak to render our applications. Depending on their hardware, clients are classified into *complete* or *minimal* clients. *Complete clients* contain enough resources for standalone applications that renders the components by themselves. Therefore, the server needs to transfer data and information, used for synchronization, to these clients. *Minimal clients* do not have enough resources for the whole application and therefore, the server pre-renders the current view of the client according to the device, which is then send and visualized as image. Basic controls are also available in the *minimal clients* such that the image contains the location and type of controls. Thus, *minimal clients* are not updated in real time, however, they have access to the full crisis scenario.

6 Scenarios

Three scenarios were designed to highlight the need for such systems. Therefore, we designed multiple catastrophes that affect critical infrastructures, such as a mass disease, a cyber attack, as well as a flood-scenario that is presented here. The flood disaster is caused by heavy rain and thunderstorms in a region of Germany. The region was already flooded in 1987 and the scenario is inspired from these events. The region is employed with a (model) smart grid by one of the project partners.

The scenario starts with heavy rain and thunderstorms. Especially the high grounds of the scenario region are soaked and the soil already begins to be unstable. A thunderstorm in the early morning hours increased the danger for this area, which alarms the command center and site-commanders of the power grid domain. Due to the social media analysis, which detected messages about unusual water levels, the site-commander sends a power grid technician (first responder) to the region for on-site information. Figure 12 b) shows the detected messages of amblers about high water levels ("hochwasser"). Also the command center reacts with alarming the regional response teams; in this case regional fire-fighters that are equipped with specialized *mobile analysis devices*.

A debris avalanche hits a small village in the suburban region. Many homes are flooded and separated from the center of the region. The debris avalanche also hits a bridge and the flood is jammed. Transformer stations in the eastern parts are immediately destroyed and the blackout cascades from the east to the south (violet stations). The remaining power circuit in the west and north is suffering from this immediate loss of consumption, which raises the voltage levels (see yellow stations in Figure 12). The fire-fighters cannot reach the casualties since streets and fields are not passable. They update the flood-endangered areas and the status of streets and casualties (see Figure 12 (a): "Gefahr! Überflutung!", "Verletzte"). They request the context information, which is enriched by the technician who supports additional information from his location. Both teams update the information, which is synchronized between their clients and the central control room. The endangered areas and evacuation routes are coordinated by the command center and send to site-commanders and first responders. The fire-fighters begin to evacuate people to the south. The water level still raises and reaches over the bridge. The endangered area where one team of fire-fighters is located is flooded. Evacuation routes and endangered areas are dynamically updated by first responders and command center. Also the whole region is now suffering under a regional blackout since the central transformer station is hit. The response teams need to inform the civilians about evacuation routes.

After the situation stabilizes, the power grid site-commander and the technician coordinate how to ensure that most of the region can be power supplied. Therefore, the technician updates the status of transformer stations and informs the site-commander, which station can be switched on. Further, the repair and response teams need to organize, which streets and transformer stations have to be repaired. Updating the on-site information to the command center and site-commanders enables their situational awareness and allows directing forces where they are needed. They can consider local incidents in their decision making, which would be ineffective without our system.

7 Qualitative Evaluation

Field studies for crisis management systems are hard to conduct. Realistic crisis data is often not available or classified. Therefore, the project partners decided to simulate realistic scenarios, which are then analyzed by target users with our system. We conducted a qualitative study based on expert feedback rounds and interviews. The first evaluation round was a main objective in the project and was conducted within ten staff months. The process and results are discussed in the following.

7.1 Process

Evaluation Teams: We formed four teams within the nationwide interdisciplinary project (blinded for review): 1) *Data team* consists of two members of the Federal Office of Civil Protection and Disaster Assistance of Germany (henceforth, BBK), as well as four representatives of power suppliers, and further two simulation experts for smart grid technology and social media. They designed the scenarios mentioned in the section above and provided the data. 2) The visual *design team* (represented by the authors) consists of eight visual analytics experts who designed the system based on the scenarios and data. 3) An *interview team* of two persons with backgrounds in visual design conducted the qualitative interviews with domain experts. 4) An external *experts team* consisting of two members of the BBK with experience in crisis response and ten power grid operators of different regional power grids in Germany. We selected these different but related domains since they can be considered as the future target users. The crisis managers are part of the command center level, whereas the operators can be considered as site-commanders that are focusing on one particular region and domain affected by the crisis.

Interviews: The *interview team* was involved in the creation of the scenarios but not in the design process and is therefore considered independent of the design decisions. However, this team was trained by the *design team* on the system components and supported with documentations. Further, they prepared questionnaires for the qualitative interviews. The *interview team* visited seven control rooms of regional power grid suppliers in Germany and also interviewed experienced crisis managers of the BBK. This *expert team* did not know the system and scenarios. The interviews were conducted in concrete steps: First, the *interview team* presented the single components of the system. After the experts were familiar with the system the interview team presented one scenario as a use case. The scenario was stopped at critical events and the experts were asked to analyze the crisis and to draw decisions with the help of the system. Then interviews according to the questionnaires were conducted. The *interview team* analyzed the results and summarized the findings, which were reported back to the *design team*, who carefully analyzed the findings and improved the system for a second iteration, which is still future work.

7.2 Results & Discussion

We found two major facts in the interviews with the target users:

1) Domain experts reported that such systems are needed for crisis response: The experts reported that there is an urge for information of the crisis site, because in most cases the command centers are blind and wait for phone calls to update their status: “The social media analysis is great. In most cases, first responders are too busy to update the crisis center. Direct access to Twitter messages that are linked to the crisis would give us a clearer and faster overview of the crisis.”, “We want a direct visual communication with first responders”. We found that they are not only interested in the information of first responders about the crisis but also how the affected civilians are describing their status: “Some people may just feel to be forgotten by the government and we could respond with sending in some teams to show our presence”. Further, we found that first responders do have an urge for the context and development of the crisis, because they do not know what may hit them within the next minutes. They highlighted that in-situ and social media analysis can improve to narrow down root causes on-site and also to effectively steer first responders:“The control center does not even know, which transformer station can be reached by repair teams. We would need information about the status of streets and areas around stations. If we could use these tools today, we could directly send the teams where they really could make a difference”. The *interview team* found that the tools were efficiently understood by the expert team, however, they highlighted that a target user of such systems would require a significant amount of training. They conclude that the concepts of our system are sound, however, will require further investigations to integrate this in future crisis response centers.

2) Crisis-Managers and Site-Commanders disagreed on the level of detail: All experts agreed that social media and linked communication with first responders is important and that our system could be used in crisis scenarios. We found that power grid operators and crisis managers disagreed on the level of details of such visualizations. The crisis managers wanted to perceive the crisis and simulate different alternatives in abstract manner. Thus, they were satisfied with our components. However, the power grid operators requested more domain details and domain standards in the overviews. They reported that the whole system is interesting, however, the visualizations do not meet the requirements of power grid monitoring. We conclude that visualizations have to adapt to the level of command structure and may be adapted to special domains.

7.3 Discussion, Limitations & Future Work

We found that the target users were convinced of the system and its applicability. However, we see that the system does not fulfill all requirements to be an operational system for power grid control. In these domains, research has evolved over decades to develop customized solutions for this particular infrastructure. Interestingly, the operators that were involved in the flood disaster that hit Germany in 2013 said that they were almost blind after the water destroyed the first transformer stations. Therefore, they highlighted an urge for on-site information and social media analysis. In the future scenario of interconnected infrastructures such as smart grid technology, we see a higher complexity as in todays power grids. Command centers must overlook and perceive the full context of a crisis. Therefore, abstract visualizations are needed, which was approved by our crisis managers. We argue that our system exemplifies a means for future central crisis managements to integrate different critical infrastructures, social media and in-situ analysis. It will however be interesting to discover the correct level of detail to satisfy each role in the command structure. For this, we plan to conduct direct user assessments to improve our components to the needs of particular site-commanders.

Our scenarios are limited to the available resources within the project and thus, to the region that is employed with a smart grid. This might cause that our scenarios overlook large scale effects. Further, in our study the effects of time pressure was not considered, which may have significant influence on decision makers.

Another issue is security. The architecture might be vulnerable although we encrypt the communication between clients and server. Further, the issue of infeeding wrong information with, for example, a stolen device or misleading Twitter messages was raised in our interviews. Therefore, we see an urge to include security protocols into this architecture.

8 Conclusions

In this paper, we present a visual analytics system that combines multiple critical infrastructures, social media and in-situ analysis to support the different levels of command structure in crisis response. We present specialized equipment and visualizations for control rooms and mobile devices. Further, we discuss means for interactive simulation and visualization of the subsequent development of a crisis. This enables an interdisciplinary and distributed teams to understand and respond to crisis situations. Our system was applied in realistic scenarios and presented to real crisis managers, who conclude that there is an urge for such systems for crisis response.

REFERENCES

[1] N. V. Andrienko and G. L. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.

[2] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl. ScatterBlogs2: real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.

[3] B. Campbell and C. Weaver. Rimsim response hospital evacuation: Improving situation awareness and insight through serious games play and analysis. *IJISCRAM*, (3):1–15.

[4] J. Dietrich, C. Trahan, M. Howard, J. Fleming, R. Weaver, S. Tanaka, L. Yu, R. L. Jr., C.N, J. Dawson, Westerink, G. Wells, A. Lu, K. Vega, A. Kubach, K. Dresback, R. Kolar, C. Kaiser, and R. Twilley. Surface trajectories of oil transport along the northern coastline of the gulf of mexico. In *Continental Shelf Research*, 2012.

[5] H. Edelsbrunner and E. P. Mücke. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics*, 9:66–104, 1990.

[6] F. Fischer, J. Fuchs, P.-A. Vervier, F. Mansmann, and O. Thonnard. VisTracer: a visual analytics tool to investigate routing anomalies in traceroutes. pages 80–87. ACM, 2012.

[7] J. Guest, T. Eaglin, K. Subramanian, and W. Ribarsky. Visual analysis of situationally aware building evacuations, 2013.

[8] S. Kim, Y. Jang, A. Mellema, D. Ebert, and T. Collins. Visual analytics on mobile devices for emergency response. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 35–42, Oct 2007.

[9] J. Kohlhammer, T. May, and M. Hoffmann. Visual analytics for the strategic decision making process. In *GeoSpatial Visual Analytics*, pages 299–310. Springer, 2009.

[10] J. Liu, Y. Xiao, S. Li, W. Liang, and C. L. Chen. Cyber security and privacy issues in smart grids. *IEEE Communications Surveys & Tutorials*, 14(4):981–997, 2012.

[11] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. pages 181–190. IEEE, 2011.

[12] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, M. Wade, and D. S. Ebert. Understanding syndromic hotspots-a visual analytics approach. pages 35–42. IEEE, 2008.

[13] A. B. Marques, G. N. Taranto, and D. M. Falco. A knowledge-based system for supervision and control of regional voltage profile and security. *IEEE Transactions on Power Systems*, 20(1):400–407, 2005.

[14] S. Mittelstädt, D. Spretke, D. Sacha, D. A. Keim, B. Heyder, and J. Kopp. Visual analytics for critical infrastructures. pages 1–8. VDE, 2013.

[15] T. J. Overbye and J. D. Weber. New methods for the visualization of electric power system information. pages 131–16c. IEEE, 2000.

[16] V. Pascucci, D. E. Laney, R. Frank, G. Scorzelli, L. Linsen, B. Hamann, and F. Gygi. Real-time monitoring of large scientific simulations. In *Proceedings of the 18-th annual ACM Symposium on Applied Computing*, pages 194–198, Melbourne, Florida, March 2003.

[17] P. Proulx, S. Tandon, A. Bodnar, D. Schroh, R. Harper, and W. Wright. Avian flu case study with nSpace and GeoTime. pages 27–34. IEEE, 2006.

[18] J. J. Rooney and L. N. V. Heuvel. Root cause analysis for beginners. *Quality progress*, 37(7):45–56, 2004.

[19] E. Santos, J. Freire, C. Silva, A. Khan, J. Tierny, B. Grimm, L. Lins, V. Pascucci, S. A. Klasky", R. D. Barreto, and N. Podhorszki. Enabling advanced visualization tools in a simulation monitoring system. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 358–365. IEEE, December 2009.

[20] D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. pages 41–48. IEEE, 2012.

[21] J. J. Thomas and K. A. Cook. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[22] B. M. Tomaszewski, A. C. Robinson, C. Weaver, M. Stryker, and A. M. MacEachren. Geovisual analytics and crisis management. pages 173–179. Delft, the Netherlands, 2007.

[23] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991.

[24] L. Wang, J. Giesen, K. T. McDonnell, P. Zolliker, and K. Mueller. Color design for illustrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1739–1754, 2008.

ClusteRim: Maintain Context-Awareness via Aggregated Off-Screen Visualization

Submission #286

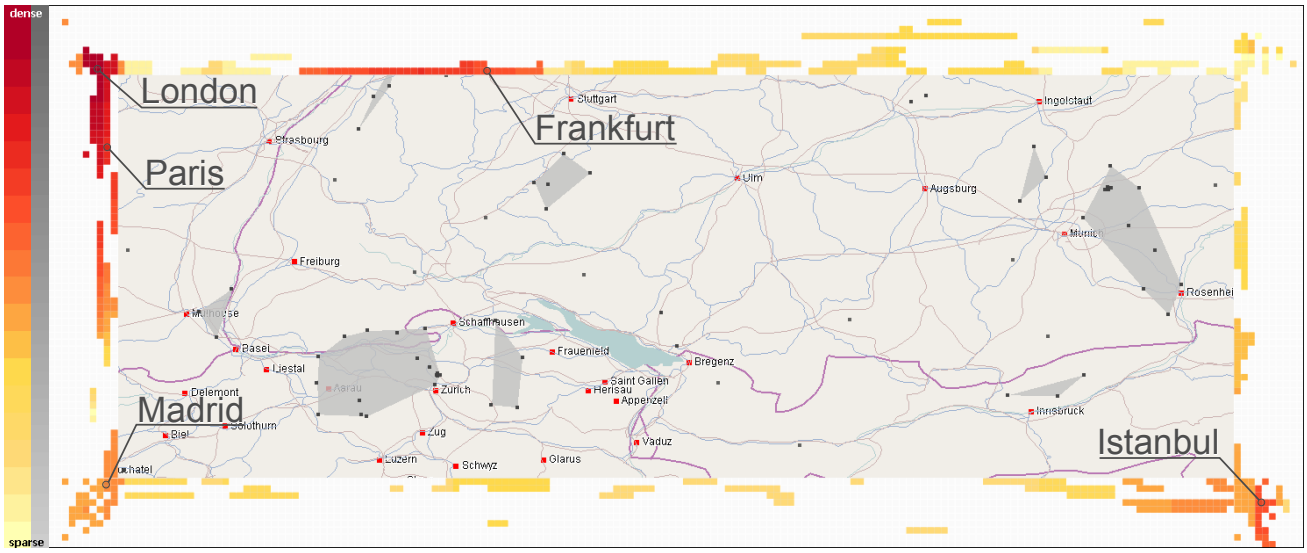


Fig. 1. ClusteRim applied to the European aviation data set. Our technique easily reveals areas of interest outside the viewport through the visualization of off-screen flight clusters. The clusters that can be quickly identified by color and size are Frankfurt, London, Paris, Madrid, and Instanbul.

Abstract— When exploring large data sets, we often lose the contextual overview. Existing approaches supporting navigation, such as overview-plus-detail and focus-plus-context techniques, offer expedient solutions but do not scale to large data sets. While overview-plus-detail techniques suffer from overplotting and sacrifice overview for augmented inset graphics, focus-plus-context techniques result in partial information loss due to distortion. To overcome named issues we present ClusteRim, a novel focus-plus-context technique that extends state-of-the-art off-screen visualization methods. While zooming areas of interest, ClusteRim preserves the contextual off-screen information as color grids within a designated space around the focal area. It does not only support navigation within large data spaces but provides context-preserving views on spatial data. This paper describes the methods to generate the ClusteRim visualization: hierarchical aggregation and mapping of off-screen objects to surrogate visual representations on color grids. In two application examples we demonstrate the usefulness of our technique in exploring large spatial data sets. A user study proves the effectiveness in context tasks compared with a state-of-the-art off-screen visualization technique.

Index Terms—Off-screen Visualization, Focus+Context, Zooming and Navigation, Zoomable User Interfaces, Contextual Views.

1 Introduction

When visualizing vast amount of data, we need to compromise between overview and detail. As the size of data sets continuously increases, it becomes even more difficult to achieve both at the same time. One widely used approach is to filter data based on user-defined conditions [1]. Eliminating non-relevant information and highlighting query-based findings may result in an inevitable loss of overview and a cluttered representation. Sampling [33], on the other hand, refers to the process of selecting a data subset. We encounter the same problems as for filtering: eliminating information might result in a loss of context. Further, sampling reduces clutter but does not solve this problem.

Existing approaches, such as focus-plus-context and overview-plus-detail techniques try to preserve overview while giving the user the possibility to inspect the data in detail. These techniques do not scale to large data sets: Overview-plus-detail techniques suffer from overplotting and make use of inset graphics which may result into a loss of overview. In contrast, focus-plus-context techniques lose information due to distortion and also do not solve clutter issues.

An example that illustrates the problems is a scatterplot visualizing several thousands points, which most probably results in a cluttered display. This means, some areas of the scatterplot are so dense that it is impossible to perceive details of the dataset. Common solutions which enable the exploration of the dataset, but provide context at the same time, are based on focus-plus-context and overview-plus-detail techniques. For instance, focus-plus-context techniques maintain a cluttered context view while providing details of the scatterplot. Also, the data between focus and overview is distorted to a certain extend, causing perceptual problems that ultimately lead to a loss of context. Overview-plus-detail approaches mainly rely on a zoomed representation of the data. This allows the inspection of data points but also does not resolve the perceptual problems coming along with the cluttered scatterplot. Due to clutter, both approaches fail to maintain context since they do not address problems caused by the visual representation

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

of the data.

To overcome these issues, we propose a novel, generalizable technique, called *ClusteRim*, to enhance context-aware navigation in large data spaces. The viewport is surrounded by a grid with colored cells, which represent aggregated off-screen objects. The viewport shows details utilizing a zoomable visual representation. This technique can be applied for navigation of data spaces in any visual representations (e.g., navigating along a route while maintaining traffic information in the surrounding area). This technique finds a proper balance between distortion-based techniques and common overview-plus-detail approaches.

Our contributions are: (1) A novel context-aware off-screen visualization technique; (2) A novel approach to combine aggregation and representation of off-screen objects; (3) A terminology bringing together data and pictorial space in the context of off-screen visualization. We also compare the performance of our technique with HaloDot, a state-of-the-art off-screen visualization technique, and prove its usefulness via application examples.

2 Related Work

So far, various techniques try to resolve the named problem of maintaining context-awareness while investigating a data set in detail [25]. Distortion oriented approaches and scalable user interfaces represent related work but are not the focus of this work. Following, we give a brief overview, before we point out related off-screen techniques.

Graphical and distortion-oriented approaches, formally known as *Focus-plus-Context*, manage to define a focal area without losing context. Some of the best known examples are the Bifocal Displays [2], the Perspective Wall [34], the Elastic Presentation Space [9], Melange [12], and Generalized Fish Eyes [15]. According to Leung and Apperley, the main feature of distortion-oriented approaches is to provide an overall context while allowing users to examine a local area in detail [32]. Despite recent improvements [8,38,39], some weaknesses are inevitable in large data spaces (e.g., the transition between focal and non-magnified areas). This area, which is known as drop-off area, is unmagnified and distorted to a certain extent, which results in partial information loss. Even with the use of different transition functions, the switchover between focal area and overview remains. Distortion-oriented techniques work well within a clutter-free environment, but fail to provide an overview in dense data. By design, distortion-oriented techniques interfere with any task that requires precise judgments about scale or distance [3].

Unlike distortion-based techniques, scalable user interfaces [11, 16] follow the Visual Information Seeking Mantra presented by Schneiderman [40], but aggregate visible information to avoid clutter. Zooming interfaces are able to reveal details, but fade out the overview. To preserve the overview, scalable interfaces make use of mini-maps, which are usually built by inserting inset graphics. However, inset graphics may overlay areas, preventing the perception of overlaid information. For the same reason, in-screen visualizations (visualization of non-clipped objects), which typically are created on top of the actual visualization, can decrease the amount of information gained through overview, in which such extra components need additional space on the viewport of the visualization. Burigat et al. name the drawback of limited space and further name the insufficient knowledge about its application to mobile devices [4].

2.1 Off-Screen Visualization

Many prior studies have developed and evaluated off-screen visualization techniques. Following, we give a brief overview of related off-screen techniques. We highlight areas for improvement in which we claim our contribution may fit as a new technique for context-aware off-screen visualization.

2.1.1 Cue-based Techniques

Cue-based off-screen techniques have mostly been used for navigation tasks and highlight points of interests (POIs). For instance, Aroundplot [28] shifts the perception of distant off-screen objects to a 3D environment through the use of context magnification. Clipped objects in 3D space are mapped to the border region of the display in 2D space. Hoffmann et al. [26] go one step further and propose visual cues on large displays. This technique addresses the problem that users sometimes fail to notice all changes occurring outside their visual span when using multiple displays. The authors therefore use visual cues to enable a fast window switching to targets outside their field of view.

A common technique to visualize POIs are arrows. Arrows are usually placed at the border region of the display and encode attributes like distance and direction by changing their size and orientation. Apart from arrows, Zellweger et al. [41] present a family of techniques called City Lights, that make use of the border region of the display. It consists of two techniques: City Light cues and Halos. City Light cues are thick lines on the border region and inform the user of clipped objects. A major drawback is the missing distance perception. Halo [3] visualizes off-screen objects on small displays using arcs, which intersect the border region of the display. The technique enables users to perceive the distance from POIs to the center of the current viewport. A comparative study by Burigat et al. [5] revealed that arrow-based visualizations outperform Halo in certain tasks; these are: pointing out a pair of POIs which are closest to each other and the ordering of POIs in increasing distance. This study was conducted with a data size of five and eight POIs. Due to the small size of the dataset, the effects of large datasets, for example perceptual problems because of cluttered visualizations, are not taken into account. Gustafson et al. [22] improved Halo and presented Wedge, which aims to overcome clutter and distance perception problems. Wedge uses acute isosceles triangles to represent off-screen objects. Distances are mapped to the legs pointing towards the target and clutter is avoided through repelling. Their evaluation shows that Wedge is more accurate than Halo but the cues perform similar considering distance mapping. Burigat et al. [6] demonstrate that Wedge performs worse in dynamic environments with high amounts of data than a simple overview and detail technique. EdgeRadar [23], a fisheye-based visualization, performs also better than Halo by projecting all POIs directly to the fixed border region of the display. We also see ClusteRim as improvement to City Lights and EdgeRadar; we enable context-awareness by addressing clutter issues and distance perception, among other named improvements.

These techniques are designed to work with small amounts of data. HaloDot [21] tries to solve the problem for large data sets. It aggregates POIs on a grid to the clipped area. The corresponding grid cell is then visualized as halo with color, transparency, or thickness indicating it's relevance. According to Gonçalves et al. [19], aggregation combined with relevance clues improves search for relevant POIs. Nevertheless, HaloDot does not consider different classes of POIs. If a noticeable amount of grid cells contain POIs, the visualization will produce considerable clutter. Gonçalves et al. [20] further conducted a comparative user study with a reduced amount of POIs that shows overview-plus-detail as well as adapted scaled arrows outperform HaloDot. Unfortunately, this study did not take large data sets into account.

HaloDot meets the requirements to visualize large data sets better than scaled arrows and overview-plus-detail techniques. HaloDot reduces clutter by taking the distance into account; the diameter of the arc is mapped to distance whereas arrows and overview-plus-detail remain in clutter. HaloDot does also not solve problems of clutter-free context-aware navigation all at once. It is obvious, to reduce clutter in the visualization of large datasets, the grid cell size can be increased. As a downside of this, the level of detail is reduced, which inevitably results in a loss of context.

Off-screen visualization has mainly been used as navigation or orientation guide for spatial data. The application of off-screen techniques to mathematical diagrams represents an interesting field that has not been visited yet. Games et al. [17] take first steps and apply cue-based visualization to barcharts and scatterplots with restricted view space. However, this technique does not address clutter and distance mapping.

2.1.2 Node-Link Diagram-based Techniques

In this section, we reference POIs as nodes that build up a node-link diagram. Node-link diagrams also use off-screen techniques to visu-

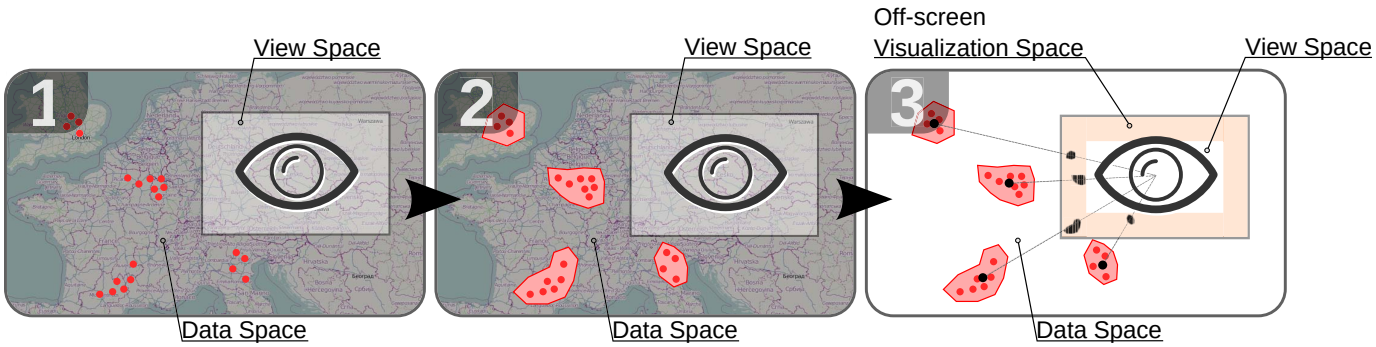


Fig. 2. ClusteRim pipeline. The map represents the whole data space whereas the opaque window inside with the eye at its center represents the viewport, visible to the user. From left to right: (1) First, the data outside the view space is determined by intersecting view space and data space. (2) Second, the clipped data are clustered using a density-based algorithm. (3) Third, the determined clusters are mapped to the border region of the viewport and visualized in the off-screen visualization space.

alize connected, clipped Nodes. Within a diagram, the connection of edges as well as nodes might suffer from clutter caused by overlap or intersections. Typically, off-screen visualization in node-link diagrams is used for navigation tasks and context-awareness, as for instance proposed by May et al. [35].

Frisch and Dachselt [14] propose an approach for node-link diagrams with application to UML class diagrams. They use the border region of the display to represent clipped UML elements through visual cues. Though, they do not take the distance to these clipped objects into account. In contrast, Myers and Duke [37] apply a lens-like technique to graph navigation. A lens, positioned and sized by the user, is set on top of the graph indicating regions of the network and providing navigational support. Moscovich et al. [36] do not use the border region of the display. Instead, they let the user decide through interactions to visualize and navigate to off-screen objects. They realized two techniques: Link Sliding and Bring & Go. Link Sliding animates the path to another node and preserves topology-awareness. Bring & Go zooms distant objects to the surrounding area of the focal point mapping direction and distance. An obvious disadvantage of this approach is the lack of context-awareness. The user loses the overview when clipped objects are pulled into the viewport. An interplay of interaction and context-awareness has been presented with Dynamic Insets [18]. Using a degree-of-interest function, the authors determine interesting off-screen objects and visualize them as insets at the border region of the display. Distance as well as direction is mapped to the insets to preserve context-awareness.

Off-screen techniques have already been widely implemented in node-linked environments. We therefore adapt ideas from both, cue-based and node-link diagram-based techniques, and apply them to a new technique to visualize vast amounts of off-screen POIs.

3 Terminology

In this chapter, we define a terminology that describes different spaces used throughout this paper. Figure 2 outlines four different spaces: the data space, the view space, the off-screen space, and the off-screen visualization space.

Data space. We define the data space D as follows:

$$D := \{(x,y)|(x,y) \in \mathbb{R}^2\} \tag{1}$$

The data space represents all data points contained in the data set to be visualized.

View space. The view space V is defined as set of points, where x and y are restricted to the area enclosed by a rectangular frame with the coordinates $x = v_x, y = v_y$ and $width = \hat{v}_x - v_x$, $height = \hat{v}_y - v_y$. That is, V contains all visible data points from the data space.

$$V := \{(x,y)|(v_x,v_y) \leq (x,y) \leq (\hat{v}_x,\hat{v}_y) \wedge v,\hat{v} \in \mathbb{R}^2\} \tag{2}$$

In large data spaces, the view space usually is a subset of the data space. The view space is further restricted in size by the screen resolution. For example, if the whole data space is presented on screen, it is smaller than or equal to the view space. The view space can be manipulated by the user by translating the x and y coordinates or changing the scale. In this paper the visual representation of the view space is also denoted as **viewport**.

Off-screen space. The section of the data space, which is not part of the view space is formally referred to as off-screen space.

$$O := D \backslash V \tag{3}$$

Objects contained in this space are called *clipped* or *off-screen objects*.

Off-screen Visualization Space. The off-screen visualization space, often denoted as *rim*, refers to the pictorial space that is the area of the viewport, which is used to display the off-screen objects. Actual size and appearance depend on the visualization technique and the users.

4 ClusteRim – Off-Screen Visualization

In general, off-screen visualizations tackle the problem of displaying two different data spaces in one visualization. Such techniques fall into the category of Focus+Context techniques, as described by Card et al. [7].

Focus+Context is very beneficial to users for geographic visualizations. Figure 1 displays details of an aviation data set in a very small section of Central Europe using our visualization technique. At the same time, the context of Europe – the whole aviation data over the entire Europe ranging from England to Turkey – is aggregated and presented through off-screen visualization. Besides geovisualizations, off-screen visualizations can support a wide range of applications when the whole data space can not be displayed.

Our approach is based on the idea to present both, off-screen and view space within one visualization at the same time. In the rim, the border region of the viewport, the off-screen visualization space is created containing a visualization of the off-screen space.

ClusteRim is geared to the well-known InfoVis pipeline proposed by Chi and Riedl [27], that consists of four consecutive steps:

$$\textit{Source Data} \rightarrow \textit{Data Tables} \rightarrow \textit{Visual Abstraction} \rightarrow \textit{Views}$$

Our system workflow adequately fits this pipeline and is therefore applicable to any visualization of data. The first step, *Source Data* to *Data Tables* is executed when a data set is loaded. Aggregation, mapping, and projection are the three components of the *Visual Abstraction*. Afterwards, the *Views* are created based on the visual abstraction.

ClusteRim is steerable. At each point of the algorithm, users can manipulate parameters (e.g., *minpoints* or ε of the clustering algorithm or the size of the off-screen grid cells). This makes our approach also suitable for Visual Analytics applications. Hereinafter, we describe all steps that are illustrated and contained in the pipeline shown in Figure 2.

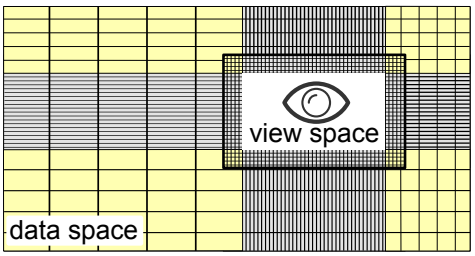


Fig. 3. Schematic view of the different areas which need to be mapped from the off-screen space to the off-screen visualization space. Each of these areas is mapped with a different projection to the visualization depending on the position of the view space in relation to the off-screen space.

4.1 Data Projection and Mapping

The combination of off-screen visualization (context) and detailed area (focus) opens a huge design space, ranging from data processing to details of the visual mapping. Our main design decisions for ClusteRim are the following:

In general, items generated by any aggregation technique (e.g., clusters) must be represented smaller than the viewport dimensions. Otherwise, the utility of the off-screen visualization is limited, because the item will be visible on the off-screen visualization space as well as in the viewport.

Partitioning the Off-Screen Space. The off-screen space is partitioned on two levels. At first, it is divided into eight different areas, as they are presented in Figure 3. In each of these areas, a regular grid with a fixed cell size is used for visual aggregation. This preserves spatial relations and distances of aggregated items in the off-screen data space with grid precision. The number of rows and columns as well as the size of the grid cells are adjustable to users' needs and application requirements. For instance, these parameters can be adjusted depending on the zoom level of the viewport, which effectively introduces semantic zooming capabilities.

From Off-Screen Space to Off-Screen Visualization Space. The grid of each of the partitioned areas in north, east, south and west of the off-screen data space is directly mapped to the adjacent area in the off-screen visualization space; Figure 3 shows these different areas. The partitioning is performed separately for each of the areas, causing the resulting distortion to depend on the position of the view space in relation to the off-screen space. The corners, north west, north east, south east, south west, are mapped to the corresponding corners of the off-screen visualization area. Eight different distortions are used depending on the off-screen data space. This is a strength of our mapping technique, because it introduces context sensitivity to each of the mapping areas. In most common distortion oriented techniques, all sides of the data space are treated equally in terms of visualizing them. We instead use the distance from viewport to the data space borders in order to adapt the mapping towards the off-screen visualization space. For example, is the viewport placed more to the western in data space, the western part of the off-screen visualization maps less data space than the eastern part. As a result, only the covered data space is mapped to the off-screen visualization space providing distance perception within the data space. Regardless of the size of the data space, both parts of the off-screen visualization space are equal in size.

In contrast, mapping the corners correctly is challenging, because in this area two parts of the off-screen visualization space with different mappings are converging. We are aware of the limitations that come along with the mapping of off-screen space corners to the off-screen visualization space, as these corners represent the largest share in terms of space but get less space in the visualization. We will pick up this issue in the discussion.

Off-Screen Space Visualization. The representation in the off-screen visualization space is, such as the data space mapping, also grid-based. Each cell is colored according to the number of points contained by the intersecting cluster.

Visually, a regular grid is drawn, while the contents *of each cell* are determined by projecting the section from the off-screen space to the corresponding grid cell. The coloring can be changed to display other data values, which should be application specific. The coloring of each cell is computed via a special mapping algorithm described in the following section.

4.2 Algorithms

We briefly explain the main parts of the algorithm used to generate the off-screen visualization (see Algorithm 1).

Algorithm 1 ClusteRim Algorithm

Require: $D \neq \emptyset$, $(D - V) \neq \emptyset$
$A \leftarrow$ aggregate(D)
$M \leftarrow$ partitionOffscreenDataSpace(A, $D \backslash V$)
$M' \leftarrow$ projectGridToOffscreenVisSpace(A, $D \backslash V$, M)
$C \leftarrow$ mapColors(A, M')
drawGrid(M', C)

Aggregation. A density-based clustering algorithm is used to aggregate the data. Algorithms of this family are capable of creating clusters of arbitrary shapes, which is desired when representing a data set without any prior knowledge of the data distribution. We therefore do not need any additional preprocessing or examination of the data set. The parameters $minpoints$ and ε are chosen in a way that the number of large and small clusters are roughly equal, leading to an off-screen area with a balanced number of large and small colored areas. We propose to use the DBScan algorithm [13], because its concept of density connected sets fits the application of off-screen visualization very well. To speed up the ε-neighborhood queries of the clustering algorithm, we create a Quadtree index structure. This ensures that the runtime of the data aggregation is in $\mathcal{O}(n \cdot log(n))$. This Quadtree is re-used in other parts of ClusteRim.

Partitioning and Projection. Before any partitioning or projection is created, we set a parameter called $maxSegments$. On horizontal areas (north, south), it defines the number of rows; on vertical areas (east, west), it defines the number of columns to use. Our partitioning and projections are thereby based on a grid, which is distorted to fit the available data space best. For each mapping area shown in Figure 3 the distortion is different (e.g., the distortion of the western part is bigger than the distortion of the eastern part). But in the off-screen visualization space both distortions are assigned equal space, which results in a homogeneous appearance in those areas. The grid creation and the actual partitioning or projection runs in $\mathcal{O}(n \cdot log(n))$.

Color Mapping. For the color mapping, we designed an algorithm, which colors every grid cell intersecting a cluster (see Algorithm 2).

Algorithm 2 Color Mapping Algorithm.

$colors \leftarrow$ getColorMap(Yellow, Red)
$c \leftarrow$ getClusterCentroid()
$n \leftarrow$ getClusterWidth()
$m \leftarrow$ getClusterHeight()
for $i = c(x) - \frac{n}{2}$ **to** $c(x) + \frac{n}{2}$ **do**
  for $j = c(y) - \frac{m}{2}$ **to** $c(y) + \frac{m}{2}$ **do**
    if hasBeenProcessed($cell[c(x) \pm x][c(y) \pm y]$) **then**
      $cell[c(x) \pm x][c(y) \pm y] + +$
    else
      $cell[c(x) \pm x][c(y) \pm y] = 1$
    end if
  end for
end for
assignNormalizedColors($cell$)

For each cluster starting at the cell containing the centroid, the algorithm iterates the cells intersected by the cluster in a circular manner, until no more cells are covered by the starting cluster. In each iteration

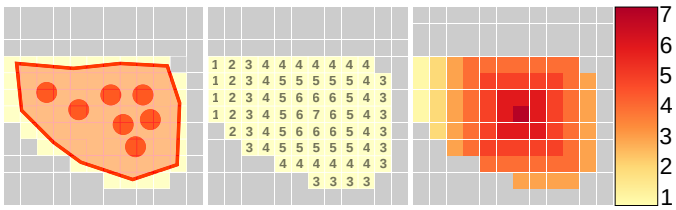


Fig. 4. The mapping of a cluster (left) to numbers (middle) to colors (right) determining the off-screen grid cell color.

of the algorithm, as the radius around the cluster centroid increases by one, the number associated with each cell intersected or filled by the cluster also increases by one. If a cell is visited for the first time, it is assigned one. In the end, this number determines the color chosen from a color map ranging from yellow to red. This algorithm runs in $\mathcal{O}(n \cdot m)$ where n and m denote the dimensions of the grid to color.

Figure 4 shows the coloring process using an example of one cluster (left). The centroid is colored as per cluster size. According Harrower and Brewer [24], we use a sequential color scheme to represent data that range from low to high values on a numerical scale. We further follow the convention "Dark equals more." This means, a cluster that contains lots of points is colored darker compared to a cluster that contains less points. The outcome of this algorithm (right) resembles a classical rasterization; the outcome is later directly mapped to the grid structure projected to the off-screen visualization. The color map we used meets our requirements to demonstrate the effectiveness of ClusteRim; the color map is application dependent and should therefore be adapted to the desired outcome.

Normalization of color depends on the data distribution and needs to be adapted based on the actual data set. For the images in this publication, we employed a min-max normalization based on the square root of the values, because the distribution of values to color were homogeneous. In such cases, square root normalization gives more details in the middle areas of the distribution.

There is also the possibility to compute a bi-cubic interpolation [30] in-between the resulting cell colors (see Figure 12 as an example). This generates a visually more pleasing visualization, but is very time consuming especially for large data sets or large cell sizes.

5 Evaluation

We decided to conduct a user study comparing our technique with HaloDot, which is a state-of-art off-screen visualization technique. Though a prior comparative study [20] found that HaloDot performs worse than Scaled Arrows and an mini-map-based, overview-plus-detail technique on navigation tasks with 40 data points, we believe that Scaled Arrows and mini-map techniques are not scalable for more than 1000 data points due to significant clutter. Thus, we excluded these techniques from our evaluation.

We conducted a two-stage user study to evaluate the effectiveness of ClusteRim compared with a state-of-the-art technique, HaloDot. HaloDot is based on the idea to apply a grid to the off-screen space and to represent each grid cell that contains POIs as a halo. Using vast amounts of data lead to enormous clutter, even for HaloDot, so it could be an unfair comparison with ClusteRim. Therefore, we made HaloDot competitive and adapted it by applying halos to the calculated clusters instead of grid cells. Furthermore, we determined the relevance of clusters by its amount of contained points and mapped it to the thickness of the arcs.

We compared ClusteRim with HaloDot on three different tasks: 1) context-interpretation comparison task, 2) context-interpretation overview task, and 3) target selection task. The detailed description on the three tasks will be explained in the method section. In addition, as prior studies (like for instance in [19]) indicate, we also tested whether ClusteRim can be influenced by the size of data. The following are our hypotheses for this experiment:

H1 Context-Interpretation Task - Comparison:

- **H1.1:** The task performance time will be less with ClusteRim than with HaloDot.
- **H1.2:** The task performance accuracy will be higher with ClusteRim than with HaloDot.
- **H1.3:** The confidence level will be higher with ClusteRim than with HaloDot.

H2 Context-Interpretation Task - Overview:

- **H2.1:** The task performance time will be less with ClusteRim than with HaloDot.
- **H2.2:** The task performance accuracy will be higher with ClusteRim than with HaloDot.
- **H2.3:** The confidence level will be higher with ClusteRim than with HaloDot.

H3 Target Selection Task:

- **H3.1:** The task performance time of ClusteRim will be comparable to HaloDot.
- **H3.2:** The task performance accuracy of ClusteRim will be comparable to HaloDot.

5.1 Method

To test the hypotheses, we designed a within-subject study. All tasks were performed for both techniques, ClusteRim and HaloDot and applied to three different data sizes: one thousand, three thousand, and five thousand data points of interest. We further conducted a two stage user study: First, to evaluate the effectiveness of ClusteRim regarding context-awareness (**H1** and **H2**) and second, to evaluate the target selection (**H3**). One of the main advantages of ClusteRim is to preserve overview while analyzing parts in detail; but we do not want to sacrifice performance in target selection which is not considered as strength of ClusteRim.

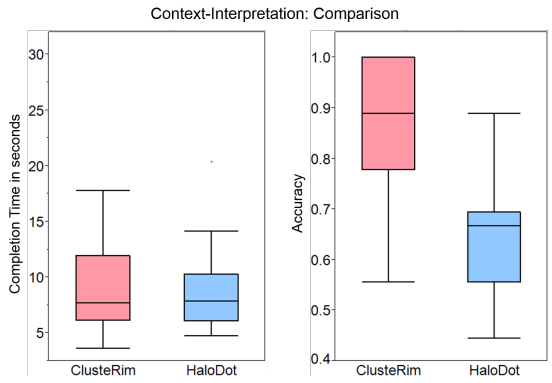
Since we conducted a two-stage user study, we also had two different user groups. The first user study addressed context-awareness. To test **H1** and **H2**, each participant accomplished 36 trials, which means 18 trials per task: context-interpretation comparison and context-interpretation overview. For the experiment, the tasks followed a random order. The second user study addressed target selection. To test **H3**, each participant accomplished 18 trials.

For each task the 18 trials are characterized as follows: 9 trials per technique and within these, 3 trials per data sizes (i.e., 1000, 3000, and 5000 data points), respectively. For the experiment, the trials followed a random order. This means, within a task ,we randomized the order of the used technique; and within a technique we randomized the data size. Before each task, the participant was presented a detailed description exemplifying the used technique within the task. Further, for each technique within a task, participants performed a training phase. A training phase consisted of three examples for respective tasks. After each example, the correct solution was presented.

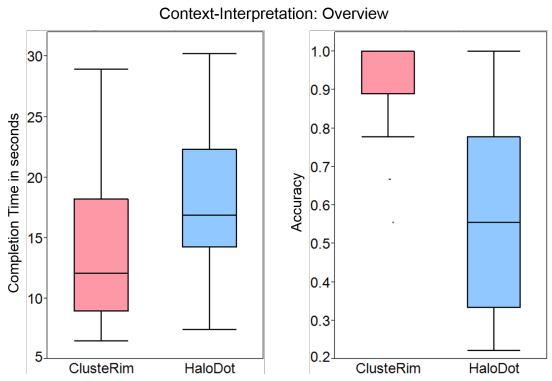
In total, 23 participants were recruited (mean age = 27.9, min age = 20, max age = 51) for the first user study. Among them, 15 participants majored in computer science, 6 in engineering, and the rest of 1 in economics and civil services, respectively. For the second user study, 22 participants were recruited (mean age = 32.7, min age = 20, max age = 71). Among them, 14 participants majored in computer science, 3 in engineering, 2 in administrative studies, and 1 in each, economics, social science, and teaching. All participants had normal of corrected-to-normal vision(20/20). All participants used computers on a regular basis (30 minutes to 8 hours a day), and had basic understanding of visualizations. The experiment was conducted on different computers that at least supported the fixed window height of the prototype of 1390 pixels, and fixed width of 710 pixels, providing each participant the same amount of screen space. A keyboard and a mouse were provided for user input and interaction.

5.1.1 Context-Interpretation Comparison Task

The context-interpretation comparison task was designed to test whether the asked POI is located in-screen or off-screen (see Figure 6). In order to make the results comparable, we used the nine different datasets in



(a) Task completion time and accuracy of ClusteRim and HaloDot for the context-interpretation comparison task.



(b) Task completion time and accuracy of ClusteRim and HaloDot for the context-interpretation overview task.

Fig. 5. Context-interpretation task experiment Results. This task consists of two subtasks: the (a) comparison task and the (b) overview task. We could prove, that in both subtasks, ClusteRim performs significantly better than HaloDot, regarding accuracy with tendency to a faster completion time.

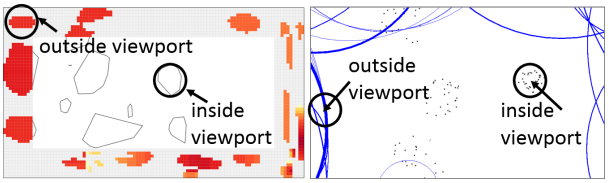


Fig. 6. Schematic representation of the context-interpretation comparison task. Participants were asked to decide whether the biggest lies inside or outside the viewport.

a randomly chosen order for ClusteRim as well as for HaloDot. In this task, participants were asked to determine whether the biggest cluster, by means of contained points, is visualized in-screen or via the corresponding off-screen technique. Therefore, a static image was presented to participants using ClusteRim or HaloDot. By investigating in-screen and off-screen POIs (visualized with ClusteRim or HaloDot), participants had to determine whether the biggest cluster lies within or outside the viewport. After each trial, participants responded to a survey question on how confident they were with their decision via a 4-level scale: *1 = unsure, 2 = somewhat unsure, 3 = somewhat confident* to *4 = totally confident*.

5.1.2 Context-Interpretation Overview Task

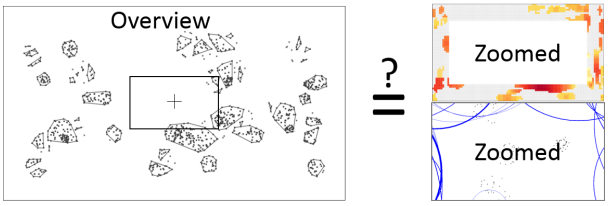


Fig. 7. Schematic representation of the context-interpretation overview task. Participants were asked to decide whether one zoomed image - out of a selection - corresponds the overview scatterplot image.

The context-interpretation overview task aimed to perceptual performances when POIs outside the viewport are visualized either by halos or by ClusteRim (see Figure 7). The goal of ClusteRim is to provide context-awareness even in large-scale data sets. For this task, we presented a static overview image of a scatterplot. Within this image we marked the area which will be zoomed. Below the scatterplot we offer a selection of three images showing the zoomed area using a given technique: HaloDot or ClusteRim. One out of the three options corresponds to the scatterplot. After each trial, the participants specified how confident they were about their decision via the 4-level scale.

5.1.3 Target Selection Task

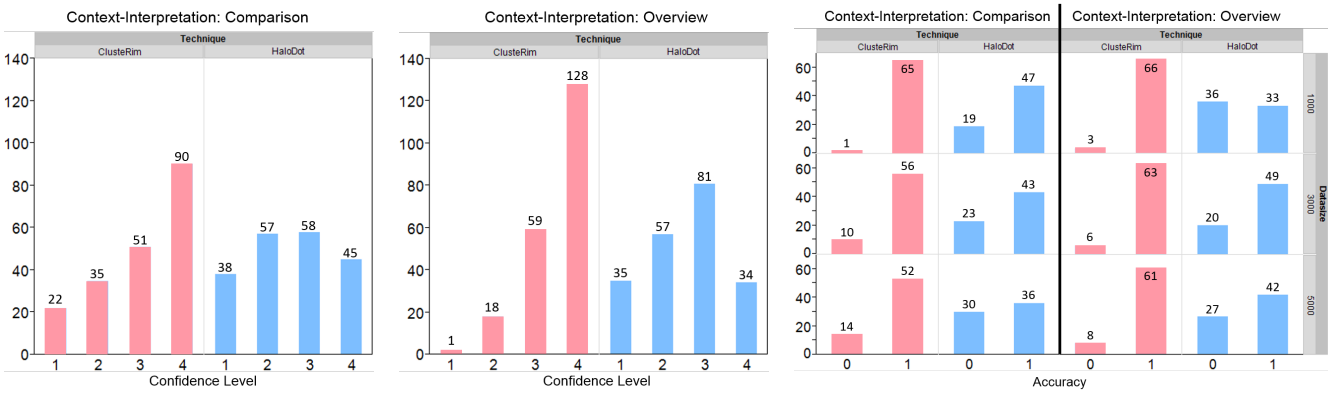
Even though ClusteRim provides additional values by providing context awareness, we also do not want to sacrifice the performance of navigational tasks for real-world applications. To show that ClusteRim is comparable to HaloDot in target selection, we designed this task. Participants were asked to interactively select a specific point of interest, a cluster with the largest number of points, among given data points randomly spread in a map space outside the focal viewport. This task intends to test the effectiveness of given off-screen visualization techniques for navigation. We collected execution time (i.e. the time elapsed until selection) and accuracy, which is defined as the correctness of selection.

After all 54 trials, participants were asked to respond to a demographic survey. We used ANOVA for the task completion time and a non-parametric Friedman's test for accuracy as well as for the confidence score.

5.2 Results

Our test results failed to confirm **H1.1**, but successfully confirm **H1.2** and **H1.3**: *in context-interpretation comparison task, ClusteRim showed higher accuracy than HaloDot with equal task performance time.* There was no significant difference between techniques ($p = 0.809$, $F(1,388) = 0.06$) and between datasets ($p = 0.318$, $F(2,388) = 1.15$) on task completion time. On the other hand, there was significant difference between techniques ($p < 0.001$, $\chi^2(1, N = 396) = 31.179$) and dataset sizes ($p = 0.002$, $\chi^2(2, N = 396) = 12.138$) on accuracy. In addition, there was significant difference between techniques ($p < 0.001$, $\chi^2(1, N = 396) = 25.372$) and between dataset sizes ($p < 0.001$, $\chi^2(2, N = 396) = 26.452$) on confidence. Figure 5(a) shows that participants performed more accurately with ClusteRim. In addition, Figure 8(a) shows that participants felt more confident as post hoc analysis showed ($1000 > 3000 > 5000$); Figure 8(c) achieved a higher accuracy even across different data sizes as post hoc analysis showed ($1000 > 5000$). It also shows tendency that the error rate increases as the dataset size increases.

Our test results successfully confirmed **H2.1**, **H2.2**, and **H2.3**: *in context-interpretation overview task, ClusteRim showed higher accuracy, less task completion time, and higher confidence than HaloDot.* There was significant difference between techniques ($p < 0.001$, $F(1,388) = 26.48$) but no difference between datasets ($p = 0.298$, $F(2,388) = 1.15$) on task completion time. In addition, there was significant difference between techniques ($p < 0.001$, $\chi^2(1, N = 396) =$



(a) (b) (c)

Fig. 8. Histograms showing the distribution for ClusteRim and HaloDot depending on confidence level and accuracy with respect to data size. (a) and (b) show for both, context-interpretation comparison task and context-interpretation overview task, the distribution of the confidence level (1- unsure, 2 - somewhat unsure, 3 - somewhat confident, 4 - totally confident). (c) shows the distribution of accuracy (0 - wrong, 1- correct) according to technique and data size.

61.429) but no difference between dataset sizes ($p = 0.002$, $\chi^2(2, N = 396) = 3.583$) on accuracy. In addition, there was significant difference between techniques ($p < 0.001$, $\chi^2(1, N = 396) = 118.624$) and between dataset sizes ($p = 0.012$, $\chi^2(2, N = 396) = 16.382$) on confidence. Figure 5(b) shows that participants performed more accurately with ClusteRim. Furthermore, Figure 8(b) shows that participants felt more confident as post hoc analysis showed ($1000 > 5000$); Figure 8(c) achieved a higher accuracy even across different data sizes as post hoc analysis showed ($1000 > 5000$). It also shows tendency that the error rate increases as the dataset size increases.

The test results confirm our hypothesis **H3**: *in target selection task, the two techniques showed comparable performances in terms of task performance time and accuracy*. Our test results did not show significant difference between techniques ($p = 0.264$, $F(1, 257) = 1.25$) and between datasets ($p = 0.344$, $F(2, 257) = 1.07$) on time. In addition, we did not find significant difference on accuracy between techniques ($p = 0.140$, $\chi^2(1, N = 264) = 2.178$) as well as data sizes ($p = 0.211$, $\chi^2(2, N = 264) = 3.109$).

In this experiment, we showed that the benefit of ClusteRim on context-interpretation tasks compared with HaloDot. The following chapter further describes useful applications of this technique in various kinds of context.

6 Application Examples

In this section, we apply our technique to a geo-spatial as well as a spatial data set and highlight the context-preserving capabilities of ClusteRim.

6.1 Exploring Large Geo-spatial Datasets

For the geo-spatial representation of datasets using ClusteRim, we consecutively present three different application examples.

6.1.1 Flight Radar Data

First, we apply ClusteRim to a snapshot of the European aviation dataset retrieved from `flightradar24.com`. Figure 9 shows the aviation data for day and night from the southern part of Germany point of view. We found out that Amsterdam has no ban on night flights whereas Munich and London follow a noise quota and Frankfurt and Zurich do not allow night flights at all. The viewport shows plane clusters located in the off-screen space. Our technique reveals a huge difference regarding air traffic around airports between day- and nighttime. Daytime is regarded from 05:00 to 23:00 and consists of circa eight thousand flights at a time; nighttime is regarded from 23:00 to 05:00. The color coding from yellow (clusters that contain a small amount of planes) to red (clusters that contain a large amount of planes) makes it easy to spot the most interesting areas within Europe. The distance to these spots is nearly homogeneous, as the visualization depends on the distance mapping from the off-screen space to the off-screen visualization space.

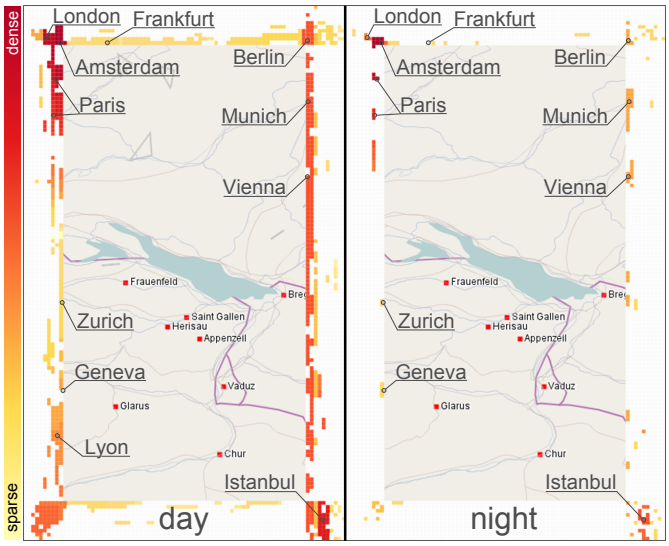


Fig. 9. ClusteRim applied to the European aviation data set. This figure reveals airports with and without a ban on night flights by a day and night comparison.

6.1.2 VAST Challenge 2011

In this example, we showcase the geo-spatial visualization of off-screen data over time using the dataset of the VAST Challenge 2011. The dataset contains microblog entries of 21 days, sent by users located in the fictive city Vastopolis. It is known that an epidemic happend during this time.

ClusteRim has been applied on a subset of the provided dataset, which has been created by selecting only those messages containing symptoms of illnesses like *pneumonia*, *fever*, *flu*, and more. The list of symptoms has been created based on the solution provided by the creators of the dataset[1]. We assume that these messages are directly connected to the epidemic. The center of the viewport has been set to Downtown Vastopolis, which is roughly the geographic center of the

[1] Available at `http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/solution`

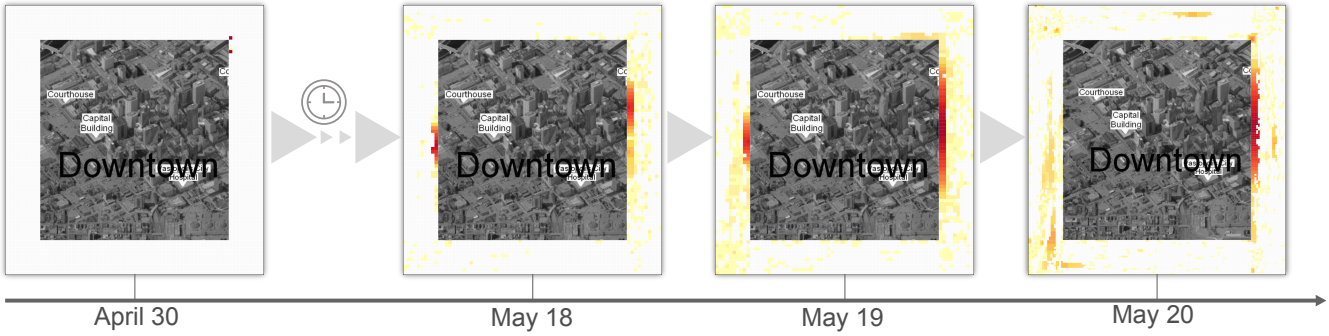


Fig. 10. Off-Screen visualization of the VAST Challenge 2011 microblog data subset. On the left, the situation before the epidemic is shown. The three images on the right show the development of the epidemic, ending at the last day for which data is available.

city. This guarantees that the messages are distributed in all areas of the off-screen space, while maintaining the focus at the crowded center of the city. The visualizations for each of 21 days are created (Figure 10 shows snapshots of four days).

On the leftmost visualization of Figure 10, the off-screen visualization area is almost empty. This means, that on April 30, only very few data points referring to an illness has shown up. The days between April 30 and May 18 are not visualized, because they look very similar to the visualization created from the data of April 30.

The visualization of data from May 18 shows the beginning of the epidemic. Yellow grid cells are popping up all over the off-screen visualization area. The red areas on the western and eastern area of the visualization stick out since the epidemic starts in Downtown. This is reflected in the small distance from the viewport to the red cells in the off-screen visualization area.

On May 19, one can see that the colored grid cells are visible, compared to the day before. These new cells are noteworthy, because they represent areas of the city which had been affected by the epidemic over night. The larger number of reddish cells on the western and eastern visualization area of the viewport show the rising number of messages dealing with symptoms of the epidemic.

The last visualization displays the data from May 20. On this day, the epidemic is on its peak, which can be easily seen when comparing the visualizations from the days before. It is also clearly visible, that the epidemic has been spread to a bigger area. The single reddish cells on the off-screen visualization area represent the hospitals distributed in the city. The area on the south west is also getting affected, the messages sent from along the lower river area are clearly standing out.

We have shown that ClusteRim visualizations of the VAST Challenge are clear in their visual representation and allow easy interpretation and fact finding. This example also shows that the visualization technique is suitable for time-oriented visualizations.

6.1.3 Mobile Application

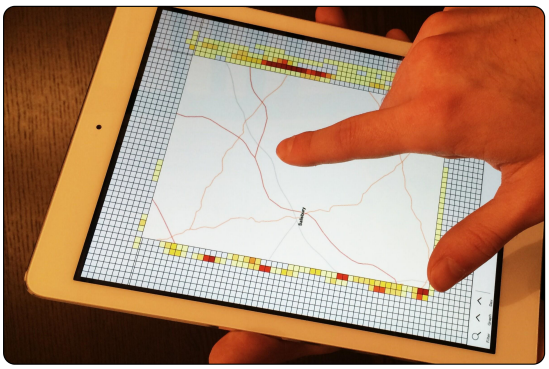


Fig. 11. ClusteRim technique on a mobile device.

To enhance context-awareness en route, ClusteRim also meets the requirements of mobile applications. Figure 11 shows our technique on a mobile device using a large geo-spatial dataset that contains critical infrastructure information in North and South Carolina. The information consists of 48 different infrastructure features such as the cellular towers, transportation, water and sewer, petroleum and oils. A total of 500,000 feature points were tested in our mobile environment.

6.2 Exploring Large Spatial Data Spaces

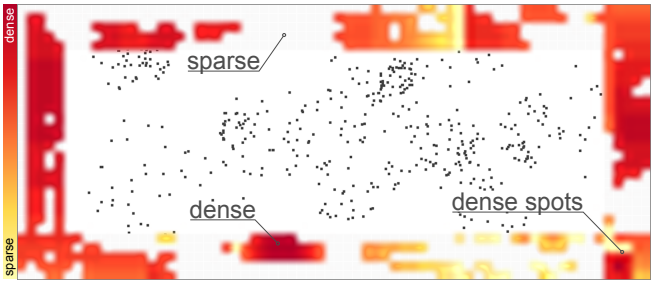


Fig. 12. Exploration of a scatterplot consisting of five thousands data points. Context is maintained while the viewport is updated. The analyst can easily identify dense spots as well as dense and sparse areas.

In our last example, we stress the generalizability of ClusteRim. Figure 12 presents a scatterplot consisting of twenty thousand data points. Here, the continuous representation in the off-screen visualization area is used, which interpolates the cell colors bicubically [30], as described in chapter 4. The off-screen space holds a total of 780 clusters, which are not visible in standard scatter plots.

Despite the high zooming factor of approximately one tenth of the data space, the off-screen visualization manifests spots of interest as well as dense and sparse areas. On the bottom right, five very dense spots that stick out can be identified. Also, the differences between the top right and the bottom left area are striking. Whereas the bottom left area exposes very dense point clusters, the top and top right areas is very sparse by means of the amount of points in total.

Also, sparse areas can contain dense spots, as it can be seen in the bottom right corner. These characteristics are helpful in getting a detailed idea of the data which is not displayed in the current viewport.

7 Discussion

Following, we address the advantages and limitations of our approach. As we showed in the user study, ClusteRim has many advantages when the user is overloaded by the amount of presented data. It is particularly useful when the data space is much larger than the view space. In this case it allows effective and efficient navigation in the data space. Our technique does not make use of the viewport itself but adds an extra border area. The off-screen data visualized is first aggregated in a

clutter-free manner, which makes it easy to represent it in a pleasing way on this border region. When the data space is too large for the view space, the distortion of the off-screen data might cause a suboptimal result. Thus, the parameters can be steered toward user's needs. We might additionally use bifocal display techniques [2] for a subset of grids based upon corresponding cluster's densities.

Special attention must be given to corner areas of the off-screen visualization space. In this paper, we applied the same mapping strategy to the corner and non-corner areas. But, as we pointed out in chapter 4, the off-screen visualization spaces in the corners cover a majority of the entire data space. To address this issue, we propose the following solutions: Instead of using a rectangular viewport, we can treat the entire off-screen space equally with a circular viewport. Another possibility is to assign more grid cells to the corner regions and to shrink the amount of remaining cells according to the off-screen space.

Besides advantages of ClusteRim in context-aware tasks, informal interviews conducted along with the user study revealed, that ClusteRim is more intuitive and the context-preservation worked for the participants. Even with a small viewport, the context remains. The context also depends on the size of the aggregated clusters. If the cluster size exceeds the dimensions of the viewport, we can not guarantee an ideal visualization. Not only the cluster dimensions but also the color coding contributes to context-awareness. The normalization as well as the color map can be customized according to users' needs, data characteristics, and application requirements.

Steering multiple parameters makes ClusteRim applicable and supportive to visual analytic systems that deal with large spatial data sets and look out for context-awareness whilst exploration. Especially in case of streaming data there is need to preserve context while exploring continuously updating data. Visual analytics systems struggling with this problem can be found in many interdisciplinary domains, for instance for text data [29] and event data [31]. Instead of squeezing the space when new updates catch on, we can simply apply further improved off-screen techniques like ClusteRim to maintain focus and context at the same time.

8 Conclusion and Future Work

In this paper we present ClusteRim, a novel technique for off-screen visualization. In contrast to current state-of-the-art techniques like HaloDot, our technique aggregates off-screen objects data and places them on a dedicated rim outside the main viewport, which leads to a clutter-free visualization. The projection of off-screen objects onto the view space preserves the spatial context and distribution of the data points according to the grid precision. The color mapping algorithm guarantees that the visualization maintains the approximate shape and density of points of interest. In addition, each part of our technique can be steered via different parameters, which makes ClusteRim suitable for visual analytics applications. Furthermore, we introduced a terminology for off-screen visualizations and showed its applicability. This can be useful for other researchers working in this area.

To continue our work, we plan to evaluate more of the design decisions we made. For example, we want to investigate the impact of different number of segments or the cell size on the context-awareness of users. The effect of different sized cells is also an interesting area for further research. Especially in context tasks, the implications of non-linear projections from the off-screen space to the visualization can certainly be of interest to study. We also want to build a visual analytics system that incorporates ClusteRim to visualize large data spaces with classic scatter plots, geo spatial, and text based visualizations.

References

[1] C. Ahlberg and B. Shneiderman. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In C. Plaisant, editor, *CHI Conference Companion*, page 222. ACM, 1994.

[2] M. D. Apperley, I. Tzavaras, and R. Spence. A bifocal display technique for data presentation. In *Proceedings of Eurographics*, volume 82, pages 27–43, 1982.

[3] P. Baudisch and R. Rosenholtz. Halo: a Technique for Visualizing Off-Screen Objects. In G. Cockton and P. Korhonen, editors, *CHI*, pages 481–488. ACM, 2003.

[4] S. Burigat and L. Chittaro. On the effectiveness of Overview+ Detail visualization on mobile devices. *Personal and Ubiquitous Computing*, 17(2):371–385, 2013.

[5] S. Burigat, L. Chittaro, and S. Gabrielli. Visualizing Locations of Off-Screen Objects on Mobile Devices: A Comparative Evaluation of Three Approaches. In M. Nieminen and M. Röykkee, editors, *Mobile HCI*, pages 239–246. ACM, 2006.

[6] S. Burigat, L. Chittaro, and A. Vianello. Dynamic visualization of large numbers of off-screen objects on mobile devices: an experimental comparison of wedge and overview+detail. In E. F. Churchill, S. Subramanian, P. Baudisch, and K. O'Hara, editors, *Mobile HCI*, pages 93–102. ACM, 2012.

[7] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Academic Press, 1999.

[8] S. Carpendale, J. Ligh, and E. Pattison. Achieving Higher Magnification in Context. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*, UIST '04, pages 71–80, New York, NY, USA, 2004. ACM.

[9] S. Carpendale and C. Montagnese. A Framework for Unifying Presentation Space. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 61–70. ACM, 2001.

[10] M. Czerwinski, A. M. Lund, and D. S. Tan, editors. *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008*. ACM, 2008.

[11] N. Elmqvist and J.-D. Fekete. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Trans. Vis. Comput. Graph.*, 16(3):439–454, 2010.

[12] N. Elmqvist, N. Henry, Y. Riche, and J.-D. Fekete. Melange: space folding for multi-focus interaction. In Czerwinski et al. [10], pages 1333–1342.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *KDD*, pages 226–231. AAAI Press, 1996.

[14] M. Frisch and R. Dachselt. Off-Screen Visualization Techniques for Class Diagrams. In A. Telea, C. Görg, and S. P. Reiss, editors, *SOFTVIS*, pages 163–172. ACM, 2010.

[15] G. W. Furnas. Generalized Fisheye Views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[16] G. W. Furnas and B. B. Bederson. Space-scale diagrams: Understanding multiscale interfaces. In I. R. Katz, R. L. Mack, L. Marks, M. B. Rosson, and J. Nielsen, editors, *CHI*, pages 234–241. ACM/Addison-Wesley, 1995.

[17] P. Games. Visualization of Off-Screen Data on Tablets Using Context-Providing Bar Graphs and Scatter Plots. *Boise State University Theses and Dissertations*, Aug. 2013.

[18] S. Ghani, N. H. Riche, and N. Elmqvist. Dynamic Insets for Context-aware Graph Navigation. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'11, pages 861–870, Aire-la-Ville, Switzerland, Switzerland, 2011. Eurographics Association.

[19] T. Gonçalves, A. P. Afonso, M. B. Carmo, and P. P. de Matos. Evaluation of HaloDot: Visualization of Relevance of Off-Screen Objects with over Cluttering Prevention on Mobile Devices. In P. Campos, T. C. N. Graham, J. A. Jorge, N. J. Nunes, P. A. Palanque, and M. Winckler, editors, *INTERACT (4)*, volume 6949 of *Lecture Notes in Computer Science*, pages 300–308. Springer, 2011.

[20] T. Gonçalves, A. P. Afonso, M. B. Carmo, and P. P. de Matos. Comparison of Off-screen Visualization Techniques with Representation of Relevance on Mobile Devices. page 9, 2013.

[21] T. Gonçalves, A. P. Afonso, M. B. Carmo, and P. Pombinho. *HaloDot: Visualization of the Relevance of Off-Screen Objects*. SIACG, 2011.

[22] S. Gustafson, P. Baudisch, C. Gutwin, and P. Irani. Wedge: Clutter-Free Visualization of Off-Screen Locations. In Czerwinski et al. [10], pages 787–796.

[23] S. Gustafson and P. Irani. Comparing visualizations for tracking off-screen moving targets. In M. B. Rosson and D. J. Gilmore, editors, *CHI Extended Abstracts*, pages 2399–2404. ACM, 2007.

[24] M. Harrower and C. A. Brewer. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, June 2003.

[25] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis.

Communications of the ACM, 55(4):45–54, 2012.

[26] R. Hoffmann, P. Baudisch, and D. S. Weld. Evaluating Visual Cues for Window Switching on Large Screens. In Czerwinski et al. [10], pages 929–938.

[27] E. H. hsin Chi and J. Riedl. An Operator Interaction Framework for Visualization Systems. In *INFOVIS*, pages 63–70. IEEE Computer Society, 1998.

[28] H. Jo, S. Hwang, H. Park, and J. hee Ryu. Aroundplot: Focus+context interface for off-screen objects in 3D environments. *Computers & Graphics*, 35(4):841–853, 2011.

[29] D. A. Keim, M. Krstajic, C. Rohrdantz, and T. Schreck. Real-Time Visual Analytics for Text Streams. *IEEE Computer*, 46(7):47–55, 2013.

[30] R. Keys. Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(6):1153–1160, 1981.

[31] M. Krstajic, E. Bertini, and D. A. Keim. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2432–2439, 2011.

[32] Y. K. Leung and M. D. Apperley. A Review and Taxonomy of Distortion-Oriented Presentation Techniques. *ACM Transactions on Computer-Human Interaction*, 1(2):126–160, 1994.

[33] S. Lohr. *Sampling: Design and Analysis*. Advanced (Cengage Learning). Cengage Learning, 2009.

[34] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The perspective wall: detail and context smoothly integrated. In S. P. Robertson, G. M. Olson, and J. S. Olson, editors, *CHI*, pages 173–176. ACM, 1991.

[35] T. May, M. Steiger, J. Davey, and J. Kohlhammer. Using Signposts for Navigation in Large Graphs. *Comput. Graph. Forum*, 31(3):985–994, 2012.

[36] T. Moscovich, F. Chevalier, N. Henry, E. Pietriga, and J.-D. Fekete. Topology-Aware Navigation in Large Networks. In D. R. O. Jr., R. B. Arthur, K. Hinckley, M. R. Morris, S. E. Hudson, and S. Greenberg, editors, *CHI*, pages 2319–2328. ACM, 2009.

[37] C. Myers and D. Duke. Navigational overlays for network analysis.

[38] E. Pietriga and C. Appert. Sigma lenses: focus-context transitions combining space, time and translucence. In Czerwinski et al. [10], pages 1343–1352.

[39] M. Sarkar and M. H. Brown. Graphical Fisheye Views of Graphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pages 83–91, New York, NY, USA, 1992. ACM.

[40] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Visual Languages, IEEE Symposium on*, volume 0, page 336, Los Alamitos, CA, USA, 1996. IEEE Computer Society.

[41] P. Zellweger, J. D. Mackinlay, L. Good, M. Stefik, and P. Baudisch. City Lights: Contextual Views in Minimal Space. In G. Cockton and P. Korhonen, editors, *CHI Extended Abstracts*, pages 838–839. ACM, 2003.

Ensemble Visual Analysis Architecture with High Mobility for Large-Scale Critical Infrastructure Simulations

Category: Research

Abstract—Nowhere is the need to understand large heterogeneous datasets more important than in disaster monitoring and emergency response, where critical decisions have to be made in a timely fashion and the discovery of important events requires an understanding of a collection of complex simulations. To gain enough insights for actionable knowledge, the development of models and analysis of modeling results usually requires that models be run many times. Central to the goal of our research is, therefore, the use of ensemble visualization of a large scale simulation space to appropriately aid decision makers in reasoning about infrastructure behaviors and vulnerabilities in support of critical infrastructure analysis. This require the bringing together of the computing-driven simulation results with the human decision-making process via interactive visual analysis. We have developed a general critical infrastructure simulation and analysis system for situationally aware emergency response during natural disasters. Our system demonstrates a scalable visual analytics infrastructure with mobile interface for analysis, visualization and interaction with large-scale simulation results in order to better understand their inherent structure and predictive capabilities. To generalize the mobile aspect, we introduce mobility as a design consideration for the system. The utility and efficacy of this research has been evaluated by domain practitioners and disaster response managers.

Index Terms—Disaster Forecast, Critical Infrastructure Simulation, Visual Analytics, Mobile Interface

1 INTRODUCTION

Forecasting the destructive impact for a volatile hurricane to a network of vulnerable critical infrastructure is a central challenge for emergency planners and responders in hurricane-prone areas. After witnessing the devastating destruction from Hurricane Sandy, decision makers in coastal US cities are on high-alert for threats to their critical infrastructures (e.g, power lines, food networks, shelters, etc.); they are requesting more robust simulation analyses to depict the potential impacts from another tropical storm.

Much prior research has focused on using simulations and predictive modeling to anticipate hurricane movement and suggest possible landfall and impact locations [36, 22, 8]. Due to the complexity and scalability of simulation runs, understanding these modeling efforts and their predictive capabilities from large collections of simulation results is challenging [30]. On the one hand, many of the traditional hurricane modeling approaches depend on a trail-and-error effort that is not always feasible for generating simulations consistently. While this type of approach is widely adopted, each simulation run requires analysts to fine-tune the parameters, which can be very time consuming and less methodological [6].

On the other hand, a new simulation approach is based on the idea of data-farming, which prepares many possible simulation outcomes in bulk [14]. However, this approach is largely limited by the simulation models, for which very few modelers have sufficient computing resources available to do sensitivity studies, validation and verification, effectiveness analysis, and related necessary activities. Exacerbating this challenge is that the large amount of simulation results is far outpacing decision makers capability to analyze and make use of them.

To meet the challenges of dealing with disaster forecast and preparation, automated simulation methods are essential. However, they are not sufficient. There must be human input and direction, specifically for the interpretation of results and in some cases for directing the simulations. It is important for the decision makers to gain enough actionable knowledge such that they can respond in a timely and correct manner to possible failures of crucial infrastructure.

A key design component for our system is *Ensemble Visualization* of a large simulation space generated as a result of the aforementioned data-farming simulation approach. An ensemble, in this case, is a high-dimensional collection of attributes aggregated from raw simulation results that are centered around a single feature (e.g., Power Station, Airport, or Hospital). Details of the attributes are illustrated in Figure 2.

Another key design consideration is *Mobility*, which, in our case, refers to mobile computing devices or environments that enable analysis in the field. It doesnt limit to just personal devices (e.g., iPad), but more broadly includes moveable equipment in general. The demand for mobility has been demonstrated in in our previous police and evacuation exercises [13], where our campus police employed a networked tablet, smart phones, and a mobile command center to provide situationally-aware support.

Central to the goal of our research is, therefore, the use of ensemble visualization of a large scale simulation space to facilitate decision makers reasoning about impact on critical infrastructures with mobility. It is in this spirit that we architected our scalable visual analytics system for analysis, visualization and interaction with large-scale simulation results in order to better understand their inherent structure and forecast capabilities. Our system captures the interplay between cascading simulations of critical infrastructure, ensemble analysis, and a networked mobile visual analytics interface. It aims to balance both *human and computer intelligence* and provide situational awareness to decision makers in both planning for natural disaster response and taking direct action during a disaster.

1.1 Research Procedures and Aims

In line with our design goals, the first activities of our research focused on establishing a simulation space based on computationally traversing possible simulation permutations within Cascading Infrastructure Simulation (CIS). Specifically, our system relies on massive amount of computer generated simulation results. Access to these results is crucial because a working system requires information about the variance of infrastructure impact within large geospatial areas, which are produced from the cascading model by varying input parameters.

As detailed in Section 4.1, we have developed a CIS model that runs multiple hurricane paths to generate a raw simulation space with millions of infrastructure failures events. Currently, our simulation includes infrastructures such as electric power, telecommunications, water, and railroad transportation. Large amounts of failures events from these infrastructures provide the fundamentals for ensemble analysis, which enables a multi-scale exploration of the high-dimensional simulation space connecting the complementary insights from global and local analysis of the data. In addition, our simulation space is situational because it depends on the properties of the particular coupled infrastructure and will also depend on the properties of the occurrence (which can vary) that brings stress to the critical infrastructure.

By using an ensemble of simulation runs, we enable users to sample the space of input conditions that is presumed to cover all possible starting conditions in a particular range, and then employ multiple models that provide greater or lesser fidelity in some aspect of the process. The resulting ensemble encompasses the range of plausible outcomes, and the variation within the ensemble members exposes information on simulation uncertainty and the sensitivity of input parameters.

To incorporate human-centered analysis, we further provide an interactive visual analysis interface for exploring the simulation space. The designed visual interface combines a variety of statistical visualization techniques to allow decision makers to quickly identify areas of interest, ask quantitative questions about the ensemble behavior, and explore the uncertainty associated with the data.

To meet the needs for mobility, we have encapsulated the interface into mobile devices, such as an iPad, with multiple coordinated visualizations to provide a cohesive view of the simulations and permit analysis at multiple scales. The coordinated visualizations and interactions are optimized for mobile display, following the design guidelines from Apple [2]. The complexity of the ensemble data will be mitigated by a flexible organization of the information, and the coordination between views will permit users to focus on the formulation and evaluation of hypotheses. Specifically, our system is designed to help domain users address the follow questions:

- Which facility is more vulnerable than others?
- Howe can responders in the field check buildings along a hurricane path in their local area?
- How is the hurricane's temporal development affecting the cascading infrastructure failure?

The rest of the paper is structured as following. We first characterize the research domain in Section 2, then describe related work in Ensemble Visualizations in Section 3. Details of our system will be introduced in Section 4. We provide our informal evaluation with disaster prevention planners in Section 5 and provide discussions in Section 6.

2 Domain Characterization and Design Considerations

2.1 Ensemble Analysis of Simulation Space

The architecture we present in this paper aims to help monitor and adapt to infrastructure changes by addressing the hurdle brought by the sheer size and heterogeneity of the relevant critical infrastructure simulations for the state of North Carolina. Recent disasters have highlighted the great vulnerability of both coastal and inland communities, such as power outages and road blockages caused by Hurricane Irene and Hurricane Isabel (which surged inland through Charlotte). It is important to identify the weak links within the massive infrastructure networks before the next hurricane hits. Road maintenance plans, development policies, hazard mitigation, and emergency response plans depend upon an understanding of the scale and linked cascading effects from hurricane impacts on critical infrastructures (e.g.,links between power substations and water services).

Managing critical infrastructure simulation is a complex, multistage process. Understanding the impact on infrastructures requires an effective workflow that includes data acquisition, de-noising, analysis, visualization and interaction. Based on inputs from our collaborators, one major challenge within a CIS process is that available tools are inadequate for analyzing and managing such large simulation results. Based on a survey conducted by the National Oceanic and Atmospheric Administration (NOAA), 57 out of 198 coastal infrastructure managers showed the need for a decision-support systems that can help them monitor the potential impacts for each infrastructure from all simulation results [28].

Our ensemble visualization is, therefore, setup to help understanding how changes in the hurricane discourse alter the stability of infrastructures. We emphasized on providing the ability to effectively and (semi-)automatically depict temporal and geospatial changes of coastal infrastructure caused by both historical and simulated natural disasters. Furthermore, we aim to enable users to view and interact with the simulations datasets, as well as the analysis results, in an unencumbered and intuitive fashion.

2.2 Interactive Visual Analytics Interface with Mobility

Mobility, on the other hand, is also another important aspect in our research effort. Through prior work in evacuation exercise, we have first hand experiences with our campus police to design and deliver networked mobile visualizations system using iPhone [13]. Specifically, the campus police chief and his force requires on-the-go analysis while events are occurring. The police chief stressed the benefits to have a mobile app on his iPad that he could use in conjunction with the setup he already employs in mobile command center. This would greatly improve their response time and keep them constantly connected while in the field.

Other portable devices like laptops are not as mobile as a tablet and are encumber some for the first responders if they were to walk around on foot. A tablet is easy to carry around and allows the responder to access what he needs at any time and place. To this regard, the chief even went further and claimed analysis system with mobility as an invaluable tool to integrate into his current methods.

During the first two years of our collaborations with responders and planners, we enumerated some of the simulation and data analytics challenges, and consider it critical to identify the infrastructure vulnerabilities at greatest risk before a disastrous event, so that natural disaster impact can be planned for and mitigated. Motivated by the above domain characteristics, our partnership, therefore, focuses on the research on ensemble visual analytics technology over massive amount of infrastructure simulations, with the goal of providing mobile visual interfaces that support analytic workflows which yield more accurate and effective results.

3 Related Work

Our work is built upon the concept of Ensemble Analysis. A specific research area that deals with ensemble data sets is climate and weather data visualization. An ensemble dataset is defined as a collection of simulation features that are generated by computational simulations of one or more state variables across space and time [24]. As temporal analysis assumes an ordered sequence of functions, a slightly different analysis comes into play when no ordering is imposed on a set of functions. This type of data often arises not from time-series data, but rather an ensemble of simulations. For example, in our targeted critical infrastructure simulation, stochasticity or randomness within the simulation setup may results in different observations following different runs. Such varying observations form an ensemble of functions.

The variation among the ensembles arises from the use of different numerical models, input conditions, and parameters. The complex nature of ensemble data sets leads to numerous possible approaches to visualization. Multivariate correlation in the spatial domain is a common approach for reducing the complexity of the task of data understanding, as is reducing the data to a hierarchical form which is conducive to 2D plots [17]. Büerger and Hauser [4] present an overview of techniques for multivariate data and Buja et al. [3] discuss a taxonomy of interaction with high-dimensional data.

One important challenge is to understand how the stochasticity or randomness within the simulation design impact the infrastructure analysis of multiple runs. This could be considered as an uncertainly analysis that explores the relationship between input parameters and the output [11, 25]. One strategy is to derive a statistical description of these ensemble structures; for example, summarizing each simulation runs and describe the statistical distribution of these entire ensemble simulation space. There has been some recent effort in conducting statistical ensemble research, such as computing certain statistics on topological summaries [18, 35].

Ensemble analytics and visualization is a new and rapidly growing field where researchers in statistics and computational topology have just begun to investigate, and a wealth of questions remain open.

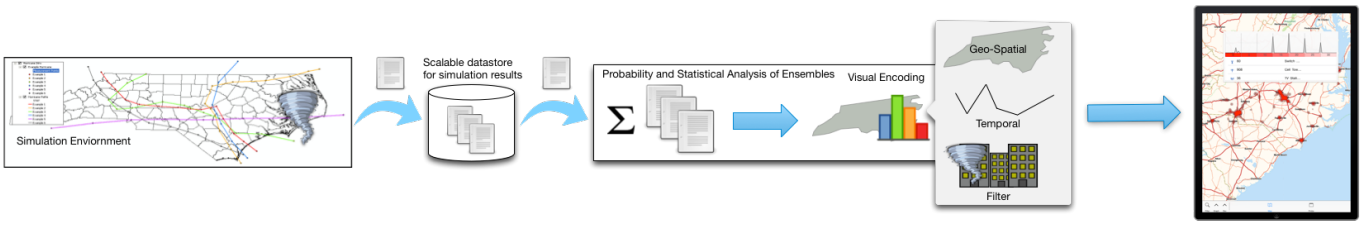


Fig. 1. System Architecture of our Ensemble Vis.

Software systems, such as SimEnvVis [21] and Vis5Ds [15], are designed to handle atmospheric data formats and include 2D geographical maps with color maps and contours, as well as more sophisticated techniques such as iso-surfacing, volume rendering, and flow visualization. The Noodles system [31] provides the aforementioned capabilities and adds uncertainty contours and glyphs. The strength of ensemble analysis is that it includes little pre-processing cost and easy extension and adaptability to new domains and techniques. In addition, Ensemble-Vis [23], uses meteorological data to provide a visual exploration of short-range weather forecasting. It adapts visualization methods common to the domain of meteorology; it also adds indications of uncertainty and enables the user to drill-down to ensemble data.

Given the need to integrate human knowledge in the analysis process, our work also relies on the ensemble feature integration. Here, we focus on the combination of interactive visualization and feature analysis that has resulted in a set of feature integration and exploration techniques for analyzing multi-dimensional simulation space. Prior work in this area can be grouped into three categories, two-variate visualization, multivariate visualization, and analysis animation [38], depending on the nature of the simulation data. The utility of these techniques has demonstrated great success in the analysis fields of volumetric data [29], terrain change detections [5], and medical imaging [20]. However, these traditional feature integration techniques no longer scale with the increasing size and complexities of the simulation datasets. Our research aims to address several challenges in this research area including feature selection and comparison, interactive exploration and semi-automated annotation.

Need Update Through further integrating knowledge gained by one of our previous work done for a mobile crowd sourcing application to model 3D buildings [7], we established our networked visual analytics architecture to build a better way to handle such mobility requirements and enable data traffic, user interaction, and visualizations. Multiple linked views of data relieve the need to present all data of interest in a single window [1]. Such approaches let the user interactively select regions of interest and reflect those selections in all related mobile views. The resulting mobile interface provides a collection of views which provides complex investigation of the data by allowing the user to drive the data analysis.

4 System Components and Implementation

The focus of our system design is to facilitate domain experts in analyzing millions of simulation results in order to identify the vulnerability and resiliency of a critical infrastructure. As shown in Figure 1, our networked visual analytics system is comprised of three integrated components: A cascading CIS model connected to a shared Oracle database that generates the raw simulation space (Section 4.1); An ensemble analysis module handles statistical computation of ensemble structures, and provides optimization and data retrieval api for the visual analysis module (Section 4.2); and finally a mobile visual analysis interface that encodes the ensemble results into multiple coordinated visualizations to ensure mobility when conducting field examinations and preparations (Section 4.4). The following sections describe each of the components in greater details.

4.1 CIS: Generating the Raw Simulation Space

To capture these complex, multifarious, and dynamic effects from a hurricane, we utilize a CIS simulation model [blinded for review] that takes into account the interrelationships among critical infrastructures. Built within a rule-based framework for integrating multiple infrastructure components at a high level, our model allows the user to create a model which represents the users assumptions about the world.This results in a dependency/interdependency ontology. Thus, for example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a school were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down.

In addition, our model handles the concept of output requirements which allows the user to specify which inputs are necessary to produce the specified output from the feature. This is especially useful to help users gain focus on the commodities that matters most to their planning, rather than noised by other simulation results. For example, a power-grid specialist can select a subset of inputs that only applies to the multi-linked power-station with his jurisdiction and run partial simulation around it.

4.1.1 Model Inputs and Simulation Concepts

Our CIS model consists of a wide range of data inputs, such as selection sets, commodities, relationships, networks, and latencies. To keep the model results as realistic as possible, we spent over 6-month period and conducted a thorough collection of a range of infrastructure datasets within the Carolinas.

Specifically, *selection sets* are the feature's involved in the model such as communication (e.g., cell towers), electric network, and services network (e.g., hospitals and restaurants). The majority of this data comes from FEMA HAZUS data disks from 2005 [10]. The cellular towers, communication facility and TV stations were provided by the FCC [33]. We further generated the switch control data for the communication network based on ESRI's random point generator [9].

Our electrical network features are a mixture of HAZUS data points for power generation facilities and a human-annotated network, where we employed undergraduate students to digitize the transmission power lines and the substations from from orthophoto satellite images [12]. Once the network was completed Thiessen polygons were created to create electric service area's operating under the assumption that electric is provided from the nearest potential provider.

The service network data are largely collected through out domestic and international collaborators [19]. The food features where provided by a project partner. In our current implementation there are 48 different infrastructure features such as the hurricane paths, transportation, water and sewer, petroleum, and oils.

To simulate the cascading effect, our model takes in the concept of *commodity*, which is a tangible or conceptual good flowing between selection sets such as electric, water, natural gas, manufactured goods, personal relationships, wealth, etc. Commodities are the foundation of the model.

As shown in Figure 2, we have collected a total of 15 commodities and 80 relationships between them. *Networks* are, therefore, the systems of selection sets that all share similar connection properties. The electric grid, road and rail systems pipeline are all modeled as different networks with distinct properties that best reflect their actual

name	Meaning
id	Feature ID stored in DB
Object_Class_Name	Commodity Type (e.g. Fire station)
Simulation_ID	Index Number for Simulation
Commodity_Name	Type of the Commodity
Indirect_Cost	Is the commodity disable indirectly?
Direct_Cost	Is the commodity disable directly?
Relative_Time_In_Hours	Time of Disablement
Number_Hours_Disabled	How long is the commodity disable during the simulation
Cause_Event_ID	Disable Event ID
COA_ID	Cause of Action ID
Model_ID	Which Hurricane Path Model
Network	Which Infrastructure Network (e.g.,Power or Transportation)

Fig. 2. Detailed dimension information for each Ensemble.

applications.

4.1.2 Cascading Relationship for Model Accuracy

This collection of networks and commodities gives us a good ground of performing realistic cascading critical infrastructure simulation. Within each network, a commodity flows between two selection sets through the creation of a relationship. The relationship establishes the provider and consumer of the commodity, the criticality of the good transfer and the method for establishing relationships.

These interlaced critical infrastructures are captures in a set of networks with each node having a set of properties according to its category and the edges providing a dependency rule according to the category and state of the two connected nodes. For example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a hospital were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down. In this way, our CIS model takes cascading events into account.

To enable domain experts to hypothesize different cascading conditions, we have built four methods create a relationship for our model: explicit, nearest neighbor with/without redundancy, and finally spatial. The users can directly specify the direct link between two features through an *explicit relationship* between the origin and destination. While the explicit method specifies a strict point-to-point relationship, the users can also use the *nearest neighbor method* to create connections between the origin and destination features that are within a specified distance. *Redundancy option* is built to further narrow down the connectivity between features. A relationship with redundancy means that the destination feature is connected to ALL sources within the distance (e.g., cell towers), while without redundancy means the destination feature is only connected to the closest provider (e.g., water pipelines). For example a hospital could connect to any sewer pipe within 1000 meters but it only creates a relationship with the sewer pipe closest to it. Finally, a *spatial relationship* are instances where the origin and destination share a relationship based on their geo-spatial locations. In our model all electric relationships are based on spatial relationships.

4.1.3 Simulation Runs and Results

Our simulation runs with on a set of Courses of Actions (CoAs), which is an pre-determined encapsulation of the above factors in to reflect the state changes for particular critical features in the storm path. Through a COA, a user can specify events that occur during the simulation and inform the CIS of What (selection set instance) and When (relative time into simulation) the selection set instance undergoes a state change (e.g., intact or affected). The user can make the COA actions permanent or temporary. A permanent event means that the feature will stay at the new state until another COA event changes the state. A building that has been physically destroyed is disabled and cannot be re-enabled even if all of the buildings inputs are available to it until the user specifies that the building has physically been rebuilt.

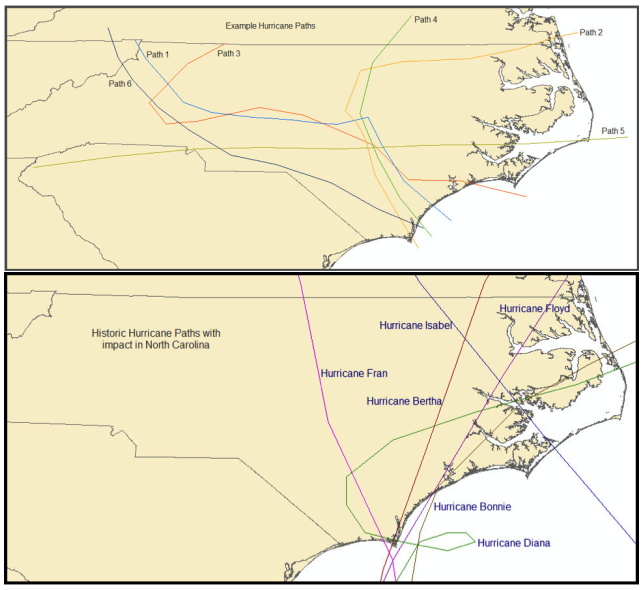


Fig. 3. Two models simulated with our critical infrastructure modeling. Top is the six notional Hurricane Paths. Bottom is the six Historical Hurricane Paths, including Bertha, Bonnie, Diana, Floyd, Fran and Isabel.

Latencies is an important part of the CoA, which is designed to represent the delays of a selection set undergoes a state change. These can account for backup electrical power or, in the case of food stores, time periods until food supplies are exhausted. For example, a hospital can have a disablement latency of 24 hours for blood supply but a 14 day electric disablement latency.

We further constructed six automatic model variations to ensure the possible conditions are simulated. This set of model variation allows the users to test their assumptions on distance and the importance of redundancy within the system. Specifically, the six model variations only represent changes to the relationships but not to any other aspect of the model. The first model in each set is the base model which represents the users best guess on structure and type of relationship between the selection sets. The second model converts all nearest neighbor without redundancy (NN) and converts them to nearest neighbor with redundancy (NNR). The third model doubles all relationship distances. The fourth model takes those new distances and converts any NN to NNR. The fifth model takes the distances in the base model and halves them. The six model takes those new distances and converts the NN to NNR.

In order to cover the simulation space, we run a large number of simulations with different hurricane paths and different intensities and spreads. In total, we have run the six models across 12 different hurricane paths, including six historic hurricanes paths (see Figures 3) and six notional paths to simulations. Doing so allowed the users to gain a comprehensive view of what happens for different storm strengths or by changing our damage radius assumptions. This hurricane paths are created together with our collaborators and give our simulation model a good coverage across the Carolinas.

With over 50,000 features, so far, we have completed 425 simulations over the 12 hurricane paths and created 14.6 million simulation events. As shown in Table 2, each event is a high-dimension collection of fields that are centered around a single feature (e.g., Power Station, Airport, or Hospital). All this data are stored in an Oracle database for remote access and analysis.

4.2 Extracting Ensembles from the Simulation Space with Statistical Analysis

While our CIS methods are essential of creating a large simulation result space, however, just the data is not sufficient. There must be human input and analysis, specifically at the interpretation of results.

It is important for the decision makers to gain enough insights to actionable knowledge such that they can respond and react to a possible failure of crucial infrastructure due to impacts from natural disasters. This became the primary reason for us to perform ensemble analysis for the raw simulation space.

Our ensemble analysis is used to abstract and reduce the complex and vast amount of information of each infrastructure. This enables the decision maker to understand the full crisis in its context and to detect potential cascading effects. It further permits the users to select the paths and CoAs most likely to occur for the most likely range of conditions. This will in turn indicates the most likely parts of the infrastructure to be disrupted and shows the likely cascading effects. For example, this gives the city planners information on "how likely is my airport going to be hit by this storm?"

We define our ensemble as a high-dimension collection of fields aggregated from raw simulation results that are centered around a single feature. Specifically, we compute an ensemble as a summation of disablement cross the entire simulation space for each specific infrastructure. The most important value is a probabilistic outage for each infrastructure. This value represents the likelihood for a specific infrastructure to be disabled across the entire simulation. We also compute the breakdown of causes for a specific infrastructure along with the minimum, median and maximum monetary cost imposed by the infrastructure being disabled.

In addition, multivariate correlation [17] is used to capture the stochasticity or randomness within the simulation setup and reduce the complexity of the simulation event data. As a result, we are able to abstract and encapsulate the ensemble of every infrastructure into an undirected graph. Its nodes are represented by symbols, such as the glyphs (see Section 4.4.2) for power stations and cellular transmission stations.

4.3 Scalability Optimization for Mobile Analysis

Due to the inherent hardware limitations of commodity mobile devices, creating the requested mobile visual interface with millions of simulation events is both a computation and a visualization challenge. Where desktop computers can perform in-memory operation of large data, a commodity mobile device (e.g., iPad) is quite limited in its computing power. Therefore, a significant part of our research has been devoted to conduct scalable mobile optimization, which heavily built around computing optimization and visualization scalability.

4.3.1 Offload Ensemble Encoding to Computing Server

Our first optimization aims to accommodate the need to interactively visualize the entire ensemble space and its temporal changes. To this regard, we have employed a computing server that is designated to performs specific statistical calculations and data retrieval;these operations would otherwise be too time consuming on any mobile device.

Our computing server is connected to the database that contains all the raw simulation results. It handles data requests sent from the mobile device to the server, extracts the necessary data and finally computes the ensemble abstraction and encoding. We use a hash map for data transfer between the mobile device and the server. Each infrastructure has a unique key identifier. Using a hash map allows for quick lookups when doing additional computation on the mobile device.

In order to examine the temporal changes, our server performs statistical analysis on the ensemble through kernel density estimation [16]. This is done using GPU parallel computing and significantly offloads the demand for performing on-device computing. We use OpenCL to compute each cell of the kde. This provides exponential speed up, but we have gone a step further since copying the computed data back from the GPU to main memory is usually more time consuming then the computation itself. To minimize data transfer between main memory and GPU memory we take advantage of being able to hand off the data computed from OpenCL to OpenGL for rendering. In the OpenCL kernel we can create geometry and then render it very quickly with OpenGL since it all exists in GPU memory. This allows us to not just create geometry very quickly but also pre-rendered bitmaps.

Data Points	10	100	1,000	10,000	100,000
Cells	4mil	4mil	4mil	4mil	4mil
Server	0.011s	0.026s	0.079s	0.475s	7.07s
Mobile	0.28s	4.497s	34.807s	355.953s	~

Fig. 4. This table shows the approximate time to perform kernel density estimation on varying data size in seconds. The spatial region was discretized into approximately 4 million cells. The mobile computation was performed on an iPad air using a CPU parallel implementation. The server computation was done using a Nvidia GeForce GTX 680MX using OpenCL.

4.3.2 Utilize Multi-core and Graphics Engine for Mobile Rendering

Some processing, however, must be done on the mobile device. These tasks usually involve data loading and retrieval as well as user interaction. Handling data can often be very time consuming and slowing the user interface is not conducive to our goals. When interacting with a mobile device any latency or slowness results in the user becoming frustrated and hinders their productivity. Mobile devices often have multiple cores like their desktop counter parts so it is absolutely crucial to utilize any available resources. We utilize parallelism for all our web requests to create asynchronous data pulls. We also load all data from the CPU to the GPU through a background processes. We purposely leave the user interface to it's own separate process in order to maintain a consistent experience.

Our mobile system also utilizes a low-level graphics engine specifically built for rendering the ensemble simulation space. It enables the device to render vast data sets very quickly. The graphics engine uses a scene graph style data structure for storing elements. It also utilizes graphics techniques like double buffering and constructing texture atlases on the fly for improved rendering.

Moreover, mapping systems included on most mobile devices are not capable of displaying lots of data very quickly and efficiently. They also are not portable between different platforms making cross-platform development difficult. We chose to develop our own mapping system in OpenGL making it very fast but also portable between platforms.

4.3.3 Remote Communications

Our mobile system is dependent upon the data pulled from the computing server. Any latency with this communication will effect the entire system. Due to this dependency we rely extensively on compression and aggressive caching to memory and disk. This limits potentially costly network requests and speeds up the mobile application. Since a smooth user interaction and experience is necessary.

We started with web services to query data on the fly, package the data and send it to the device. This method worked well with moderate dataset;as the data grew, however, the query time and packaging time began to slow down greatly. The response time had reached a certain threshold and at that point the application becomes unusable.

We, therefore, created a persistent environment, instead of solely rely on a temporary service that queries the database each time and repackages it. This persistent environment caches the packaged data in main memory. Querying the database can be improved through minor tweaks with caching and other features, but what substantially slowed down the entire process was each time the web service is called a new instance was spun up in main memory to load the queried data in order to package it into a form usable by the mobile device.

We utilize parallelism for all our REST requests to create asynchronous data pulls and provide three possible return types for scalability. Each type is specified by the client based on their needs and the clients capabilities. The first is a list of all points returned from the query stored in a compressed JSON. Returning all the points allows the client to run additional operations that might not be computationally

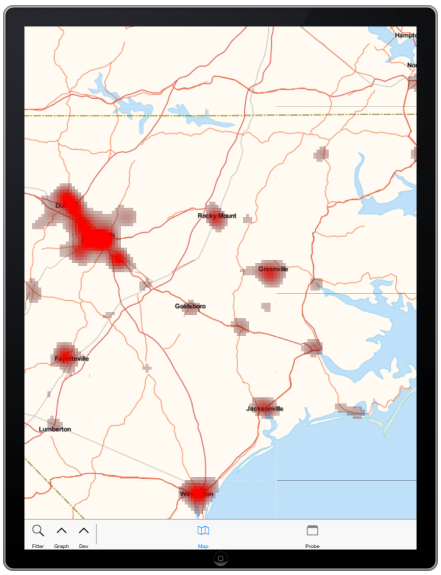


Fig. 5. The Geospatial View of our mobile interface. The Kernel Density Estimation is represented in Red clusters.

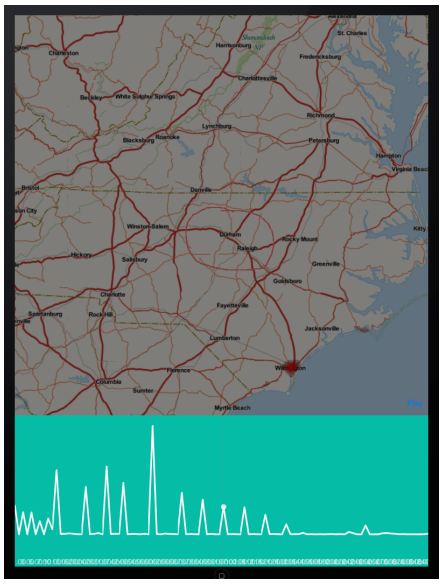


Fig. 6. This image shows the interactive Temporal View. In the view, the users can depict the peak time of the outages and select a range of time to further their investigations.

constrained. The second is a precomputed kernel density estimation that is computed into indexed triangulated geometry and compressed into binary for transfer. This data type is for the a client system that can not computationally calculate a spatial kde in sufficient time, but is capable of rendering the triangulated result. The triangulated result is resolution independent in order to provide a higher level of detail for visual analysis. The last type is a pre-rendered bitmap of the kernel density estimation. This is for client systems incapable of handling raw data points or even triangulated geometry.

All this scalability optimizations setup a feasible platform for us to encode and visually represent the ensemble analysis into a mobile visual interface.

4.4 Ensemble Visualization with Mobility

As the first step in our ensemble visualization design, we followed prior visual analytics design studies (e.g., [37, 32]) and collaborated with first responders to understand their workflows and analytics needs. Based on our discussions, it becomes clear that a key aspect of our development effort is to provide them with a visual analytics system that can accommodate their in-field analytics needs.

4.4.1 Geospatial-Temporal Overview for Ensemble Analysis

Customizable Geospatial View: Our simulation data set revolves around geospatial information. While mobile OS incorporates mapping services they are not specifically designed for displaying large data sets or any complex 2D drawing. For this reason we incorporated a lightweight tile map server specifically for our needs. A tile map server [34] simply allows a client to fetch pre-rendered geospatial map tiles on the fly from a server and display them in an arranged fashion much like Google maps.

Our customized tile map system, as shown in Figure 5 (A), bears the similarity with commercial mapping system, but also provide additional functionalities that fits our visual analytics needs. Since the map tiles are stored in the cloud and the visualization components are implemented as layers that can be stacked onto the map, our mobile interface remains lightweight. This enables us to completely customize the maps and it also allows us to imbed certain features in the tile map textures to reduce the redundancy of rendering by having all the essential map details baked into pre-rendered tiles.

As shown in the video, we provide two different visualizations the first is a time based color coding of each individual infrastructure based on the average time it went out. This allows first responders to see when specific events were disabled during a hurricane. The second visualization is an Kernel Density map []. The intensity map shows the probability of outages for a particular region. This way users can quickly focus on regions with a high percentage of outages.

Temporal Visualization: Understanding temporal behavior of a disaster is another important aspect for city planners to isolate time critical infrastructure. We have developed an interactive temporal analysis view that enables the selection of specific time ranges to depict what infrastructure was disabled as the hurricane passed through. As shown in Figure 6, the user can filter the simulation space on time and further examine the outages leading up to a peak or the peak itself.

Our system also enables the users to examine the result of cascading outages and interactively analyze the complex cascading relationship between different infrastructures. This allows the users to visually depict the interdependency of infrastructures and understand the effect of one infrastructure on others. Through the interactions, our collaborators were able to observe hurricanes with two distinct impacts patterns, namely the immediate sparse outages following high peaks of outages as well as the latency outage pattern where an infrastructure (e.g., hospital) will stay on with its backup generator for a longer period of time before becoming disabled. These are all key temporal factors for evacuation and planning effort.

Animation: We have further setup animations to help the users more effectively analyze the outage patterns. Previous research [27] has suggested animation plays an excellent roll in revealing changes where these cascading events could take place. This gives an overview of the events that took place and hotspots during the simulation. However, enabling animation posed an computational challenge given the 50,000 features and millions of events we have in our simulation space. As shown in Figure 4, computing each time segment visually exceeds the computation power of any existing commodity mobile device.

To address this challenge, we utilized streaming techniques that offloads much of the computation on our server and streaming the visualization in quick succession to a mobile device. As shown in our video, we have created animation of all the simulation events over a thirty day period with spatial kernel density estimation.

In addition, the same three data types: data points, geometry, and pre rendered bitmaps that are available for single queries are also available for animations. A client can request all the raw data points across the entire animation span or the triangulated geometry or pre-rendered bitmaps. This allows the animation to be flexible on a wide range of client systems and scalable even down to mobile devices.

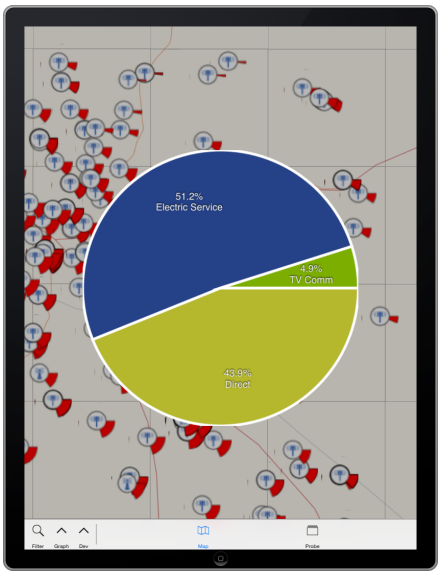


Fig. 7. Detail view shows breakdown of impact to a feature based on Ensemble Analysis.

Fig. 8. Example of glyphs used to visualize specific critical infrastructure.

4.4.2 Detailed View for Ensembles and Infrastructures

Each infrastructure in our simulation has a probabilistic chance that it will be disabled across the entire simulation space. When the user zooms in close enough we display a map icon with the probability of each Infrastructure as a pie chart, as shown in Figure 7. This allows the users to quickly see which infrastructure have a higher probability of being disabled. More importantly each feature might be disabled from multiple other events. By selecting an individual infrastructure a secondary pie chart will be displayed that shows individual causes for disablement broken down by percentage. This way a user can isolate the main cause for the infrastructure being disabled.

The glyphs for each infrastructure are created programmatically on demand based on the type of infrastructure and the summation of it's potential outage. Figure 8 shows a few examples of different types of glyphs. The inner icon is associated with what type of infrastructure. This allows users to easily discern what specific infrastructure they are looking at. The example shows a hospital, cell towers, a port, train station, and an airport. Secondly, the outer ring displays the probabilistic chance the infrastructure is disabled. A full circle would indicate a 100 percent chance of being disabled. This outer ring allows users to quickly identify what features have the greatest incident of being disabled.

4.4.3 Interactions

Probe Selection: Probing is the culmination of all interactions that allows users to visualize a selected region in a free-form manner. As shown in Figure 9, a user can directly draw onto the map with his or her finger drawing a bounding area around a region. This is an extremely important analysis tool because it allows the user to drive the analysis and focus on what is important to their needs. Anything within this bounding region will then be analyzed and returned to the user in a separate window.

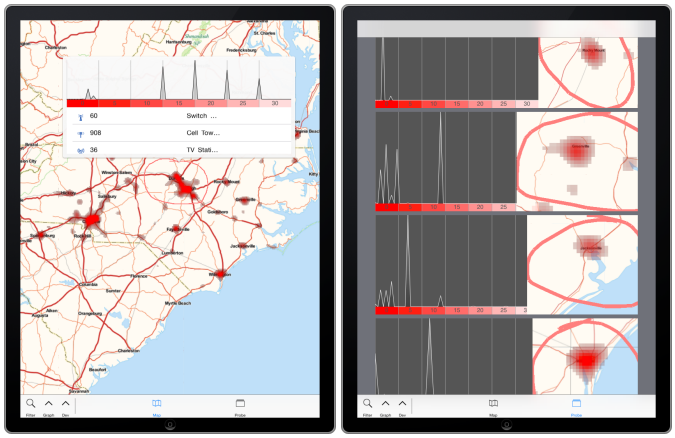


Fig. 9. The probe selection view (Left) and a comparative view for all the probe selections (Right).

When selected, a separate window will appear that initially shows the time distribution for all the features within the region. A user can quickly see any peaks in time from their selection. Secondly, a user can then view the break down of what feature classes are in the selection and how many were disabled (Figure 7). This allows users to see what features were disabled, how many were disabled and the time distribution of those features. This allows for a very fast analysis of specific regions that might be of interest to a user. It also allows users to pinpoint features that might have a high probability of being disabled. Most importantly it allows a user to compare multiple regions together to analyze.

Each probe selected by the user is copied and saved for additional viewing, as shown in Figure 9(Right). This way users can select multiple probes across many different areas and filters. Users can probe different time segments as well as different hurricane paths and commodities. Along with the selected features a screen shot of the region is attached to the data to give geospatial context to the region that was selected. All this information is presented in a secondary window where each probe is represented as a card with a map image of the selection and a time distribution of the events from the probe.

When the user selects a specific card the view transitions to a detailed break down of that specific probe. This way users can return and further analyze individual events and time distributions from prior selections they have made.

Filtering: In addition, filtering is a key function that allows for the removal of unwanted features or events, as shown in Figure 10. In such a large simulation space that we have it is crucial for a user to be able to focus on specific commodities and hurricane paths. More particularly the combination of the two together allows for complex filters. This helps users navigate the simulation space by visualizing different parts of the simulation in a very easy manor.

The first type of filter we have is a commodity based approach. Each commodity is separated into specific categories that share a common area. For example all the infrastructure like ports, train stations, airports, etc. is separated into a transportation category. By having abstract categories users can more easily pick what areas they would like to filter instead of navigating through very specific features. Grouping similar features into categories also allows for easier comparison. We allow users to filter specifically on commodity types like transportation and electricity. This allows for easy comparison between different commodities.

The second type of filtering is hurricane path based. Our model has numerous hurricane paths that create a probabilistic outcome. Users can select specific hurricane paths to see the probabilistic effect of each impact. But, the real core of this feature is the filtering of multiple hurricane paths. By filtering on multiple hurricane paths pertinent to a region a user can get very encompassing results for that region.

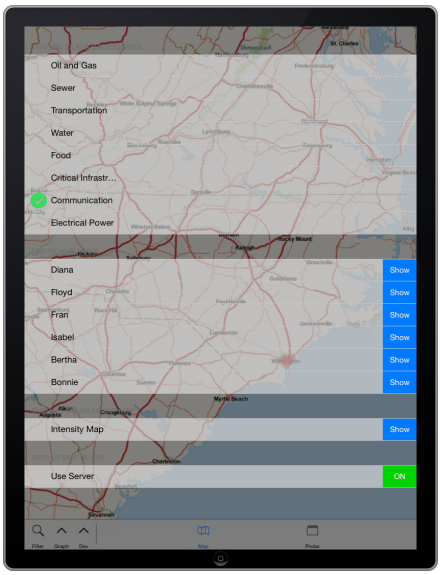


Fig. 10. The Filter Overlay in our system. User can use this view to focus on different feature groups (e.g., Communication) or individual hurriance path

4.4.4 View Coordination on Mobile Device

We took very careful consideration when designing view coordination on the mobile interface. As far as we know, no prior research has been done that studies multiple-coordinated views (MCV) for mobile. Due to the limited screen-space, our view coordination has to occur within the same display area, rather than in a spread out fashion as mentioned in the original MCV guideline [26].

We ultimately relied on the transparency of each view to make the view coordination apparent, as shown in the video for this paper. Since there is limited screen space on mobile devices the use of overlays and transparency is necessary in order to maximize every inch of the screen. This allows the user to intuitively make selections and understand the effects of their changes.

Given our focus on disaster response domain, we choose the geospatial view as the main entry view where all ensembles are presented at the initial stage. The coordinated views for temporal analysis and filtering are split into two transparent overlays. Specifically, the user is able to select either a filter view or a temporal view that contains a transparent overlay in order to see any changes made to the spatial representation. The user can also very quickly hide the overlays with a simple swipe gesture in order to return to spatial navigation.

5 Case Study and Expert Feedback

Thorough evaluation for our mobile ensemble visual analytics system is challenging. Such evaluation requires a group of responders from multiple domains working on a large collaborative disaster response exercise. Due to the limited availability of the domain experts, we have demonstrated our system's utility through simulated scenarios, which are then analyzed by targeted users. In this process, we conducted two evaluation sessions with 3 City Emergency Responders and 4 Power Grid managers from a large regional power company, respectively. For each evaluation session, we sought users' understanding about the ensemble analysis environments that our architecture enables and solicit their expert feedback and comments. The use cases and results are discussed in the following.

5.1 Forecasting Events prior to Hurricane Season

Scenario: The center focus of this scenario was Wilmington, NC, a coastal city that is historically vulnerable to hurricanes. Specifically, over 1 million cascading infrastructure events were examined over 8 hurricane paths that had directly passed through the greater NC region. As shown in Figure 9, our mobile ensemble visualization provided users an exploratory environment to identify where is the most vulnerability facility.

In this case, an emergency responder was interested in isolating critical infrastructure that would likely be financially distressful if it were destroyed or rendered inoperable from a hurricane. The city planners need to specifically find these critical points like power substations and water services that could be very costly to the community if they are knocked out. He quickly examined the region using the probing function by drawing a bounding area around the different areas of the city. This action brings up a secondary window with the time distribution of events in that region, as shown in Figure 9. With multiple probing windows opened within the different quadrants of the city, the planner was able to see the aggregate view of the outages over time, but more importantly he was able to compare the peak times of outages side-by-side. Within the same window the planner was able to view the complete break down of events, which contains outages of airports, ports, highway bridges, and train stations. It is to the planners surprise that the airport in this region shows a significant outage at the early stages of our simulations, suggesting that a high chance for such facility to be disabled given the simulated hurricane paths.

He hypothesized this may be caused by road blockage and direct impact and navigated through the interface and selected the detailed breakdown for the airport. The emergency responder was surprised to see the report on what critical events caused the outage, as shown in Figure 7 (B); although it complies with his experience that likely outage may be caused by direct impact (44%), there still was an unexpected 51.2% outage caused by Electric Service failure. This provided him more insights and helped him to take action on investigating weak links in power connections to the airport by working with power company participants.

Expert Feedback: One of our architecture's advantage is the ability to compute massive amount of physical based infrastructure simulations ahead of a hurricane season. All collaborators consider this is of great value as they can start their planning process a couple of months ahead and be fairly comfortable with the probability level for the resiliency of their infrastructures. As summarized by one of the emergency responders, "this system gives us the ability to look forward and be prepared...It could help us to find a vulnerable building and plan ahead the scenarios when a hurricane actually hits".

However, since our analysis is build upon simulations, managers from the energy company suggested further extensions to fusion of heterogeneous datasets into the mobile ensemble visualization. In particular, they are interested in learning how we could effectively associate public information, such as census and demographics data. This will certainly be on our agenda moving forward along this research direction.

5.2 Investigating Cascading Effort of Infrastructure Breakdowns

Scenario: Our mobile visual analytics environment enables users to explore and follow the impact caused by the evolvement of a natural disaster. A city planner using the system inspected a couple possible hurricane paths that made a pass through the Wilmington area (Figure 5.

Using the direct filtering methods, the user can quickly narrow down the analysis scope to the paths within close proximity to the city and examine critical areas along it. As a transportation expert, her area of focus is on how to bring back the transportation infrastructure after an catastrophic impact. Hence she further filtered down with transportation infrastructure and noticed an interesting spreading pattern that presents an elongated trailing effect as it's still impacting the infrastructure after a month after the initial landfall. She was surprised to see this trailing effect and suspected this may due to the a less robust road structure. In deed, a quick examination of the transportation network near the coast revealed that this may be one of the possible cascading effects due to the less redundant pathways from the outer

coastal region to the inland. Once the hurricane passes through, getting relief and assistance back into the city requires outlets for transportation. Our system allows the user to then apply a specific filter just on transportation to see just those particular events in conjunction with the hurricane paths selected already. In this scenario, the mobile interface permits the user to focus on a specific infrastructure breakdown of her particular analysis interests.

Expert Feedback: One of the benefits that all these experts see in our architecture is its capability in helping to depict the overall impact on critical infrastructure as well as directly assess individual commodities (e.g., transportation structure). Using this visual interface, they can select specific hurricane paths that would pass through a certain city. Especially to planners, who are responsible for correlated information from various simulation models, the capability to identify and filter the infrastructures from a massive simulations space is of great value. As commented by a power grid manager that "...this is very useful for me to follow and compare my power lines with possible hurricane paths. I can then re-route my power transmissions away from the danger zone."

Many collaborators mentioned the need to conduct search-by-example functions due to the limited time constrains when a disaster unfolds. As stated by an emergency respond manager that, "it would be great if we can select one [impact] pattern...and the system can automatically suggest other similar ones...this would save a lot of time for me to navigate through the whole dataset." This requires our architecture to be able to perform comprehensive ensemble structuring, producing a more similar space for all the ensemble structure. His comment is well received, and we are working extensively on researching a quantifiable ensemble structure for ascertaining where attention is needed and resources should be deployed.

6 Limitations and Future Work

There are limitations to this research that must be addressed. Although our research was conducted towards the goal of ensemble analysis of simulation results, the complexity, the scalability and the mobility needs of the targeted dataset exacerbated the challenges on a computational-level and visualization-level. Especially, we found it very challenging when designing the multiple-coordinate views on a mobile device. While such research activity is upcoming, there is no gold standard to follow when considering the small screen space in mobile devices. We did attempt to mitigate the challenge a) by following mobile emergency design practices we have accumulated experiences [13] and b) by projecting ensemble abstractions to space-temporal visualizations that can best utilize the mobile displays. Nevertheless, multiple coordinate visualization in different domains still engender different design considerations. We acknowledge this challenge and are working closely with domain users to provide a more customized visual interface.

Moreover, our scenarios and expert feedback are limited to the available resources within the project. A more in-depth assessment of our system utility can only be derived when it's deployed in the field with the responders. Further, in our study the effects of time pressure were not considered, which may have significant influence on decision makers (Section 5.2). We will look into incorporating techniques, such as search-by-example, to reduce users the time pressures.

In future work we would like to run simulations in the cloud that are driven from a mobile device. There are a number of limitations that prevent us from generating new simulations in real time. The first limitation is a computational problem. Running these exhaustive simulations is very time consuming and would require new more efficient methods to run in real time. Secondly, we would have to develop effective methods for generating a simulation model on a mobile device then communicate that model to our simulation environment in order to evaluate the simulation. Currently the simulation environment is a closed system that is self contained. Additional modifications and designs would have to be made to allow for two way communication with the simulation environment.

We recognize these limitations and consider the support for disaster forecast and preparation as an important visualization, analytics, and interaction research topic. We hope the presented approach illuminates the role that mobile ensemble interfaces play in such complex problem-solving environments.

7 Conclusion

Understanding large cascading simulation datasets is vital in disaster monitoring and emergency response. But, gaining insight from these simulations requires extensive tools and analysis in order to provide planners and emergency responders with valuable information. To gain enough insights to actionable knowledge, we utilize ensemble visualization of a large scale simulation space. We have developed a general critical infrastructure simulation and analysis system for situationally aware emergency response during natural disasters. Our system demonstrates a scalable visual analytics infrastructure with mobile interface for analysis, visualization and interaction with large-scale simulation results in order to better understand their inherent structure and forecast capabilities. The utility and efficacy of our research has been evaluated by domain practitioners and disaster response managers.

References

[1] L. Anselin, I. Syabri, and O. Smirov. Visualizing multivariate spatial correlation with dynamically linked windows. In *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, 2002.

[2] Apple, Inc. ios human interface guidelines. Sep 2013.

[3] A. Buja, D. Cook, and D. F. Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5:78–99, 1996.

[4] R. Bürger and H. Hauser. Visualization of multi-variate scientific data. In *Eurographics 2007 STAR*, pages 117–134, 2007.

[5] T. Butkiewicz, R. Chang, Z. Wartell, and W. Ribarsky. Visual analysis and semantic exploration of urban lidar change detection. *Comput. Graph. Forum*, 27:903–910, 2008.

[6] J. Dietrich, C. Trahan, M. Howard, J. Fleming, R. Weaver, S. Tanaka, L. Yu, R. L. Jr., C.N, J. Dawson, Westerink, G. Wells, A. Lu, K. Vega, A. Kubach, K. Dresback, R. Kolar, C. Kaiser, and R. Twilley. Surface trajectories of oil transport along the northern coastline of the gulf of mexico. In *Continental Shelf Research*, 2012.

[7] T. Eaglin, K. Subramanian, and J. Payton. 3d modeling by the masses: A mobile app for modeling buildings. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 315–317, 2013.

[8] H. Edelsbrunner and E. P. Mücke. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics*, 9:66–104, 1990.

[9] Environmental Systems Research Institute. Arcgis desktop. March 2014.

[10] Federal Emergency Management Agency. http://www.fema.gov/hazus.

[11] H. Feild and K. Emery. An uncertainty analysis of the spectral correction factor. In *Photovoltaic Specialists Conference, 1993., Conference Record of the Twenty Third IEEE*, pages 1180–1187, Louisville, KY, USA, May 10–14 1993.

[12] Google Inc. http://map.google.com.

[13] J. Guest, T. Eaglin, K. Subramanian, and W. Ribarsky. Visual analysis of situationally aware building evacuations, 2013.

[14] G. Horne. Beyond point estimates: Operational synthesis and data farming. *Maneuver Warfare Science*, 2001.

[15] B. Hubbard, J. Kellum, B. Pual, D. Santek, and A. Battaiola. Vis5d. http://vis5d.sourceforge.net.

[16] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, Mar. 2010.

[17] A. Mascarenhas, R. W. Grout, P.-T. Bremer, E. R. Hawkes, V. Pascucci, and J. Chen. *Topological feature extraction for comparison of terascale combustion simulation data*. Mathematics and Visualization. Springer, 2010. to appear.

[18] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27:124007, 2012.

[19] S. Mittelstdt, D. Spretke, D. Thom, D. Jckle, A. Karsten, and D. A. Keim. Situational Awareness for Critical Infrastructures and Decision Support. In *Proceedings NATO STO IST-116 Symposium on Visual Analytics*, 2013.

[20] D. Ni, Y. Chui, Y. Qu, X. Yang, J. Qin, T. Wong, S. Ho, and P. Heng. Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT. *Computerized Medical Imaging and Graphics*, 33:559–566, 2009.

[21] T. Nocke, M. Fleshig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *IEEE 2007 Water Simulation Conference*, pages 703–711, 2007.

[22] V. Pascucci, D. E. Laney, R. Frank, G. Scorzelli, L. Linsen, B. Hamann, and F. Gygi. Real-time monitoring of large scientific simulations. In *Proceedings of the 18-th annual ACM Symposium on Applied Computing*, pages 194–198, Melbourne, Florida, March 2003.

[23] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. *International Conference on Data Mining*, 0:233–240, 2009.

[24] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, V. Pascucci, and C. Johnson. A flexible approach for the statistical visualization of ensemble data. In *Proceedings of IEEE Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes, and Impacts*, December 2009.

[25] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

[26] J. C. Roberts. State of the art: Coordinated and multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.

[27] G. Robertson, K. Cameron, M. Czerwinski, and D. Robbins. Animated visualization of multiple intersecting hierarchies. *Information Visualization*, 1(1):50–65, Mar. 2002.

[28] T. Safford, J. Thompson, and P. Scholz. Storm surge tools and information: A user needs assessment. NOAA Coastal Services Center.

[29] R. Samtaney, D. Silver, N. Zabusky, and J. Cao. Visualizing features and tracking their evolution. *Computer*, 27(7):20–27, July 1994.

[30] E. Santos, J. Freire, C. Silva, A. Khan, J. Tierny, B. Grimm, L. Lins, V. Pascucci, S. A. Klasky", R. D. Barreto, and N. Podhorszki. Enabling advanced visualization tools in a simulation monitoring system. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 358–365. IEEE, December 2009.

[31] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16:1421–1430, 2010.

[32] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12):2431–2440, 2012.

[33] The Federal Communications Commission. http://www.fcc.gov/data/download-fcc-datasets.

[34] Tile Mill. Tile mill https://www.mapbox.com/tilemill/.

[35] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Manuscript*, 2012.

[36] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991.

[37] X. Wang, W. Dou, T. Butkiewicz, E. Bier, and W. Ribarsky. A two-stage framework for designing visual analytics system in organizational environments. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 251 –260, oct. 2011.

[38] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies

Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky

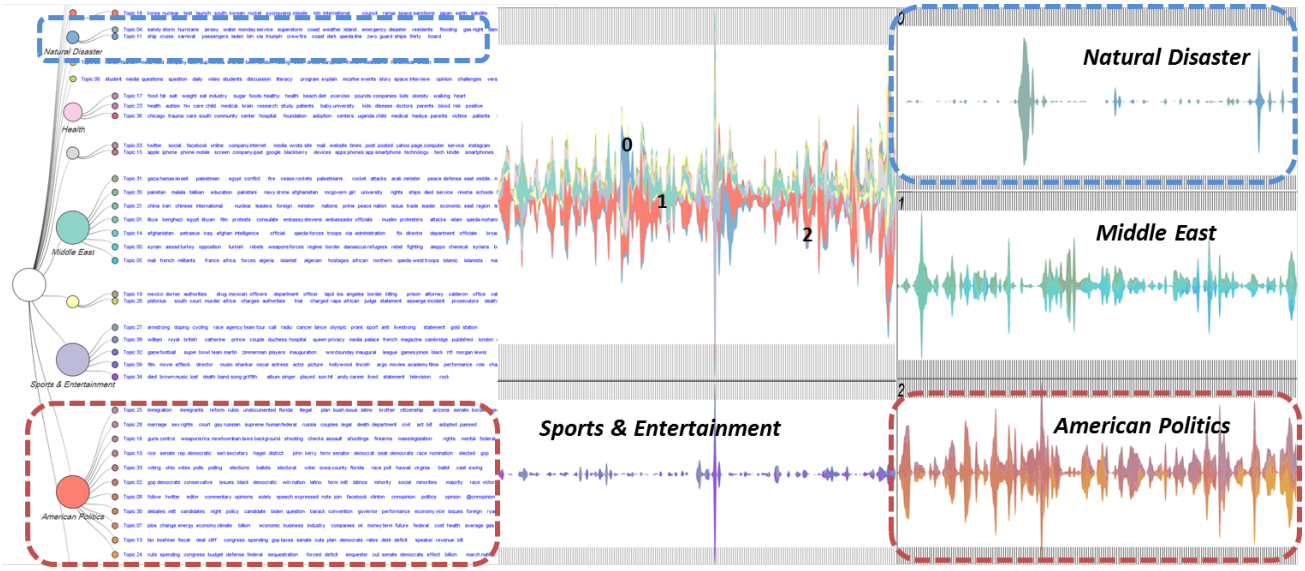


Fig. 1. Overview of the HierarchicalTopics system. The Hierarchical Topic structure is shown on the left in a tree visualization. The Hierarchical ThemeRiver view on the right presents the temporal pattern of topics in a hierarchical fashion. The dataset being visualized is the CNN news corpus. Topics are organized into 5 categories and annotations are attached to describe each news category. The corresponding categories in both view are outlined with same colors.

Abstract—Analyzing large textual collections has become increasingly challenging given the size of the data available and the rate that more data is being generated. Topic-based text summarization methods coupled with interactive visualizations have presented promising approaches to address the challenge of analyzing large text corpora. As the text corpora and vocabulary grow larger, more topics need to be generated in order to capture the meaningful latent themes and nuances in the corpora. However, it is difficult for most of current topic-based visualizations to represent large number of topics without being cluttered or illegible. To facilitate the representation and navigation of a large number of topics, we propose a visual analytics system - HierarchicalTopic (HT). HT integrates a computational algorithm, Topic Rose Tree, with an interactive visual interface. The Topic Rose Tree constructs a topic hierarchy based on a list of topics. The interactive visual interface is designed to present the topic content as well as temporal evolution of topics in a hierarchical fashion. User interactions are provided for users to make changes to the topic hierarchy based on their mental model of the topic space. To qualitatively evaluate HT, we present a case study that showcases how HierarchicalTopics aid expert users in making sense of a large number of topics and discovering interesting patterns of topic groups. We have also conducted a user study to quantitatively evaluate the effect of hierarchical topic structure. The study results reveal that the HT leads to faster identification of large number of relevant topics. We have also solicited user feedback during the experiments and incorporated some suggestions into the current version of HierarchicalTopics.

Index Terms—Hierarchical Topic Representation, Topic Modeling, Visual Analytics, Rose Tree

1 Introduction

Digital textural content is being generated at a daunting scale, much larger than we can ever comprehend. Vast amounts of content is accumulated from various sources, diverse populations, and different times and locations. For example, 1.35 million scholarly articles were published in 2006 alone [18]. With an average annual growth rate of 2.5% [30], research articles are currently being published at the pace of approximately 4400 titles per day. In the social media world, people are contributing to the accumulation at an even faster pace. By June 2012, Twitter is seeing 400 million tweets per day [31]. Meanwhile, 900 million active Facebook users have been busy sending 1 million messages every 20 minutes [28]. Today, part of the content (e.g, tens of thousands of different sites, Twitter, digitized books) is archived in the US Library of Congress with more than 300 terabytes in size, which keeps on growing [11].

It is generally agreed in government and industry that valuable but latent information is hidden in the vast amount of digital textual con-

• *Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky are with University of North Carolina at Charlotte. E-mail: wdou1, lyu8, xwang25, zma5, ribarsky@uncc.edu.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

tent. For instance, in scientific research, one of the crucial investigations is on the development of science. To this aim, researchers have created maps of science [25, 27] and evaluated the impact of science funding programs [14] by analyzing research publications and proposals. For emergency response agencies, sifting through massive amount of social media data could help them monitor and track the development of and response to natural disasters, as illustrated in the use of Twitter to reach victims from Hurricanes [35]. Last but not least, the emergence of numerous social media startups shows that profitable marketing and business analytics insights that can be extracted from such content. To extract insights and make sense of large amounts of textual data, efficient text summarization is therefore much needed.

In this regard, topic models have been considered as the state-of-the-art statistical methods to extract meaningful topics/themes for summaization. Although powerful, topic models do not provide meanings and interpretation; human must be involved [7]. To enhance the interpretations of topical results, visual text analytics researchers have designed algorithms and visual representations that make the probabilistic topic results legible and exploratory to a broader audience [8, 9, 10, 14, 15, 26, 34]. Examples of the utility of these topic-based visualization interfaces include the analysis of social media users based on the content they generated [22], depiction of the temporal evolution of topics [14, 26], and identification of interesting events from news and social media streams [8, 15]. Many of these topic-based visualization systems have been studied through use cases and regarded powerful in aiding text analysis processes.

However, current visual text analytics systems have limitations. In contrast to the common practice of extracting hundreds of topics from large document corpora in the topic model community [2, 4, 21, 29, 32], current systems usually only manage to effectively represent a small number of topics. As more textual data becoming available, the number of necessary topics for interpretable text summarization will grow inevitably. Only extracting a small number of topics, therefore, won't capture the nuances in the corpora. As the number of topics increase, sifting through and comprehending all the topics becomes a time-consuming and laborious task, which will be further hampered by the visual clutter introduced when displaying the temporal evolution of hundreds of topics with no organization.

In particular, three challenges must be met to effectively analyze document collections that are summarized by large number of topics:

1. **How to organize the topics to facilitate the navigation and analysis within the topic space?** Without organization, sifting through a hundred topics with each topic consisting of 20 or more keywords could be intimidating. One example that highlights the problem is that when developing the NSF Portfolio Explorer, it took days for a researcher to manually examine a thousand topics to select 30 topics for further analysis and visualization [12]. Since certain topics are closer in meaning than others, organizing semantically similar topics into topic groups will ease the navigation in the topic space. Having an automated classification of topics could potentially jumpstart the analysis of text collections based on large number of topics, however, the automated classification may not always conform to individual users' mental model of the topics space.

2. **How to visually convey and permit user interactions with the organized topic results so that users can classify the topics based on their interests?** It is essential to place users in the center of the topic analysis process, allowing users to leverage and modify the topic classification results. For example, when analyzing a news corpus, a user may want to organize the topics into a hierarchical structure through first categorizing the news topics into either domestic or foreign news. In addition, for domestic news topics, the user may want to further divide the topics into groups such as politics, sports, entertainment, etc. Similarly, when analyzing topics from Twitter streams, a business analyst may be interested in grouping all topics related to sales and customer services and further divide them into more refined categories. Therefore, intuitive topic visualizations and user interactions are needed to support the analysis and modification from an initial topic organization provided by an automated algorithm.

3. **How to modify existing visual metaphors to accommodate the organization of a large number of topics?** After a user has identified a desirable hierarchical topic structure, the third challenge lies in tailoring existing visual representations. Visualizing temporal evolution of topics has been considered essential to understanding various domains (e.g. scientific fields, breaking news, etc.) over time. However, ThemeRiver [16] and stack graph that are commonly used to present the temporal trends of the topics do not convey hierarchical information. To enable the analysis and comparison of temporal behavior of topic and topic groups, it is essential to extend the current visual metaphors to incorporate hierarchical structure of topics.

To tackle the three challenges, we propose HierarchicalTopics (HT), a visual analytics system [1] that supports scalable exploration and analysis of document corpora based on a large number of topics. HierarchicalTopics *addresses the first challenge* by integrating a novel algorithm that automatically classifies topics into a hierarchical structure. Through joining similar topics into the same group, the new organization of topics provides scalable representation and navigation in the topic space. HierarchicalTopics further incorporates visual representations and interactions that embrace the hierarchical organization of the topics, and enables the users to depict the temporal evolution of topics or topic groups. In addition, user interactions are provided in HT to *address the second challenge*. Along with the visual representations of the topic hierarchy, HT allows users to modify and update the automatically computed topic groups. It therefore supports the customization of the visualizations based on the users' analytical interests. *To address the third challenge*, a new Hierarchical ThemeRiver has been designed to accommodate the hierarchical organization of the topics. The Hierarchical ThemeRiver eases the exploration of temporal behaviors of topic groups, and enables the comparison of topic groups on a temporal dimension. Through tight coordination between the visualizations of topic hierarchy and hierarchical temporal trends, we intend to provide an inviting interface that supports making sense of large document collections via navigating through large number of topics and their temporal evolution.

We have assessed the HT through both qualitative and quantitative evaluations. To evaluate the system in a qualitative manner, we present a case study in which an expert user performed in depth analysis on a collection of 11,961 NSF awarded proposal abstracts. To evaluate HT in a quantitative fashion, an 18-participant user experiment is conducted to compare the HierarchicalTopics system to a non-hierarchical representation based on a CNN news corpus that contains 2453 recent news articles. The experiment results reveal that the hierarchical topic visualization leads to faster identification of a large number of relevant topics. Constructive user comments were also collected during the experiment. After the user study, some suggestions on improving the visualization and interactions from the participants have been incorporated into the current version of the HierarchicalTopic system.

The rest of the paper is structured as follows: we introduce the previous work that inspired the design of HierarchicalTopics in Section 2. Section 3 focuses on introducing the HierarchicalTopics, including its system architecture and interactions. We present a case study in Section 4, followed by descriptions of a user study in Section 5.

2 Related Work

Two lines of work inspire the design of HierarchicalTopics, namely topic models and topic-based visualizations.

2.1 Topic Models

Topic models can be effective tools for text summarization and statistical analysis of document collections [2]. The number of topics needed is typically determined by the size of the text corpora. The

[1] A video of the HierarchicalTopics can be found at `http://youtu.be/Vi1FP5kAbOU`.

larger the size the more topics are preferred to ensure topic comprehension and human interpretability, typically tens of thousands of articles will require topics in the scale of hundreds. Specifically, one school of topic models is based on a human-defined number of topics. Researchers and practitioners usually generate a large number of topics to capture the themes that pervade the text collection as well as the nuances. For instance, in the experiment of evaluating the collaborative topic model [32], the authors extracted 200 topics from a paper-abstract collection with 16,980 articles and a vocabulary size of 8000. In other non-parametric Bayesian topic models, such as the hierarchical Dirichlet process (HDP) [29] and the discrete infinite logistic normal distribution (DILN) [21], the number of topics is determined by the model. However, it is evidenced that such algorithmically generated number of topics is typical rather large. For example, in the experiment evaluating DILN, the model produced 50 to 100 topics given a fairly small dataset with only 3000 to 5000 news articles.

Such large number of topics creates challenges to human interpretations and the sense-making process. Much research has been focused on revealing the correlations between latent topics and organizing topics into more human interpretable structures. Work in this area aims to facilitate the navigation through the topic space and enables the discovery of documents exhibiting similar topics. While most of the existing topic models do not explicitly model correlations between topics, a few exceptions have directly accounted for relationships between latent topic themes. For example, both correlated topic model (CTM) [4] and DILN [21] have demonstrated better predictive performance and have uncovered interesting descriptive statistics for facilitating browsing and search. Although the topic correlations have been modeled, it is still difficult for users to take advantage of the descriptive statistical relationship of topics without an effective organization and visual representation of the topics.

Many researchers consider that organizing topics into a hierarchical structure presents a scalable solution to improve human-interpretability of topic. To this aim, Blei et al. have proposed a hierarchical topic model (hLDA) that learns topic hierarchies from data to accommodate a large number of topics [3]. The hLDA is a flexible, general model for extracting topic hierarchies that naturally accommodates growing data collections. However, the topic hierarchies hLDA produced are rather rigid since the depth of such hierarchies is predefined and fixed throughout the modeling process. In addition, the higher level topics generated by hLDA usually consist of stopwords, therefore less meaningful for human users.

In order to leverage the scalable hierarchical structure without enforcing rigid restrictions on the topic models, we developed an algorithm, Topic Rose Tree, to construct a multilevel hierarchical structure with any given number of generated topics. Together with interactive visualizations, our HierarchicalTopics system enables users to explore and iteratively update the topic hierarchy. Our system aims to improve human-interpretability by enabling users to tailor the hierarchical topic results to their analytical interests or mental models of the topic space.

2.2 Visualization based on Topic Models

The power of topic models in summarizing and organizing large text corpora has been widely recognized in the visualization community. A good number of visualization systems have been developed based on topic models for users to comprehend document collections.

As one of the pioneer visual text analysis systems, TIARA [34] combined topic models and interactive visualization to help users explore and analyze large collections of text. Specifically, TIARA utilized a stack graph metaphor to represent temporal change of topics over time. Similarly, another system ParallelTopics was also developed to depict both temporal changes of topics using ThemeRiver and the characteristics of documents based on their topic proportions via Parallel Coordinates [14]. Since temporal evolution of the topics has been considered one of the most useful features of the topic-based visualizations, researchers have extended a great deal in this direction. TextFlow [13] presented a novel way to visualize topic birth, death, and merge that signify critical events. In a similar vein of identifying events, LeadLine [14] applied event detection methods to detect "bursts" from topic streams and further associate such bursts with people and locations to construct meaningful events. Furthermore, Chae et al. proposed a visual analytics approach that supports the analysis of abnormal events detected from topic time series [8]. Instead of representing and analyzing topics along the temporal dimension, Lee et al. proposed a visual analytics system for document clustering based on topic modeling [19]. Users could guide the clustering process through adjusting term weights in the topics.

These topic-based systems have demonstrate the effectiveness of combining topic models with interactive visualizations in facilitating analysis of text corpora. As indicated in in most of their reported case studies, however, these systems only dealt with a fairly small number of topics. This is quite contrary to the common practice in the topic modeling community, where a lot more topics are generated for a text collection of similar size (Section 2.1).While a greater number of topics will inevitably introduce visual clutter and legibility issue to the visualization systems, limiting the topic number may also hamper users' ability in comprehending the text collection.

Therefore, more scalable approaches to organizing the topics and visual representations based on the topics are much needed to support real-world challenges of analyzing large text corpora. To meet this need, HierarchicalTopics provides a scalable solution that allows iterative analysis of document collections with a large number of topics and further supports the exploration of temporal evolution of those topics in a hierarchical fashion.

3 HierarchicalTopics

3.1 System Pipeline

As illustrated in the overall system architecture in Figure 2, HierarchicalTopics is a user-centered analysis system that integrates computational methods with interactive visualizations. HT systematically incorporates both online and offline computations and utilizes scalable infrastructures described in [33], including MapReduce and Parallel Processing. There are four key processing stages in the HT architecture including two offline computation modules (e.g., Data collection, and preprocessing and Parallel Topic Modeling) and two online components (e.g., Topic Rose Tree and Hierarchical Visualizations).

In particular, HT accommodates digital text content from various sources including social media, research publications, news, etc. Once the data is collected, it is streamlined into HT's data cleaning and preprocessing step, as shown in Figure 2A. In this process, HT first unifies the formats of input data and converts certain documents (PDFs) to proper topic-model-readable text files. It then prepares the documents for parallel topic models by removing stopwords and emojis.

The cleansed data then goes through the topic modeling stage (Figure 2B), which extract topics from the document collection. It is worth noting that the choice of the topic model component in HT is rather flexible. The architecture of HT is set to utilize a variety of topic models and can leverage their unique strengths such as interpretability [7], convenience of non-parametric models [21, 29], and accounting for additional metadata [23, 24], etc. As reported in paper, HT has successfully incorporated both the vanilla LDA [5] and the Author Topic Model (ATM) [24] to handle the natures of different text corpora.

After the first two stages are accomplished offline, the rest of the computation and visualization are computed online. The Topic Rose Tree (TRT) shown in Figure 2C organizes the probabilistic topic results into a hierarchical structure, as detailed in next section. Based on the hierarchical topic organization, two coordinated interactive visualizations (Figure 2D) are designed to present and support interactive analysis of topics and temporal evolution of the topics.

The TRT and the visualizations are closely coupled through the user interactions provided by the HierarchicalTopics system. In particular, the three essential operations in the TRT algorithm (e.g., join, absorb, and collapse) are directly incorporated in the visualizations and interactions. Through direct visual manipulations, HT allows the users to perform the same operations to modify the initial topic hierarchies and iteratively derive the most interpretable topics groups based on their analytic interest.

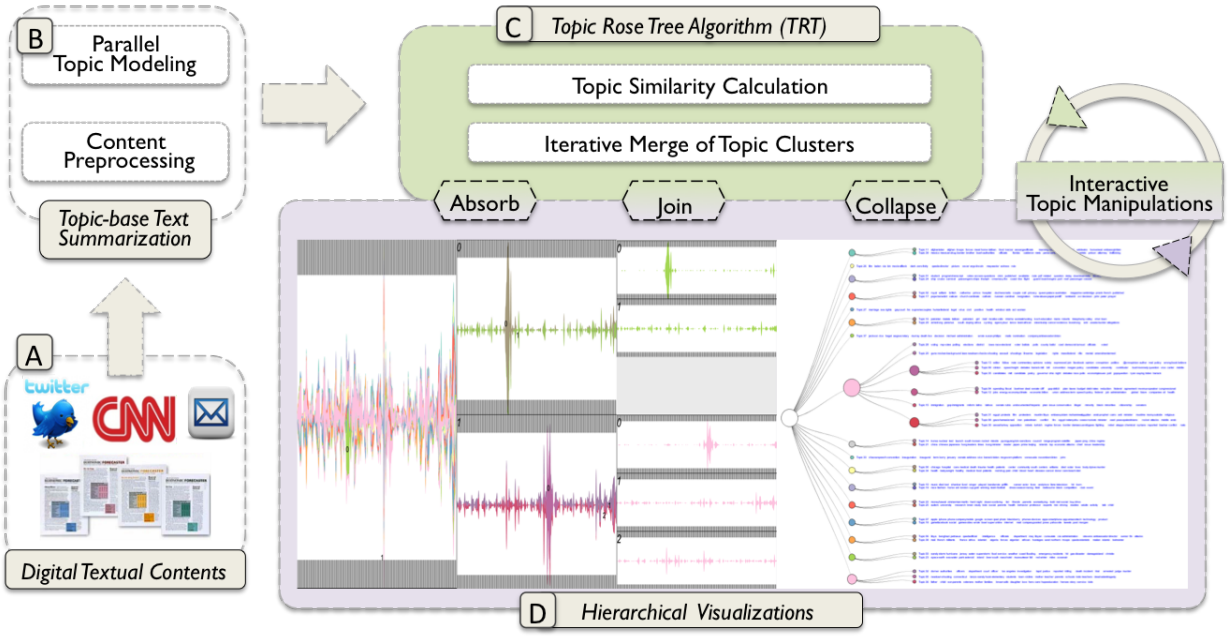


Fig. 2. System Architecture of HierarchicalTopics. Starting from bottom left, textual data is first harvested (A). The data then goes through a preprocessing stage before entering the topic model component (B). These two steps are completed offline. The resulting statistics from topic models then serve as input to the Topic Rose Tree (C), which constructs a hierarchy given a list of topics. The topic hierarchy is then visualized in the interactive visual interface (D) for users to analyze the topics and temporal trends in a hierarchical fashion to derive understanding of the text collection.

In the rest of this section, we will focus on presenting details of the online components of HierarchicalTopics.

3.2 Topic Rose Tree

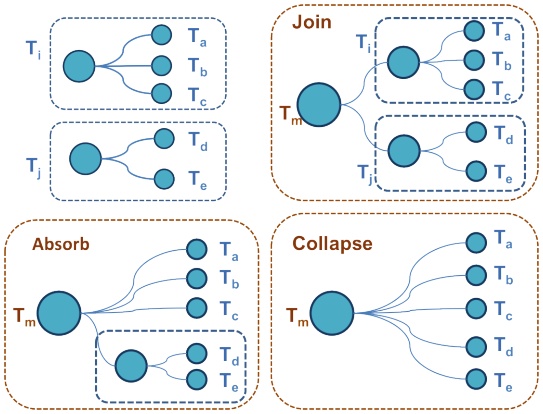


Fig. 3. The three essential operations of our Topic Rose Tree algorithm.

Our goal in designing the Topic Rose Tree is to support scalable visual representation and exploration. TRT is an automated method that can meaningfully organize a list of topics into a hierarchical structure. Its core algorithm is built upon key concepts from the Bayesian Rose Tree (BRT), which constructs a hierarchy using hierarchical clustering methods [6]. Compared to previous hierarchical clustering methods that limit discoverable hierarchies only to those with binary branching structures, BRT produces trees with arbitrary branching structure at each node, known as rose trees [6]. We consider such characteristic more natural in organizing topics, since any number of topics could be similar and should be grouped into one partition in a hierarchical structure. The essence of generating a rose tree is support of the three operations, namely join, absorb, and collapse (shown in Figure 3).

Unfortunately, simply borrowing BRT and directly applying it to topic models is unfit based on our experiments. This is primarily caused by the large number of features (words in the vocabulary) from topic models. In addition to the vocabulary size of a text corpus, which is usually in the thousands, the binarized matrix of topic distributions over the vocabulary is extremely sparse, causing problems for calculating the marginal probability of the topic groups in a tree.

Therefore, we developed TRT, an algorithm that built upon the three operations to construct hierarchies specifically from topic modeling results. TRT is a one-pass, bottom up method which initializes each topic in its own cluster and iteratively merges pairs of clusters. To construct the hierarchical structure, we first compute the similarity between any pair of clusters (topics/topic groups). TRT then merges the most similar clusters using one of the three operations. In this process, the Hellinger distance, which is a symmetric measure of the similarity between two probability distributions, is used to calculate the similarity of a pair of clusters. Intuitively, topics or topic groups that share similar distributions over the vocabulary yield lower distance. To construct the hierarchy, the most similar topic (group) clusters will be merged at each step.

In particular, each topic from the topic modeling results is represented as a probabilistic distribution over the entire vocabulary given a text collection, denoted by $X_{i,v}$, with i representing the ith topic and v representing the vocabulary of size N. To represent the probabilistic distribution of a node that contains multiple topics (children), we simply compute an average of all distributions of the children's. Details of the TRT are shown in Algorithm 1.

The complexity of the topic rose tree is the same as the BRT algorithm. First, the distance for every pair of data items needs to be computed-there are $O(n^2)$ such pairs. Second, these pairs must be sorted in order to find the smallest distance requiring $O(n^2 \log n)$ computational complexity.

To showcase how the topic rose tree algorithm could group similar topics together, Figure 4 shows a partial result from the initial grouping. In this case, we used the 2011 VAST mini challenge 1 microblog data, which contains an embedded scenario of an epidemic spread. This data is good for qualitatively evaluating the algorithm since we

Algorithm 1 Topic Rose Tree

Input: Data $\mathbf{D} = \{\mathbf{X_i}, \boldsymbol{v}\}, \boldsymbol{i} = 1, 2, ...n$; $\boldsymbol{v}$ is the vocabulary of the corpus
Output Topic rose tree T_{n+1}, a hierarchical structure with all topics
Initialize: $\mathbf{T_i} = \{\mathbf{X_i}, \boldsymbol{v}\}, \boldsymbol{i} = 1, 2, ...n$
Steps:

Denote c as cluster count
while c > 1 **do**
  for each pair of trees Ti and Tj **do**
  Calculate cost D(i,j) for 3 operations (join, absorb, or collapse): $D(i,j) = 1/2 * \sum_{v=1}^{N} (\sqrt{t_{i,v}} - \sqrt{t_{j,v}})^2$, $t_{i,v}$ denotes the probability distribution of tree node T_i over the vocabulary of size N
  Find operation m which yields lowest cost for T_i and T_j
  Merge T_i and T_j into T_m using operation m
  Delete T_i and T_j, $c = c - 1$
  end for
end while

expect similar topics regarding the epidemic spread should be grouped together. The topic group shown in Figure 4 (top) contains three topics highlighting the flu-like symptoms for the first two days of the epidemic (each tick on the x axis denotes a day). Another topic shown in Figure 4 (bottom) highlights evolved symptoms such as pneumonia for the third day of the epidemic. Note that since the words that were tweeted to describe the symptoms have changed a great deal, the topic rose tree did not put topic 31 into the first topic group. However, combining with the temporal patterns, one can identify when the epidemic spread started, and how the symptoms evolved over time. This example illustrates that the topic rose tree is able to group similar topics together, and the result is very much interpretable by human users.

Fig. 4. An example showcases the capability of TRT grouping to group topics together. The top three topics (grouped by TRT) describe all flu-related symptoms on the first two days of the disease outbreak. The bottom topic (in grey) was not grouped into the first group by TRT since it describes different symptoms on the third day. Tweets related to certain topics are shown in a detailed view upon selection.

3.3 Visual Components

After applying the Topic Rose Tree to the topic modeling results, a hierarchical topic organization is generated. To facilitate the topic analysis of the text collection, we present a visual interface that is tailored to the hierarchical organization of topics. The visual interface consists of two coordinated views, namely Hierarchical Topic View and Hierarchical ThemeRiver. The two views are coordinated through user interactions with a focus on correlating the hierarchical information.

3.3.1 Hierarchical Topic view: Depicting topics in a hierarchical fashion

While TRT computationally alleviates the topic organization issue, the Hierarchical Topic view is designed to visually address ***Challenge 1*** by presenting the topic contents in a hierarchical fashion. Such representation not only offers a scalable solution as it allows the number of topics to accrue, but also supports better navigation by grouping similar topics together. Figure 1 shows the Hierarchical Topic view with 40 topics extracted from the CNN news corpus. To provide user a familiar visual environment, we adopt straightforward tree visual representation. In this view, each leaf node represents a topic, while the non-leaf nodes denote topic groups. The first node on the left is the root of the topic hierarchy, with the rose tree spanning from left to right. The content of each topic (in the form of a group of keywords) is presented to the right of each leaf node. The size of the node is drawn proportionally to its number of children (shown in figure 1).

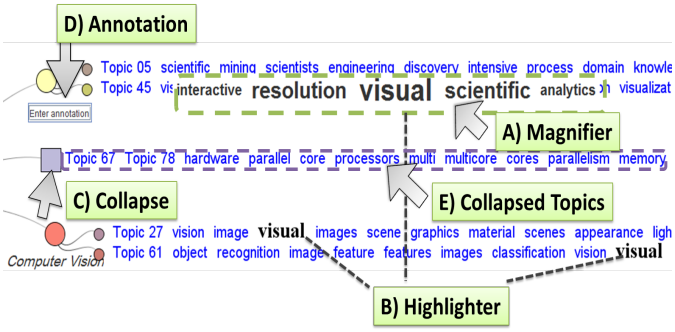


Fig. 5. Interactions provided by the Hierarchical Topic view. A) Magnifier: enlarges keywords near mouse cursor. B) Highlighter: highlight all occurrences of a selected keyword. C) Node collapsing: details of the collapsed children nodes are no longer shown. The shape of the node turns rectangular when collapsed. D) Annotation: allows users to enter annotation. E) Collapsed topic: keywords showing a summary of the two topics being collapsed.

User interactions. The Hierarchical Topic view provides a set of user interactions for effective exploration and navigation through large numbers of topics. In addition to standard panning and zooming, this view employs both an on-demand magnifier and highlighter to facilitate the examination of the topic contents, as shown in figure 5 A and B. The magnifier is designed to help users to better read the topic keywords through enlarging the font near the mouse cursor, while the highlighter aims to reveal the associations between topics by highlighting all occurrences of a certain keyword in the other topics. To further help users concentrate on the topics of interests, the Hierarchical Topic view supports interactive collapsing and expanding topic groups, shown in the square node in Figure 5C. Keywords for topics being collapsed into the same group are shown in (Figure 5E). More importantly, the Hierarchical Topic view allows users to annotate on the nodes to attach semantic meanings to topic groups (Figure 5D).

Interactive modification of the topic hierarchies. In addition to facilitating topic exploration, the Hierarchical Topic view aims to provide an intuitive way to visually classify the topics based on users' interest. In the process of analyzing a text corpus, only human users can attach semantics to the topics and provide meaningful yet sometimes subjective groupings. Therefore, it is essential to allow users to interactively modify the rose tree based on their analytical interests.

To permit such modification, the three operations that are used to construct the hierarchy in the topic rose tree algorithm are supported intuitively through drag-n-drop in the Hierarchical Topic view. As shown in Figure 6, dragging one leaf node into another constitutes the "join" operation. Drag-and-dropping any non-leaf node into another is considered as performing the "absorb" operation, while dragging multiple nodes into another node is interpreted as the "collapse" operation.

As observed in both the case study and user experiments (Section 4 and 5), the ability to iteratively refine and manipulate topic groups has demonstrated significant utility when analyzing text collections. Especially when HierarchicalTopics embodies the above three essential operations into intuitive mouse interactions, it creates a flexible text

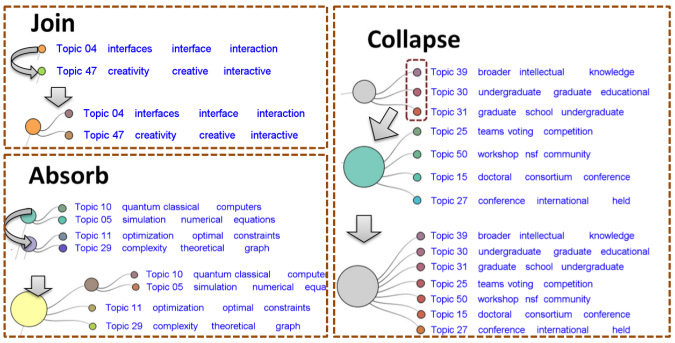


Fig. 6. Three operations supported to modify the topic hierarchy through user interactions.

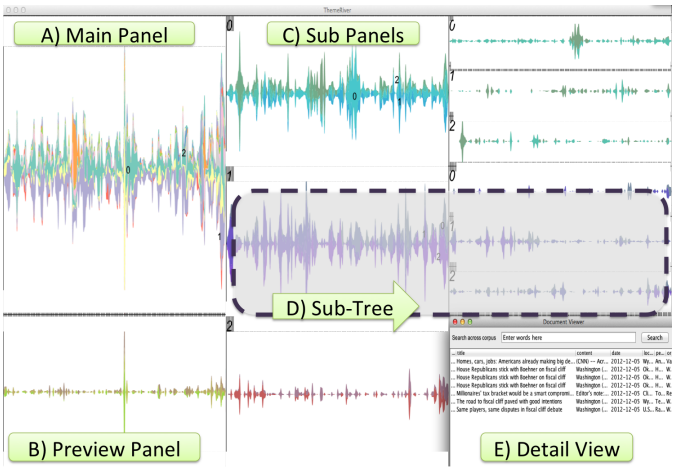


Fig. 7. Overview of the Hierarchical ThemeRiver. The dashed rectangle, in component D, highlights a sub tree created upon user interaction to view temporal patterns of child nodes.

analytics environment for users to categorize, modify, and update topics and topics groups. For example, as illustrated in Figure 1, participants in our user study have used these three operations to effectively group topics into five news categories based on the initial TRT hierarchy. In addition, the annotation interaction in HT view permits the users to attach semantic interpretations of the topic groups, and further helps them to connect the dots of a large number of topics. Many of our participants agreed that such user interactions served as a potential solution to the ***Challenge 2*** (see Section 1).

In summary, the Hierarchical Topic view provides both a visual representation of the topic hierarchy and a set of user interactions to serve as the first step to effectively analyze text collections.

3.3.2 Hierarchical ThemeRiver: Representing the temporal trends of topic groups

In addition to visually representing the topics which serve as a summarization of the document collection, visualizing the temporal evolution of the topics brings a unique contribution; it permits the discovery of the rise and fall of different topic themes, as well as identifying possible critical events [13, 15].

To this aim, we extend the widely adopted temporal visualization, ThemeRiver [16], to further incorporate hierarchical information. Our goal in designing the Hierarchical ThemeRiver is to provide users the ability to analyze and compare temporal behaviors of topic and topic groups, which address the core issue in ***Challenge 3*** (Section 1).

As illustrated in Figure 7, the Hierarchical ThemeRiver starts with the main panel (Figure 7A), where the temporal evolutions of the highest hierarchy (children of the root node) are shown; the height of each ribbon is calculated by summing the height of its leaf nodes. Once a ribbon is hovered, a preview of the temporal evolution of the child nodes will be shown in the preview panel (Figure 7B). The panels support interactively examination of the overall temporal trends of a text corpus as well as individual topic groups.

An elastic-panel structure is built into the view to enable the users' comparison of multiple topic groups. To compare different topic groups, a user can start by selecting a topic ribbon in the main panel; such interaction will create a sub panel (Figure 7C) showing the next level of hierarchy of the currently selected node. Multiple selections can be made to view the detailed temporal evolution of different topic groups, thus enabling the comparison and association of temporal patterns. Note that sub panels are always expanded to the right of the current selection, creating a coherent look and feel of the layout as in the Hierarchical Topic view.

Color assignment. To assist user exploration as well as to keep a smooth transition between panels, we have carefully chosen 12 perceptively coherent colors for the Hierarchical ThemeRiver view. This is done in an experimental fashion using the "i want hue" system [20], with the k-Means clustering and light background option. In the Hierarchical ThemeRiver view, the 12 distinct colors are first assigned to the topic ribbons in the main panel (Figure 7A). The child ribbons of each selected parent ribbon get colors of the same hue, but with varying luminance and chorma, as shown in Figure 7C. The same color scheme is also used in our Hierarchical Topic view to provide a coherent visual cue that helps correlating the two different representations of the same topic or topic groups.

Temporal selection and details on demand. To permit the examination of documents of interests, details of the text content are shown upon selection. In any panel within the Hierarchical ThemeRiver view, a user can enable the "time column" mode and interactively select a subset of documents published in a certain time period. By doing so, a detail view (Figure 7 E) will be shown to help the user validate the temporal patterns and understand its cause. During the user study, for example, this operation was demonstrated useful in examining the contributing posts to a topic burst pattern.

In summary, the Hierarchical ThemeRiver view is tailored to represent temporal patterns of topic and topics groups in a hierarchical manner. The incorporation of hierarchical information is mainly achieved through user interactions and in a way that is coherent to the Hierarchical Topic view representation.

3.3.3 View Coordination

Both views in the HierarchicalTopics system are tightly coordinated. On the one hand, selecting a node in the Hierarchical Topic view would highlight a corresponding temporal panel in the Hierarchical ThemeRiver view. This helps users to examine the temporal evolution of the selected topic group. On the other hand, selecting a ribbon in the temporal view will highlight the corresponding node and its path in the topic view. More importantly, once the hierarchy is modified through user interactions in the topic view, the temporal view will also be updated accordingly to reflect the new hierarchical structure.

In summary, the HierarchicalTopics system presents both topic information and temporal evolution of the topics in a hierarchical fashion. This system is designed to aid the exploration of topic content and temporal trends of topic groups through a set of user interactions. In addition, our system allows users to iteratively modify, define, and annotate topic groups based on their interpretation. The HierarchicalTopics provides a flexible visual analytics environment that tightly integrates computational methods with interactive visualizations for analysis of large document collections.

4 Case Study

To qualitatively access the utility of HierarchicalTopics in facilitating the analysis of text corpora with large number of topics, we recruited a senior researcher whose research interests covers HCI and Information Retrieval. This case study is set up for him to explore a collection of NSF awarded proposal abstracts to identify interesting research trends in his research domains. Eighty topics were extracted from 11,961 proposal abstracts funded by all three divisions (IIS, CCF, CNS) in

the CISE (Computer and Information Science and Engineering) directorate from 2005 to 2012.

4.1 Depicting temporal portfolio of NSF programs

Using the Hierarchical Topic view, the researcher started by visually browsing all hierarchical topic groups that are produced by the TRT algorithm. He quickly identified a few topics of interest and interactively merged them into topic groups that fits his analytic goal. The result of his customized grouping and corresponding annotation is shown in the first column in Figure 9. Specifically, two groups of topics are created through the "join" and "collapse" interactions, "*HCI*" and "*Information Retrieval and Data Mining (IR)*".

With the exploration scope narrowed down to these two topic groups, the user wanted to identify and compare the trends in research funding for individual group over the years. Therefore, he turned to the Hierarchical ThemeRiver view and selected the two topic groups so that their research funding trends can be examined and compared.

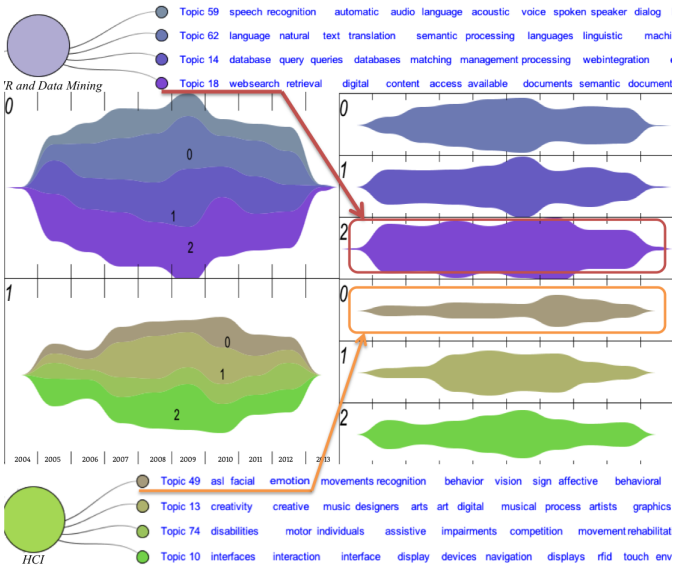


Fig. 9. Case Study: Examination of topic groups of interest. Top (with a purple hue): Topic keywords and temporal trends of the "Information retrieval and data mining" research domain. Bottom (with a green hue): Topic keywords and temporal patterns of the "Human Computer Interaction" field.

The second column in Figure 9 illustrates the overall temporal evolution of selected groups. The user noticed that the trend of proposals awarded under the *IR* group seemed steady with a slight decline over the recent two years. To examine and compare the development of individual topics in the IR group, the users further isolated three topics that are of interest. The corresponding trends for these topics are shown next to the overall trend.

Through quickly examining the volume of each topic trend, the user confirmed his hypothesis that topic 18 on "web search and document retrieval" has continued to be a more popular subfield over the years in terms of NSF research funding (Figure 9 ribbon with red border). However, the user was also surprised when found out that the *"HCI"* group exhibits a slight decline in recent years after a steady growth around 2007. Through examining individual topic trends, more interesting patterns prevailed. Although the overall trends for other topics group have subsided slightly, the research on"*affective computing and emotion related studies*" has gone up significantly in the past two years, as outlined in Orange.

This use case illustrated that the visual interface not only enables the user to view trends for a group of topics that describe a research field, but also permits the discovery of the contributions of individual topics to the overall trends as well as anomalies. According to the user, such analysis gave him valuable insights in understanding the research trends in the areas he is interested in and could potentially help him adjust future proposal focus.

4.2 Identifying program impacts in research

Given that the above two topic groups all exhibits slight downward trending, the user wanted to identify upcoming research topics that received more funding interest in the recent years. He started by mouse hovering over each topic ribbon in the main Hierarchical ThemeRiver view, looking for increasing trends.

Two topic groups caught his attention as shown in Figure 8. Both groups exhibit increasing volume in the past three years, indicating more research proposals were awarded in the two areas. The top row illustrates a topic group related to environmental related research as well as citizen science. As shown in the individual temporal trend for each topic, the user identified that the topic on citizen science and spatial temporal analysis significantly contributed to the recent growth of the focused topic group.

The second row in Figure 8 illustrates a topic group that summarizes research on medical and healthcare related research. Through enabling the time column selection, the user selected proposals related to the health care topic that were awarded in 2012, highlighted in the yellow rectangle. He then discovered that most of the proposals were related to health monitoring and were awarded by the only-recently launched program–Smart and Connected Health (2011).

The user was pleased to find out the impact of a newly established program on research trends and considered the HierachicalTopics a powerful tool in aiding the discovery of the contributors to the temporal changes and possibly the cause for such changes.

5 User Study

To quantitatively evaluate the utility of HierarchicalTopics in aiding users analysis of a text corpus, we conducted a formal user study focusing on comparing hierarchical to non-hierarchical topic structure. Our hypothesis is that the hierarchical topic structure would yield faster identification of topics that are similar in nature.

5.1 Data and Tasks

The dataset used for the user study contains 2453 news articles published between Sept 2012 to March 2013 on CNN.com. Two conditions were designed to evaluate the effect of hierarchical topic structure versus representing them as a flat list of topics. We designed two tasks for the experiment: the first task aims to group individual topics into different news categories; the second task focuses on examining the overall temporal trends for the topics in each news category. For the second task, we required the participants to group all the topics based on their findings in task 1.

Specifically, in task 1, we asked the participants to identify news topics that fall into the following five categories: American Politics, Sports and Entertainment, Natural Disaster, Health-Related Issues, and Middle-East News. An example topic grouping result produced during one of the experiments is shown in Figure 1. Each participant was provided an answer sheet to write down the topic number belonging to each category. For each topic the participants have identified, we also asked them to provide a score (1-5, with 5 as very confident) indicating their confidence of how much the topics fits into their category of choice. For the second task, we asked the participants to group the topics identified in task 1 based on their category. The grouping was done through drag-and-drop interactions within the visual interface. After each group of topic has established, we asked the participants to examine and describe the temporal trend for the topic groups.

To control the complexity of the tasks, we extracted 40 topics from the news corpus. The reason for doing so was that the participants assigned to the non-hierarchical topic organization had to go through the topics one by one. With no initial aid of organizing similar topics together, grouping large number of topics would become laborious and require a lot of repetitions of the same operations. This implies that, if the hierarchical structure proves superior in this study, it will increase its edge relative to a flat structure as the number of topics grows.

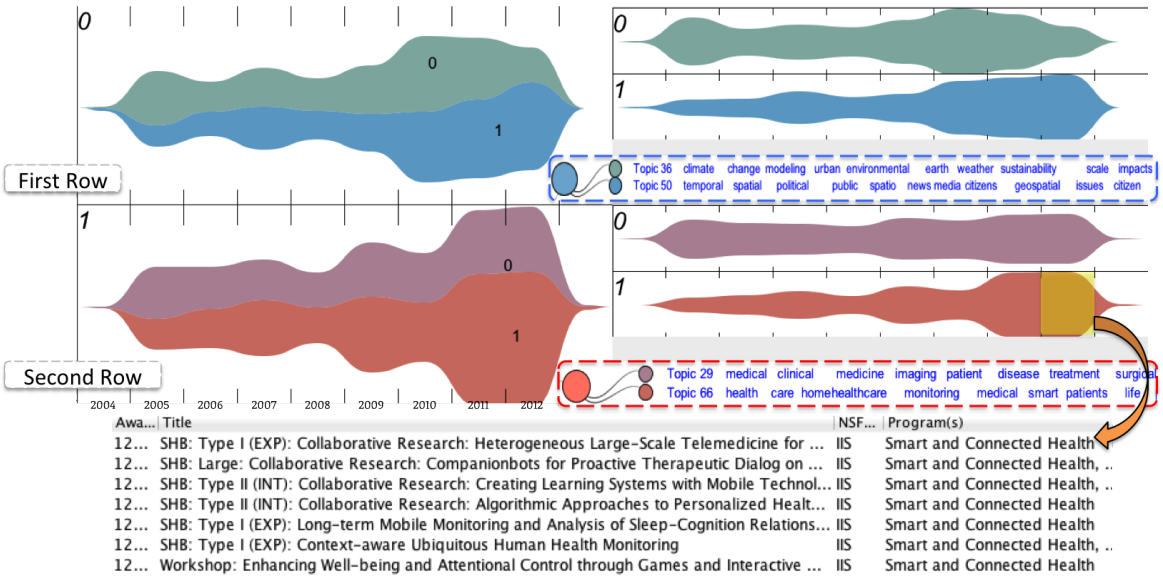


Fig. 8. Case Study: Making sense of increasing topic group trends. Top (with a blue hue): topic group of "environmental and citizen science" has seen recent growth. Middle (with a red hue): heath care related topic group exhibit growth in the past two years, with the "health monitoring" topic as the major contributor to the overall growth. Bottom: detail view showing proposals regarding the "health monitoring" topic awarded in 2012.

5.2 Experiment Design

Eighteen participants took part in the study (13 male, 5 female). The age of the participants ranged from 18 to 34.The study used a between-subjects design. All participants were first provided 10 minutes of training on the HierarchicalTopics visual interface. Each participant was then randomly assigned to one of two conditions (hierarchical vs. non-hierarchical topic organization). The participants were asked to write down their findings on an answer sheet, which records the identified topic numbers for each listed category for the first task and the pattern of the temporal trends for the second task. The experimenter timed the participants for completing each category while they were performing the tasks. The study was conducted in a lab setting, on a computer with two displays (resolution at 2560x1600 and 1920X1200, respectively), 2x 2.66GHz CPU and 12 GB memory.

5.3 Results

For the purpose of analyzing whether the hierarchical topic structure helps the analysis of large text corpora, we calculated the difference of average time for identifying topics for each news category. The average time is computed as the overall time to find all topics for each category, divided by the number of topics identified. The reason for using the average time is because participants identified different number of topics for a given category. In practice, determining whether a topic belongs to a certain category can be subjective. For instance, some participants consider a topic related to the trial of Conrad Murray (the physician for Michael Jackson) belonging to the "Sports and Entertainment" category since it's related to the pop singer. Other participants may consider this being a stretch since Michael Jackson is not the main subject of the news articles related to the topic.

For the same reason, we did not grade the accuracy of the identified topics, since arguments could be made for topics to be included or excluded from a news category. Although we did not grade accuracy of the identified topics, most of the identified topics for each news category did overlap. Two experimenters independently examined each participants' answer, and they did not find answers that are clearly not pertinent to the categories.

5.3.1 Speed: hierarchical topic vs. non-hierarchical topic organization

To measure whether the hierarchical topic organizations yield faster speed for identifying topics for each news category, we performed one-way ANOVA on each category. A significant effect was found for two categories: American Politics and Middle-East News. For the American Politics category, a significant effect of hierarchical topic organization on the time for identifying relevant topics (Task 1) was found at the $p<.05$ level for the two conditions [$F(1,16) = 4.84$, $p = .043$]. For the same category, a significant effect was also found between two conditions [$F(1,16) = 4.79$, $p = .044$] in task 2, which involves grouping the identified topics and observing the temporal trends. For the Middle-East News category, the ANOVA revealed a significance between two conditions [$F(1,16) = 5.15$, $p= .37$]. No significance was found for the other three categories. Detailed results are shown in Figure 10 (left).

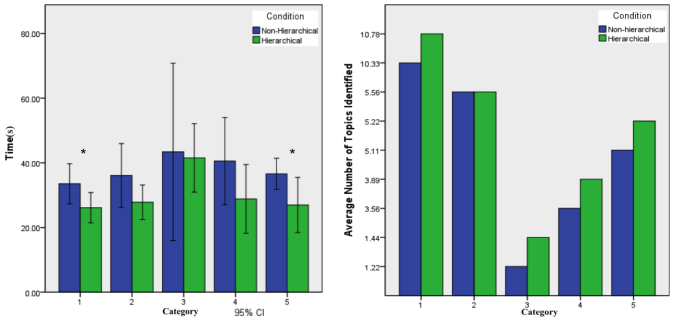


Fig. 10. Left: Average time to identify all topics for each news category during task1. Asterisk denotes significant difference. Right: Average number of topics identified for each news category.

Combining with the average number of topics found in each category shown in Figure 10 (right), the results became more informative. Significant differences were found for categories with relatively large number of topics. In other words, the hierarchical topic structure lead to faster identification and grouping of large number of relevant topics.

5.3.2 User's confidence and Response on potential scalability of the system

As mentioned in section 5.1, during task 1, when a participant jotted down the topics for each category, we have also asked her to provide a confidence value of how well the topic fits into the category. The confidence values for all participants assigned to the hierarchical condition

have a mean of 4.5, with a standard deviation of 0.52. The confidence values for participants assigned to the other condition exhibit a mean of 4.47, and a standard deviation of 0.5. Although no statistical significance was found, the participants under the hierarchal condition consistently reported higher average confidence value for each news category. Note that with 5 as the most confident, the mean values of the confidence show that all participants are fairly certain about their answer. From another perspective, the high confidence values also reflect that the participants could interpret the topics and possibly the topic hierarchies without much difficulty.

The last question on the answer sheeting was regarding the potential scalability of the system. In particular, the question asked the participants to comment on if the HierarchicalTopics could scale to hundreds of topics. We tallied the participants' response. 4 out of 9 participants assigned to the hierarchical condition answered "yes", 4 answered "maybe", and the rest 1 participants answered "no". In contrast, 0 out of 9 participants assigned to the non-hierarchical condition answered "yes" to potential scalability, while 6 answered "maybe" and 3 answered "no".

None of the participants assigned to the non-hierarchical topic condition thought the system could scale to hundreds of topics, while the participants answering "*Maybe*" under the same condition further commented that some sort of automated classification such as topic groups could make the system much more scalable. The participants assigned to the hierarchical topic condition provided more positive responses toward the potential scalability of our system. Several of constructive comments were generated based on user feedback, details of which will be described in the discussion session.

In summary, the study results reveal that hierarchical topic structure leads to more efficient identification and grouping of larger numbers of relevant topics. After performing two tasks through interacting with the visual interface, most participants consider the hierarchical system scalable and bears potential to handle hundreds of topics.

6 Discussion

In this section, we discuss possible improvements on the topic rose tree algorithm and the visual interface.

6.1 Implicit modeling assumptions and design elements

One implicit assumption of organizing large number of topics into a hierarchy is that the topics can fit cleanly into such a structure. However, in practice, such assumption may not always hold. For example, certain topics may fit into multiple groups based on users' interpretation. To address this issue, we could allow users to duplicate topics and add the topics into the corresponding groups.

Another implicit assumption is that we assume that the topic results are fine-grained enough so that the "split" operation is currently not supported in the HierarchicalTopics system. We think the "split" operation is potentially very important since it permit users to directly influence the topic models. However, there are several reasons that the splitting operation is challenging to support. First, asking users to specify how to split the topics (words that should or should not be group) could quickly turn into a laborious task if the interactions are not properly designed. Second, since the the computation of topics usually involves hundreds of interactions, rebuilding the topic model based on users' input of how to split the topics is difficult to achieve in real time [17]. Despite the challenges, we consider the "split" operation a very important option, and a great contribution for interactive visualization to potentially bring to the topic modeling community. Therefore, our future work will try to address this issue and more broadly to permit users to modify the underlying topic model in real or semi-real time.

6.2 Limitation and future improvements on HierachicalTopics system

During the study, the participants provided constructive comments for improving HierarchicialTopics. A few users mentioned the need for annotation feature, which would allow them to annotate or bookmark a general topic group. In addition, users would also like to search for a particular word in the topic view, for the purpose of discovering all topics containing a word of interest. As mentioned in Section 3.3.1, we have already incorporated both the annotation feature and the search function into the current system based on the feedback.

Another interesting comment was on possibly taking advantage of spatial organization of the topics. One participant would like to organize the topics into interested vs. not interested piles and place them on different parts of the screen. Spatial organization is commonly used when working with real objects, and has been shown to aid more complex sense-making processes [1]. Thus more flexible user interactions need to be supported for users to accomplish such task in an un-laborious manner.

During the study, a few participants raised the question of what if one topic falls into two or more topic groups. For example, the topic of human robot interaction could be categorized into both HCI related topic group and Robotics related group. Therefore, we are planning to provide additional user interactions that allow users to duplicate topics and keep track of the duplicates.

Lastly, one limitation arose from the use of tree visualization to represent the hierarchical topic structure. The concern is that tree visualizations may not scale to displaying very large number of topics or multi-level hierarchies. Our HierarchicalTopics system alleviates this issue by supporting multiple user interactions, including collapsing, annotating, and deleting the nodes in the rose tree. Nonetheless, we acknowledge the potential limits of this tree representation and will further explore other visual metaphors.

6.3 Future improvement on the Topic Rose Tree

As of the Topic Rose Tree algorithm, improvements could be added to make the algorithm more transparent and interactive to end-users. For example, when merging two subtrees in each computational step, selecting different operations would yield different results not only in terms of topic groups, but also regarding the depth of the tree. Theoretically, both the absorb and collapse operations would lead to a rose tree with smaller depth compared to the join operation. Trees with less depth may make more sense for grouping topics, since the topics were assumed to be equally descriptive in the topic models. In the hLDA [3], topics on a higher level are usually less meaningful, comprised of mainly stopwords. Thus it makes sense to control the tree depth to be as small as possible. A simple way to influence the depth of Topic Rose Tree is to encourage the absorb and collapse operation rather than the join operation. New interactions could, therefore, be designed to allow users to tweak the weight when calculating the cost of each operation. Such interactions could potentially support advanced users in influencing the topic hierarchy generation. This will be one of the future directions for our visual text analytics research.

7 Conclusion

In this paper, we present HierarchicalTopics, a visual analytics approach to support the analysis of text corpora based on large number of topics. HT is designed to address three challenges faced when analyzing large text corpora through topic based methods. HierarchicalTopics not only provides initial hierarchical structure of topics to facilitate exploration and navigation, it further allows users to modify topic hierarchies based on users' interest through intuitive interactions. In addition, the ThemeRiver in HierarchicalTopics is tailored to represent temporal trends in a hierarchical fashion. It enables the analysis and comparison of groups of topics as opposed to viewing the evolution of one topic at a time. Through both case study and user experiments, we have demonstrated the efficacy of HierarchicalTopics in helping users identifying topics groups, as well as interesting temporal patterns.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Science Foundation under award number SBE-0915528 and DHS VACCINE Center of Excellence.

REFERENCES

[1] C. Andrews, A. Endert, and C. North. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 55–64, New York, NY, USA, 2010. ACM.

[2] D. M. Blei. Probabilistic topic models. *Communication of the ACM*, 55(4):77–84, 2012.

[3] D. M. Blei, T. Gri, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Neural Information Processing Systems(NIPS)*, 2003.

[4] D. M. Blei and J. D. Lafferty. Correlated topic models. *Neural Information Processing Systems*, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] C. Blundell, Y. W. Teh, and K. A. Heller. Discovering nonbinary hierarchical structures with bayesian rose trees. *Mixtures: Estimation and Applications*, April 2011.

[7] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems (NIPS)*, 2009.

[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE VAST*, pages 143–152, 2012.

[9] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

[10] J. Chuang, D. Ramage, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*, 2012.

[11] CNN. Library of congress digs into 170 billion tweets. http://bit.ly/Uwqi7X.

[12] Committee on National Statistics. Science of science and innovation policy principal investigators' workshop. http://bit.ly/10o3via, Sep 2012.

[13] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421, 2011.

[14] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231–240, 2011.

[15] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102, 2012.

[16] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[17] Y. Hu, J. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 248–257, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[18] A. Jinha. Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

[19] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. *Comp. Graph. Forum*, 31(3pt3):1155–1164, June 2012.

[20] Medialab Tools. i want hue web color chooser. http://tools.medialab.sciences-po.fr/iwanthue/, March 2013.

[21] J. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. *Bayesian Analysis*, 7(4):997–1034, 2012.

[22] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. AAAI, 2010.

[23] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[24] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[25] D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1122–1130, New York, NY, USA, 2012. ACM.

[26] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99–106, 2010.

[27] R. M. Shiffrin and K. Börner. Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5183–5185, 2004.

[28] Statisticbrain.com. Facebook statistics. http://bit.ly/YaAVmg.

[29] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[30] The National Science Board. *Science and Engineering Indicators 2010, Chapter 5, Page 29*. National Science Foundation, 2010.

[31] The Unofficial Twitter Resource. Twitter now seeing 400 million tweets per day, increased mobile ad revenue, says ceo. http://bit.ly/JP9DXA, Feb 2013.

[32] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[33] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-SI: Scalable Architecture of Analyzing Latent Topical-Level Information From Social Media Data. *Computer Graphics Forum*, 31(3):1275–1284, 2012.

[34] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[35] ZD Net. Engaging citizens the right way: Government uses twitter during hurricane irene. http://zd.net/mS0aOU, Sep 2011.

Interactive Analysis and Visualization of Situationally Aware Building Evacuations

Jack Guest and Todd Eaglin and Kalpathi Subramanian and William Ribarsky

Charlotte Visualization Center, Department of Computer Science
The University of North Carolina at Charlotte, Charlotte, NC 28269, USA

ABSTRACT

Evacuation of large urban structures, such as campus buildings, arenas or stadiums is of prime interest to emergency responders and planners. Although there is a large body of work on evacuation algorithms and their application, most of these methods are impractical to use in real-world scenarios (non real-time, for instance) or have difficulty handling scenarios with dynamically changing conditions. Our overall goal in this work is towards developing computer visualizations and real-time visual analytic tools for evacuations of large groups of buildings, and in the long-term, integrate this with the street networks in the surrounding areas. A key aspect of our system is to provide *situational awareness and decision support to first responders and emergency planners.* In our earlier work, we demonstrated an evacuation system that employed a modified variant of a heuristic-based evacuation algorithm, that (1) facilitated real-time complex user interaction with first responder teams, in response to information received during the emergency, (2) automatically supported visual reporting tools for spatial occupancy, temporal cues, and procedural recommendations, and (3) multi-scale building models, heuristic evacuation models, and unique graph manipulation techniques for producing near real-time situational awareness. The system was tested in collaboration with our campus police and safety personnel, via a table-top exercise consisting of three different scenarios. In this work, we have redesigned the system to be able to handle larger groups of buildings, in order to move towards a full campus evacuation system. We demonstrate an evacuation simulation involving twenty two buildings in the UNC Charlotte campus. Secondly, the implementation has been redesigned as a WebGL application, facilitating easy dissemination and use by stakeholders.

Keywords: Evacuation, visual analysis, situational awareness, emergency response

1. INTRODUCTION

In any emergency or incident involving large urban structures (arenas, stadiums, college campuses), the safety of the occupants is of paramount importance. Normally, every building has a set of passive safety features (sprinkler systems, fire extinguishers) and evacuation plans or routing maps posted at various points within the building. Current research on studying evacuations has focused on large urban structures or street networks, and have been based on mathematical and algorithmic approaches. These methods study the problem of evacuating the occupants from the structure in the shortest possible time, also known as the egress time. What is lacking in these methods is the ability to handle real-world scenarios involving *large and complex structures that may involve multiple buildings,* and more important, the *ability to react in real-time to dynamic changes in the scenario*, such as blocked stairwells or hallways (due to congestion, smoke), and provide timely recommendations to responders who can mitigate damage, injury, or loss of life. For a commander overseeing the evacuation, the ability to clearly and unambiguously understand the rapidly changing situation is critical, and useful for optimal allocation of limited resources and personnel. Closely related is the ability to visualize the indoor building geometry and the current location of the responders, providing a highly intuitive spatial understanding for informed decision making.

Further author information: (Send correspondence to Kalpathi Subramanian)
Kalpathi Subramanian: E-mail: krs@uncc.edu, Telephone: 1 704 687 8579
Jack Guest: E-mail: jguest@uncc.edu
Todd Eaglin: E-mail: teaglin@uncc.edu
William Ribarsky: E-mail: ribarsky@uncc.edu

In this work, we address some of these challenges, with the primary goal of responding to dynamic events in real-time during an emergency. We begin with an existing heuristic based route planning algorithm,[1] adapt this algorithm within the spatial and capacity constraints for large buildings, and embed it as part of a visual analytic system that permits complex real-time interactions during the event. This allows dynamic changes in the building accessibility to be incorporated, and alternative routing assessed in near real-time. To accomplish this, we employ an Level of Detail(LOD) graph representation of the underlying urban structures that permits real-time recommendations to be presented to the emergency planners and responders for informed decision-making. This is combined with realistic models for building occupancy and traffic flow. Interactive visual analytic tools permit quick exploration of possible impacts of the current situation, and the increased situational awareness for emergency commanders and responders permit more optimal use of scarce resources, for instance, by dispatching responders to areas of need in congested sections of a building, handling casualties, etc.

We begin with a review of current work on evacuation algorithms, followed by a description of the techniques we use to create our 3D models and building networks. This is followed by a description of our modified capacity based route planner, followed by a detailed description of the major features of our interactive visual analytic system and its reporting and recommender functions. We describe the use of this system on typical scenarios using our campus buildings as a test case. We describe our earlier work[2] on evaluating our system and its performance via a table-top exercise, consisting of 3 scenarios: gas leak in building, active shooter, and an explosion in an adjacent structure. We next describe our extension of the system to handle a scenario consisting of 22 campus buildings, towards accomplishing a campus level evacuation that will ultimately include foot paths and the surrounding street networks. We also describe the computational challenges in scaling the system to larger groups of buildings, and our web-based implementation, that will facilitate portability and easier dissemination to end users, by exploiting cross-platform advantages.

2. PREVIOUS WORK

The study of computer based evacuation modeling has evolved with both mathematical and algorithmic approaches. These approaches are categorized as *macroscopic* and *microscopic*. We review these methods, as well as the use of interactive visualization in analysing evacuations.

2.1 Macroscopic Models

Macroscopic approaches focus primarily on minimizing egress time. The evacuees are treated as a unit or a group of units, and moved from source to destination. Interaction among these units is defined by capacity/congestion rules. Linear programming methods based on Network Flow were one of the earliest approaches that yielded optimal solutions, but at high algorithmic cost, making them impractical to use in real-world scenarios; for example, the solution to the Maximum Flow problem has been implemented, with costs as high as $O(n^3)$[1] yielding a best known cost of $O(nm \lg n^2/m)$. Naoyuki Kamiyama et al.[3] apply two initial conditions to the network flow problem. For each vertex the sum of transit times of arcs on any path takes the same value, and for each vertex the minimum cut is determined by the arcs incident to it whose tails are reachable. These assumptions resulted in a 2d grid network and they solved the transhipment problem in $O(n \lg n)$ time. Shekhar and Yoo[4] compare models relevant to the study of nearest neighbor paths. Also Kim, et.al[5] discuss contraflow in reconfigured networks for emergency route planning. In our work, this is relevant, since dynamic changes to the building structure such as blockages or heavy congestion will require modification of the building network, followed by rerouting the affected occupants.

Our model is based on a heuristic approach, the Capacity Constrained Route Planner algorithm proposed by Shekhar et al.[1] This approach attempts to find lower cost algorithmic solutions at the expense of the detail of each evacuee's egress. These approaches are interesting because they can be evaluated quickly from a user perspective as the network-flow problem is reduced to a generalized shortest path problem. The inputs are a graph of the building(or groups of buildings) and the evacuee population. The graph structure consists of nodes comprising various building elements(rooms, corridors, stairwells/elevators) represented as nodes and weighted edges(by distance) representing the relationship between the nodes(paths). The output is a route plan with start times, and a location matrix for each evacuee for each defined time segment. We have adapted this algorithm to meet the more challenging requirements for near real-time decision making in large urban environments, as well as the ability to inject situational changes during an emergency.

2.2 Microscopic models

Microscopic approaches use *agent based modeling*, where each evacuee is governed by unique rules of behavior. Interaction among individuals and their environment are defined based on spatial and social parameters. In addition to the deterministic path or goal, they also rely on behavior rules applied to each evacuee to overcome the "lack of detail" inherent in network flow or heuristic planners. The goal in these methods is also to minimize the time to evacuate individuals to safe zones. Agent based methods can be adaptations of neural networks and fuzzy-logic[6] towards building evacuation simulation. Discrete Particle Swarm Optimization,[7] use of velocity and spatially based rules of interaction[8] are other approaches, as detailed by Castle et al.[9] The most important aspect of agent based models are the rules applied to each evacuee. Castle et. al.[9] describe a detailed list of rules and attributes. We believe our system captures sufficient detail using a congestion model.

2.3 Visual Analytics

Visual analytics involves effectively combining interactive visual displays with computational transformation, processing and filtering of large data.[10] One focus of visual analytics is real-world problems involving situationally-aware decision support. Andrienko et al.[11] described the sequence of tasks that are fundamental to exploratory visualization of spatio-temporal data. They classify their work along 2 dimensions, namely temporal characteristics (spatial location, existence, etc.) versus state of the data(temporal instant, interval, progression over time, and summary). They incorporate these ideas into effectively demonstrating the movement of storks across a geographical area. Many of these ideas are also part of our system in the effective use of animation, prioritized display of significant or critical events during a simulation and their inter-relationships across differents views that make up the final display. Campbell and Weaver[12] investigate situational awareness during emergencies using two different tools: RimSim Response! (RSR) and RimSim Visualization(RSV). Extensive role playing via serious games and post-mortem analysis is the strategy to continually train responders to more effectively respond during emergencies. The RSR tool supports role-playing games on simulated emergency scenarios, while the RSV tool provides interactive visualization tools that support exploration in a non-sequential manner. The authors present results of a hospital evacuation simulation, that involved evacuating 200-250 patients to 20 other hospitals participating in the tabletop and live drills of the exercise. While their tools appear primarily for planning, the long-term goal of our evacuation modules are for use during an emergency; the effectiveness of the use does depend on live information coming into the tool (our system can dynamically adjust to new conditions). We discuss these further in the concluding sections of this article. The work of Kim et al.[13,14] focused on the use of mobile devices for situationally aware emergency response and training, and thus their approach is similar to our work. They demonstrated their system with an evacuation simulation of the Rhode Island club fire of 2003. Our system is considerably more general and scalable to large urban buildings and has the means to interrupt the simulation based on new situational information or dynamic changes.

From a visualization tools standpoint, the use of linked views is important in connecting different representations of information within a single visualization display, and applications specific to urban structures have appeared earlier.[15,16] Mieguins et al.[15] report on the use of coordinated views to analyze an evacuation simulation performed with building occupants tracked by RFID tags. This is, however, an unrealistic expectation in real-world evacuation of large urban environments, but might be of use in understanding user behavior as part of emergency planning. Ivanov et al.[16] report on experiments monitoring a 3000 sq.m office space of 80 people using an array of cameras and motion sensors over a period of 12 months. While this raises privacy issues, some level of monitoring in office spaces is becoming the norm; the acquired information can be used during emergencies to monitor movement and for estimating the number of occupants in buildings. Finally, Andrienko et al.[17] propose 2D summary views of transportation schedules for use in evacuation planning, taking into account the different categories of endangered people (sick, injured, disabled), their sources, destinations and transportation modes. These tools are useful for planning, but insufficient for use during an actual emergency.

3. METHODS

Fig. 1 illustrates the major components of our evacuation system for situationally aware evacuations. There are two major components to the system. Functions in the Route Planner involve a significant amount of a priori processing and initialization. The visual analytic system is a highly interactive system that is user driven and

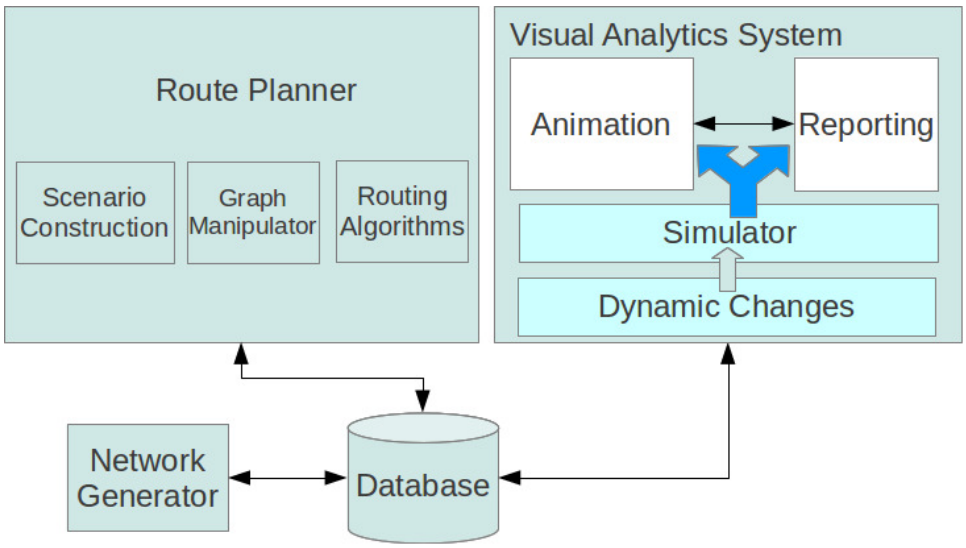


Figure 1. Evacuation System Architecture. Route planner involves preprocessing and routing calculations that serve to initialize the system, with processed results stored in a database. The visual analytic system uses visual abstraction and scalable representations that permit real-time interaction, injection of dynamic situational changes and visual analysis.

can inject and respond to dynamic changes during evacuations. It also consists of reporting and visual analysis functions that can assist an emergency planner in exploring different scenarios, as to the use and deployment of resources, dispatch responders, effect of rerouting occupants, etc.

In our earlier work,[18] we described a semi-automatic system that constructed a *building graph,* incorporating key elements of a georeferenced urban structure critical to evacuations, such as hallways, stairways, elevators and entrances/exits. This building graph generator is used to process urban structures and is stored in a PostgreSQL database. We have successfully processed over 70 buildings on our campus using the graph generator.

We next describe the main components of our evacuation system consisting of the route planner and the visual analytic system.

3.1 Route Planner

The route planner loads evacuation objects, that are combinations of urban structures, pathways, streets etc. Evacuations are built as combinations of structural objects and route planner objects and saved in the database. The Route Planner consists of the following components:

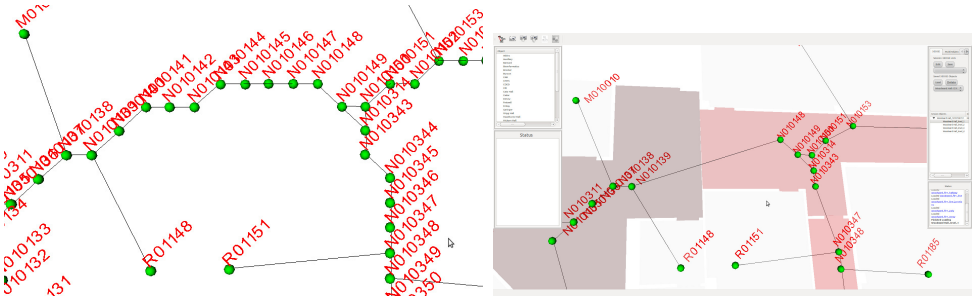
3.1.1 Scenario Construction

Scenario construction includes loading single or multi-building evacuation objects, selecting a route planning algorithm (currently limited to our modified capacity constrained route planner[1]), and setting visualization modes and evacuation parameters, such as building capacity, egress width, evacuee speed and density, and stairway up resistance.

In our scenarios, we use an egress width of 6ft.[19] Building capacity is estimated as a percent of the maximum classroom occupancy(summed up across all the classrooms in the building and known occupancies in other parts of the building). Classroom occupancies are known based on course schedules, and can be used in real-world scenarios. Classroom occupancies are 3 sqft per occupant plus 3ft average separation between seats.[19] Rooms less than 100 sqft. are assumed to have a single occupant(likely, an office). The entire campus evacuation scenario uses actual room capacities taken from the course schedule.

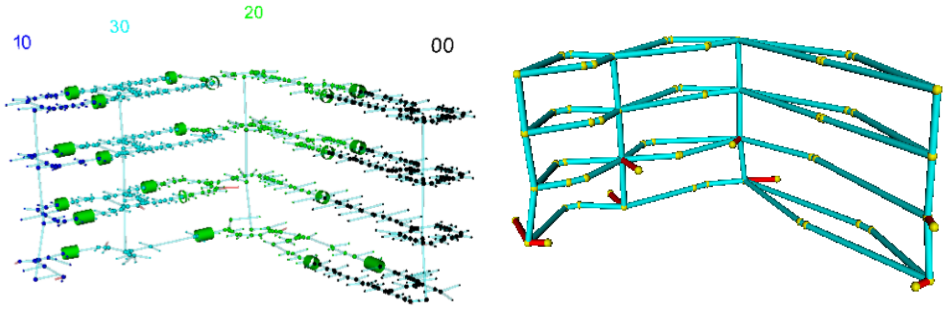
Evacuee speed is set at 3ft/sec.,[20] which is reasonable for urban structure building design considering a fire emergency. The maximum density is set at 6 occupants per 9 sq.ft., a bit on the conservative side. In our model, evacuees move only when the density within this space is less than this maximum, else they are stopped until the density falls below this threshold (when some of the occupants in front begin moving forward).

We also include a parameter to induce occupants to choose exits on the current floor, as opposed to possible exits on adjacent floors. This parameter is termed stairway resistance, and has the effect of increasing the path length up the stairs to the adjacent floor. In our experiments, we set the stairway resistance to 10 times the actual stairway path length.



(a) (b)

Figure 2. Building graph Simplification. Removing redundant nodes. (a) a section of a building floor with labeled building elements, (b) simplified graph.



(a) (b)

Figure 3. Building graph simplification to a zone graph. (a) building graph of a campus building, (b) transformation to zone graph representation. Yellow spheres are nodes and cyan tubes represent edges. The red tubes represent paths to exits

3.1.2 LOD Graph Construction/Manipulation

All computed egress paths are saved in the database as node to node connections. Nodes are defined based on their function, and can be rooms, hallways, stairways, elevators, and exits. The evacuation application also creates muster points, that represent locations where evacuees are ordered to congregate during an emergency. During the process of extracting centerline points, the geometry of hallways is sampled at approximately two foot intervals. This process is necessary to insure accuracy, particularly at corridor elbows. However, this raises computational issues with extremely large buildings or multi-building graphs, and hinders real-time performance(expensive routing calculations) when dynamic changes need to be accommodated during an event. We address this problem by simplifying the graph with two different levels of detail.

Level 1. Remove Redundant Nodes. In this step, we simplify the original building graph by removing nodes that do not impact routing, for instance, shortest path calculations. The larger sampling rate used for accurate centerline calculations results in nodes that can be removed for use in building routes. Nodes and edges are collapsed in the process. Beginning with any node with three or more edges (or any arbitrary node), edges are followed until a node with 3 or more edges is encountered. This becomes a node of the simplified graph and a new edge is created connecting to the previous node, as can be seen in Fig. 2. The process is continued until all nodes have been visited. Our route planner is executed on these simplified graphs for scenarios in which all original egress paths are available. In our experiments, we see a factor of 5-8 reduction in the number of nodes in the simplified graph.

Level 2. Zone Graphs. For rapid computation of paths when dynamic changes are injected during an event, the simplified graphs can still be large, especially in multi-building evacuations. In these circumstances, we further simplify the building graphs into *zone graphs*, by segmenting the building into evacuation sensitive zones: for instance, stairwells, elevators and exits form the critical elements of any egress path. An example zone graph is illustrated in Fig. 3, involving stairwells, hallways and exits(in red).

To compute the zone graph, we use the precomputed evacuee paths to first associate each node with a zone that is closest to it, using an iterative procedure. In the second pass, the graph connectivity is established by keeping track of the zone of related objects that are encountered in these paths (adjacency lists are maintained). Intermediate nodes(the yellow spheres in Fig. 3(b)) are also identified by paths that cross multiple zones and are further used to complete the graph construction. Zone objects span floors in multi-storey structures, with appropriate floor identification for proper path determination during routing calculations. The number of nodes in the resulting zone graph depends on the number of zones and the number of floors in the building. In our experiments, a further factor of 5-7 reduction in the number of graph nodes was seen.

Algorithm 1: Modified Capacity Based Route Planner

```
Input: ;
(1) G(N,E): Directed Graph, N nodes, E edges;
(2) Node Properties: capacity, occupancy;
(3) Edge Properties: capacity, travel time;
(4) Set of Source Nodes;
(5) Set of Destination Nodes;
(6) Set of evacuee objects;
Result: Evacuation Plan : Routes with schedules of evacuees on each route
foreach evacuee i at each source node s do
    path_found = find shortest path p from s to all destinations      with available capacity;
    if path_found then
        while p < max_capacity do
                                        /* can route evacuee via p */;
            evacuees[i].path = p;
        end
    end
    move all evacuees;
end
while evacuees not at destination nodes do
    move all evacuees;
end
```

3.1.3 Routing Algorithms

We have implemented a modified version of the Capacity Constrained Route Planner(CCRP),[1] which is illustrated in Algorithm 1. Given a directed graph with node and edge capacities, the algorithm repeatedly computes the shortest path for each evacuee with available capacity. If a path is found, then it assigns as many evacuees as possible through that path, i.e., until the capacity of any node or edge along the path is exceeded. This is followed by moving all the evacuees at that time step. The process repeats until all evacuees have found paths to exit the structure. The final step is to evacuate the remaining evacuees in the building(who already have paths, but not exited the building).

We have augmented the CCRP algorithm by specifying the movement of the evacuees(in addition to finding the paths) at each iteration. As described in Section 3.1.1, we specify spatial constraints (space occupied and capacity) for each occupant and movement restrictions as a result. Additionally, the algorithm is modified to work with our simplified graphs; in the simplified graphs, weights of the collapsed edges are accumulated and assigned to the new simplified edges. Running the routing algorithm on the simplified graphs makes it more

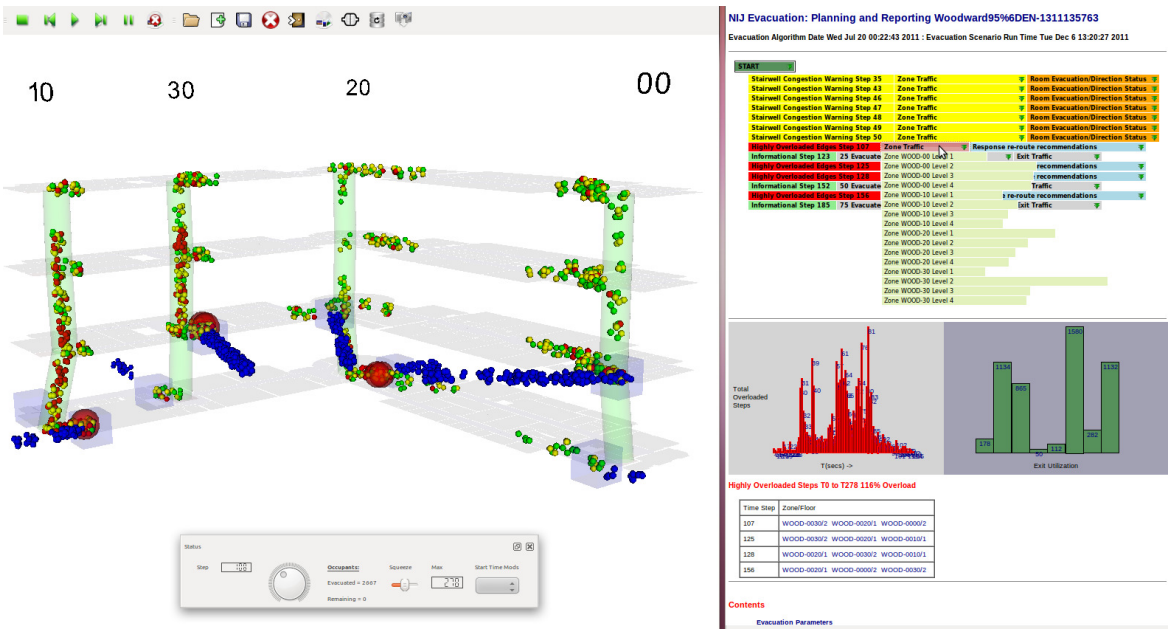


Figure 4. Visualization Design. The upper left panel is the 3D view of the building undergoing an evacuation. Spheres encapsulate evacuee population densities, permitting easy identification of congestion points in the building. Vertical tubes (light green) represent stairways/elevators. Cubes on the first and second floor indicate exits. Lower left indicates the status bar for animation control. The upper right panel is for displaying reports of significant events, that can be drilled down. Lower right panels(bar graphs) show aggregate information on exit occupancy as well as events arranged on a timeline. The report views and the 3D views are linked for immediate updates.

scalable to larger urban structures as well as facilitating dynamic changes to the graph that will require rerouting occupants around blockages or other hazards caused by the emergency event. Finally, although each evacuee has a set average speed (3 ft. per sec.), evacuees cannot exceed the set density threshold. Thus, as congestion builds up, evacuee movements are naturally slowed down. Additional data structures are maintained to make these computations efficient.

3.2 Visual Analytics System

The visual analytic system (Fig. 1) consists of a simulator that accepts user input during an emergency, a 3D interactive animated display of the ongoing evacuation, and reporting and analysis modules. All these views accept direct input and the views are linked to update automatically.

3.2.1 Simulator

The simulator loads evacuation plans that were saved for scenarios under normal static conditions. Since dynamic changes affect only a small part of the structure (for example, a blockage in a stairwell is usually localized, requiring rerouting only occupants close to that area) a large amount of preprocessed data can be reused, contributing to our near real-time performance. The simulator accepts user defined dynamic changes (specification of blockages or casualty reports through the 3D display or the report modules) and modifies computed paths, and reroutes occupants. In addition it updates the generated reports and responder recommendations or action plans.

3.2.2 Visualization Design

Single Building View. Fig. 4 presents our interactive visualization system. Here we see the occupants evacuating from a single 4 storey building. Almost all of the interaction are via *direct manipulation*. There are three components that make up the design. On the left is a 3D animated view of the urban structure, where the user can load and play evacuation simulations. Evacuees are represented as spheres and colored green,

yellow, or red based on low to high congestion. Partially transparent light green tubes represent stairwells, while purple cubes are exits. Blue polygons represent areas that can be occupied. Large red spheres of varying opacity represent edge (capacity) congestion. On the bottom left is the *status tool widget*, that allows moving around in the animation with a slider, indicating current step, evacuee counts and total simulation time. The top right panel is the *significant event* window. The rectangular bars are menus with varying levels of detail of the simulation report. The bottom right panel is a scrolling widget with interactive charts and graphs for interacting with the simulation and visual analysis.

Major features within the single building view are as follows:

- **Congestion Representation.** Congestion is a primary concern in evacuation scenarios and predictions of future congestion and mitigation is important to emergency response teams. In our system, congestion levels range from green to red (low to high).
- **Temporal Cues.** The color and size of spheres is modified at significant event times. For example sphere size is enlarged when the simulator starts moving evacuees from a source location. The resultant pulsing in the animation informs the user of new sources of traffic and likely areas of future congestion.
- **Details on Demand.** The *Significant Event* window in the reporting tool uses a colorized layered menu so that the visualization of significant information is presented as needed by the user. Significant events are evacuation events of interest to first responders in an evaluation of a given scenario. These event definitions came from discussions with UNC Charlotte Police. They include the following:
  - *Informational Events:* Limited in this implementation to a percent of the occupants that have been evacuated. We currently use 25, 50, and 75 percent.
  - *Warnings:* These are capacity related events. Currently set at 90% of calculated zone capacities. They are displayed for every zone at the threshold at any given time step.
  - *Highly Overloaded:* These are also capacity related events. Currently set at 116% of calculated zone capacity. They are displayed for every zone at the threshold at any given time step. This happens when a zone is completely full because each zone shares a node with its neighbor zones. This causes the zone density to rise above 100% because it includes the occupants of its shared node.

  These events allow a maximum amount of reporting while allowing for quick event scanning in the event report. A user sees a limited top level distribution of data unless there is a reason to drill down deeper into the event. In the top right of Fig. 4 the user has selected a *Heavy Zone Congestion* item (in red) with its time step. The user can explore further via a mouseover operation. to reveal a bar graph that shows the relative congestion of each zone in the building.
- **Interaction On Linked Views.** The simulation is manipulated by direct interaction over the 3D animation and report views. For instance, a blockage can be introduced via the 3D view, and simulation rerun to generate new (rerouted) paths for impacted evacuees; the report view is updated to reflect the situational change. Similarly the interaction with the reports menus, charts, etc. temporally updates the 3D animation view. All such operations are performed in near real-time, as the computation is performed on simplified graphs.

Campus View. Fig. 5 illustrates an evacuation simulation of our campus; in this simulation 22 academic, administrative and student resident buildings were involved in the evacuation. Rather than show the evacuation of each building in its entirety, we show a more abstract view of the evacuation so that the incident commander can get a quick overview. We thus use a 2.5D style visualization in the campus view. The building outlines have been extruded to show a 3D view of the campus; the campus buildings are viewed from above, and bars representing the total occupants in that part of the building are drawn at each time step. Each bar represents a 100×100 sq. ft area, and occupants within that area are summed up across all the floors. In order to reduce clutter, only the areas containing significant numbers of occupants are displayed (using a predefined occupancy threshold). As congestion areas tend to be concentrated near stairwells (which form parts of exit routes) this

Figure 5. Campus View. Frame from a simulation of the evacuation from 22 academic, administrative and student resident buildings of UNC Charlotte, with about 15000 occupants. Colored bars above the buildings indicate occupancy counts, each bar representing a 100 by 100 sq.ft area that spans across all the floors of the building. The color coding is normalized against the maximum building occupancy and the height of the bar is normalized against the maximum occupancy of all buildings involved in the evacuation. Large green cylinders indicate staging areas where evacuated occupants gather; the exact staging area chosen depends on the exit used by the occupant to leave the building. This frame shows the activity at 114 seconds into a 633 second evacuation simulation.

strategy is reasonable; bars are colored relative to size of the building occupancy, while the size of the bars are based on the total occupancy across the entire campus. Thus, it is possible to quickly understand the peak occupancies in each building as well as finding peak congested areas in the campus; occupancy ranges are mapped to a discrete set of colors, ranging from green(low) to red(high). Finally, selected areas represented by green cylinders are designated as staging areas for occupants as they exit the building. As the evacuation progresses, these cylinders grow taller, as a function of the number of occupants. Interactive querying for details of a part of the building or the staging areas will be supported in the final implementation, permitting the incident commander to obtain quantitative information during the emergency.

Thus, the campus view provides an overview of the evacuation for a large group of buildings; users can then select a particular building to bring up the building view(see Fig. 4). Alternately, if the focus of the evacuation involves a few specific buildings, then these can be selected and the detailed evacuation of the involved buildings can be analyzed. It is to be noted that the evacuation of all the buildings in the campus view runs concurrently, regardless of what buildings are selected in the current view; the building view simply shows a more detailed view of the occupants and associated geometry (corridors, stairwells, elevators, rooms, exits/entrances) at any instant.

4. IMPLEMENTATION

Unlike our earlier system,[2] our current system uses web based languages and tools to make it highly portable. All of 3D rendering is done using WebGL,[21] an implementation of OpenGL for web browsers. We use the PostgreSQL database with the PostGIS extensions to maintain all information related to the building geometry and for server-client communication. The database is accessed running on the local machine with direct package

calls. The reports section is an HTML/Javascript window inside a QT4 widget. This allows easy porting to mobile device browsers as well as easy dissemination to system users.

Following are the list of major changes that have been implemented as part of the redesign:

1. **Evacuation Planning.** Since the evacuation planning using the CCRP algorithm is computationally intensive, we retained our earlier approach of using the desktop application to compute the evacuation plans as a preprocessing step and store the results in the database. As described earlier, dynamic changes(blockages, rerouting) generally impact a small number of buildings, necessitating updates of their precomputed evacuation plans. All of the data(evacuation scenarios) produced in the database are then available for access by the visual analytics subsystem.

   Table 1 illustrates the performance of the CCRP algorithm performance as the number of buildings involved in the evacuation is increased. As expected, CCRP execution time is affected primarily by graph size, which impacts the shortest path computation. Recall that the CCRP algorithm performs this computation for each evacuee until all evacuees have a path. The number of exits in a building and the number of larger rooms at long distances are also factors in the CCRP execution time.

# Buildings	#Graph Nodes	#Evacuees	CCRP Time(mins)
1	1026	3200	0.7
4	3650	7600	3.8
22	12227	15200	16.3

Table 1. CCRP algorithm performance as a function of number of buildings involved in the evacuation

2. **Browser Visualization.** All of the visualizations are directly supported on the web browser; the animation view is implemented as a WebGL[21] application. WebGL is a browser based implementation of OpenGL ES 2.0, and uses the GL Shading Language[22] for shaders. The Three JS library[23] is used for graphics and rendering. Given that current implementations of WebGL can exploit graphics hardware, very little is lost in terms of geometry rendering efficiency, while application portability is greatly increased.

   The WebGL application loads campus building data (via files in Shape format); a triangulation is performed, followed by an extrusion to create a 3D model. These building models are then concatenated into a single buffer to improve rendering efficiency. The application follows the same procedure for campus footpath geometry. Animation data consisting of the occupancies centered around each area are loaded, followed by generation of 3D bars scaled and colored by the occupancy at each location. The building geometry is rendered using a lambertian lighting model with backface culling. Bars are drawn with a flat color and no lighting to reduce shader computation. Screen space ambient occlusion is used to help distinguish the buildings and bar graphs from the background.

   The remaining views that include generation of congestion reports, exit statistics are implemented using web technologies (HTML), and remain the same from our earlier work(see Fig. 6, right).

3. **User Interaction.** All of the interaction involving specification of blockages and rerouting of occupants around inaccessible areas of the building are sent as requests to the server for processing, and signaled back to the browser upon completion. This approach avoids significant data processing within the browser. The HTML5 functions related to the canvas for WebGL, websockets for persistent connectivity, and webworkers for true threaded browser operations permit this design and allow for a rich and portable evacuation application.

5. EXAMPLE SCENARIOS

Next we describe four experimental scenarios to illustrate the use of our system. The first 3 of these involve a 4 building cluster to illustrate the detailed evacuations within a building. In this scenario, there are 3500 evacuees.

The maximum egress capacity is set to 1 evacuee per cubic foot and navigation speed is set at 3ft/sec (congestion can slow down or halt evacuees during a simulation). The building is loaded to 95 percent capacity.

The fourth scenario involves a campus evacuation involving 22 buildings. This scenario involves nearly 15000 evacuees. It provides a high level view of the campus evacuation with the ability to select particular buildings to see the details of the location of occupants within the building, congestion areas, etc.

5.1 Situation 1: No Blockages

Figure 4 is a screen shot of our visual analytic system, loaded with a campus building with no inaccessible areas. Evacuees are represented by spheres, visually clustered based on egress width. A small red sphere indicates an evacuee cluster on a congested step. Large red spheres of varying opacity indicate congestion greater than 116 percent of rated capacity. In the timestep shown in Fig. 4, the user has clicked on the reporting panel(upper right), representing a highly loaded event at timestep 108 sec. The user has also rolled over the zone congestion bar to reveal the detailed zone congestion graph. As indicated by the bar at Zone 30 Level 2 this is the most traveled and congested route. The application suggests that a responder be dispatched to this area. As the user rolls over the associated 'Response Reroute Recommendations' menu bar in the list the recommended action can be made visible.

Access for responders can be found by looking at the green exit utilization bars(lower right bar chart of Fig. 4), and choosing a low utilization exit. The zone exits in this scenario that are not utilized are on the 1st level. This type of information can be a powerful dispatch tool for the emergency commander to make an informed decision.

Each bullet in the significant event list (upper right panel in Fig. 4) serves as a visual clue to the overall execution of the scenario. The list covers the entire evacuation. The yellow stairway warnings can be drilled deeper to see which areas are becoming congested. These cues are important for responders to quickly react during the beginning of an event or if a campus lock down has been released. The room evacuation, direction status options can be rolled over to indicate the direction from which the traffic is proceeding, which in turn could result in congestion at a later point.

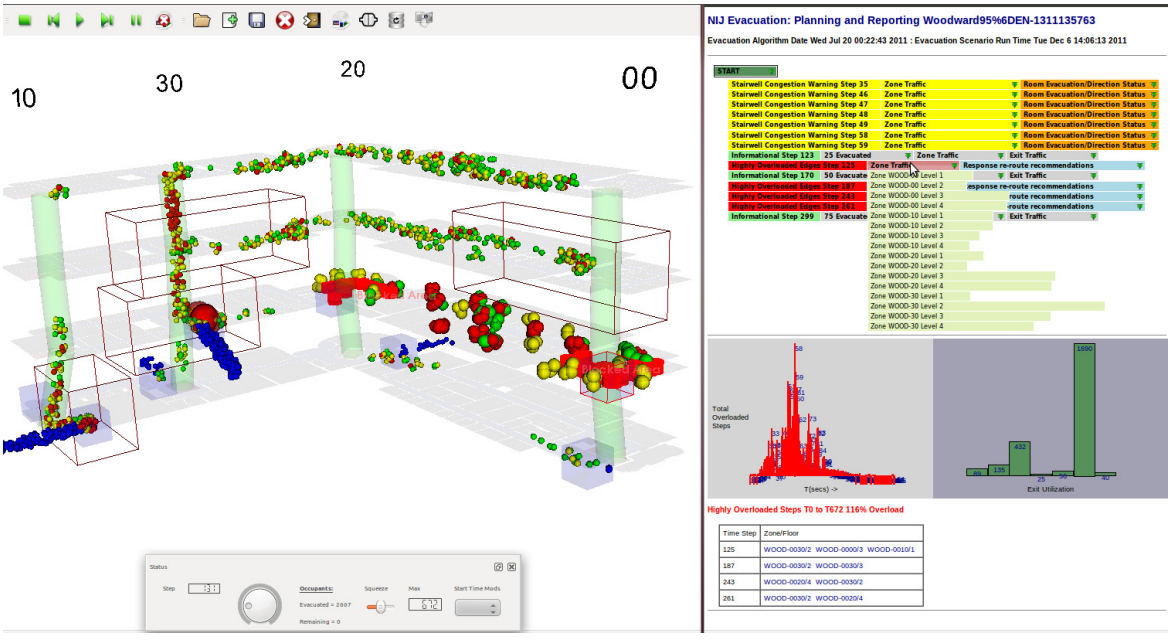


Figure 6. Two blockages have been placed in the building at 45 seconds into an emergency evacuation. (1) Floor 2 at zone 20 stairwell and (2) floor 2 at zone 00 stairwell. Individuals are shown trapped between them by enlarged spheres. Wireframe cubes indicate traffic flow areas, obtained by rolling over the bar chart at the bottom of the report window.

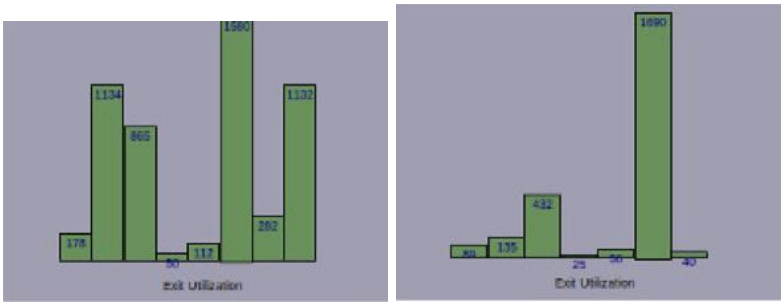
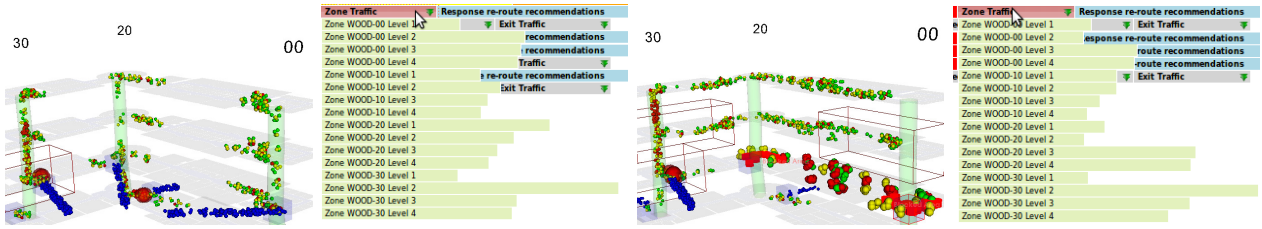


Figure 7. Before and after congestion: Exit utilization.



(a) (b) (c) (d)

Figure 8. Contrast between normal(unblocked) vs. blocked scenarios. Two blockages have been introduced. (a,b) 3D view at 108 seconds into the evacuation, with top floors mostly evacuated, (c,d) 3D view at 139 sec. Floors 3 and 4 are still heavily occupied. Zone traffic (right panels) confirm and illustrate the aggregate picture of the traffic across the entire evacuation.

5.2 Situation 2: Induced Blockages

As shown in Fig. 6, two blockages have been introduced into the scenario of Fig. 4. The blocked areas represented by red squares spread out over several square feet on the second floor at zone 00 and zone 20. In this example, a total of 3100 evacuees were rerouted and all reporting recalculated in 2.8sec. Note that in the control bar at the bottom of the 3D animation view the maximum evacuation time has increased to nearly 7 minutes from less than 3 minutes.

Fig. 7 shows the drastic shifts in the movement of people from a standard evacuation of the building. This example serves to show that evacuation modeling of normal (non-blocked) scenarios is considerably different than when blockages are introduced. In particular, notice the difference in the utilization of the exits. In the blocked case, the congestion occurs earlier and is steeper, resulting in longer times for all evacuees to exit the building.

Mouse rollover on the significant event list view in Fig. 6 shows that the third and fourth floors are getting backed up above the exit at zone 20 floor 2 which is loaded heavily even during a normal scenario. Because the event occurred early there were a number of evacuees occupying the upper floors.

Fig. 8 further contrasts the blocked and non-blocked cases. Here Figs. 8a,b illustrates the unblocked case 108 sec. into the evacuation, and Figs. 8c,d for the blocked case at 131 sec. We compare the zone traffic via the light green bar charts. The bar charts show total zone traffic from the beginning to the end of the evacuation scenario. Even though the bar chart for the blocked case is the total picture it is still clear from both the building and zone traffic charts that the traffic on the top two floors is heavier in the blocked case, and in particular, is shown by the size of the bars labeled Zone WOOD-20 Level 3, Zone WOOD-20 Level 4, Zone WOOD-30 Level 3, and Zone WOOD-30 Level 4. The bar chart and the single step picture of the building point the user to the same conclusion.

5.3 Situation 3: Rerouting Evacuees

When certain parts of a building are blocked, we can reroute evacuees in that area to other nearby less utilized exits. Also, our system permits a selected number of evacuees to be rerouted to reduce congestion at a stairwell or exit. This operation is performed in near real-time (less than 30sec.) in our experiments; however, it depends on the number of evacuees being rerouted. The simulation can then be played, to evaluate the traffic or congestion

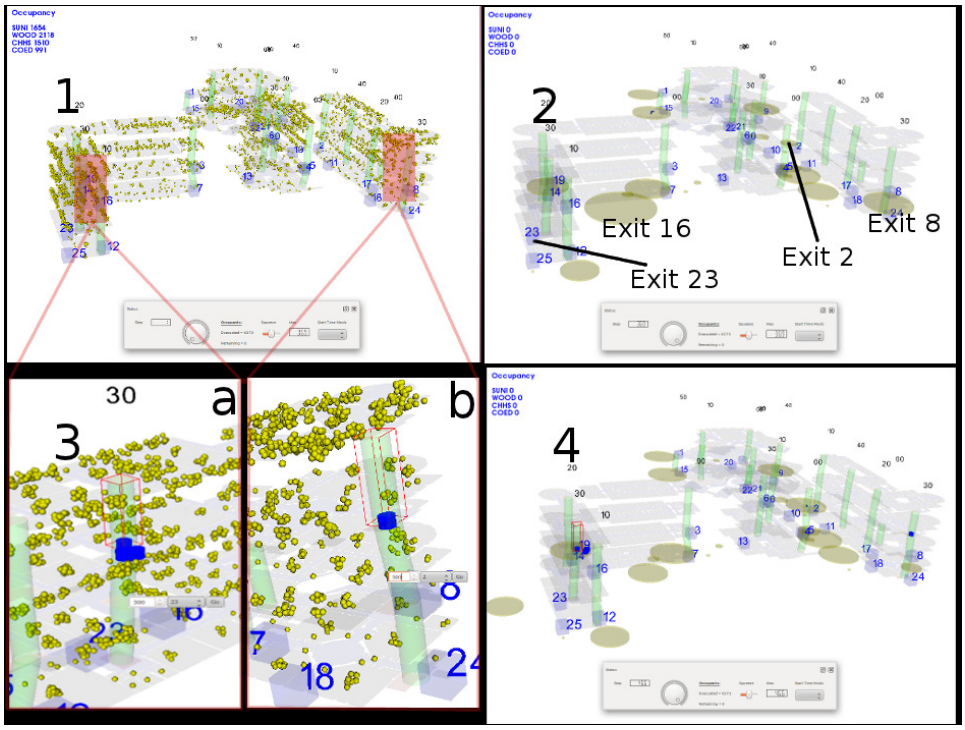


Figure 9. Rerouting evacuees from congested areas in a 4 building evacuation simulation. (1) Four building cluster, (2) Resulting evacuee density circles after simulation, (3a, 3b) User added rerouting flags as indicated by the blue discs and associated with the opaque red rectangles in (1), (4) Resulting evacuee density circles after modified simulation, changing the routes of evacuees to exit 2 and exit 23 in their respective buildings.

patterns resulting from such an intervention. If needed, a responder can then be dispatched to the affected area for assistance.

Figure 9 illustrates an example evacuation from a cluster of 4 academic buildings. The original evacuee densities are illustrated in panel 2. Exits 8 and 16 are heavily used, as indicated by the area of their exit circles. The red cubes in panel 1 have been interactively selected(panel 3 shows a zoomed-in view of these areas) for rerouting occupants within those areas. This is followed by specifying the number of evacuees to be rerouted to specific exits (here exits 2 and 23 were chosen). Panel 4 shows the results of these actions, leading to reduced densities at exits 8 and 23.

Building lockdown and release operations can benefit from such 'what-if' style scenarios that brings together rich spatio-temporal information into the hands of first responders. In this example, running the entire scenario from initiation to results and analysis took approximately 2 minutes. When large collections of buildings are involved with traffic routed to the adjacent street networks, such tools can be invaluable for effective and timely evacuation as well as optimal asset deployment.

5.4 Situation 4: Campus Evacuation

As described in Section 3.2.2 we have extended the evacuation system from our earlier work[2] to handle evacuation simultaneously from a large number of buildings. Using this campus level view, we have run a simulation involving roughly 15000 evacuees distributed across 22 academic, administration and student resident buildings of the UNC Charlotte campus. Table 2 illustrates the occupancy of these buildings used in the simulation. The evacuation was run over 709 seconds (time steps), about 11.8min.

Fig. 10 illustrates 3 key frames during the simulation. The simulation consists fo 633 time steps(seconds). In the top frame, at 114 seconds, peak activity can be seen in several buildings, with the red bars indicating high occupancy in each building, relative to its maximum occupancy. The 5 green cylinders indicate evacuees

Building	#Occupants	Building	#Occupants
Atkins	557	Grigg	483
Biol	549	Kennedy	201
Burson	471	King	56
Cato	171	Macy	117
Cedar	105	McEniry	876
Denny	123	Reese	539
Coed	837	Robinson	847
CHHS	913	Winningham	61
CRI	778	Student Union	1364
Fretwell	3245	Woodward	1748
Friday	632		

Table 2. Campus Evacuation Simulation: Involved buildings and their occupancies. There were close to a total of 15000 occupants on this evacuation simulation.

making their way out of the buildings toward their nearest staging areas. In the middle frame, at 291 seconds, the evacuation is tapering off, as seen by the large green cylinders indicating the number of evacuees that have exited the building. At this stage, there are 6-7 buildings, which started with a much large number of occupants, that remain to be evacuated. In the bottom frame, at time step 450, evacuation is almost complete, with perhaps 2-3 buildings in their last stages of evacuation.

6. EVALUATION: TABLETOP EXERCISE

The development of our application has included regular feedback and demonstrations with campus emergency and safety personnel, including the chief of police, other safety officers, and campus business continuity staff. As part of evaluating the system, we conducted a table top exercise with our campus police. We ran the application through three different scenarios to determine our system's usability, effectiveness, and need for improvements. A business continuity office staff member designed the scenarios. The campus police chief, a senior police officer, and the software team participated in the exercise.

All three scenarios involved a cluster of four campus buildings and a base scenario for the evacuation of approximately 5000 evacuees. **The preprocessing step (performed once, at the beginning) was timed at approximately 8 minutes.** All simulations used this base evacuation object. Video of each of the 3 exercises were recorded for analysis, followed by feedback from the emergency personnel. The system was operated by a member of the software team while commands were received from the police chief.

6.1 Scenario 1. Gas Leak in Building.

Figure 11 shows time sequenced snapshots of a simulated gas leak somewhere in the exercise area. Initially the gas leak was reported as "near Woodward Hall". The police chief requested a simulation start. As seen in Fig. 11(a) the buildings are being evacuated as expected with all exits being utilized. Several seconds into the simulation (sim time: 7:26:38) a report is received that the leak is in the "courtyard", as shown in the red ellipse in the figure. The simulation is halted and reset. The police chief instructed first responders to be dispatched to the building exits facing the courtyard. Also, entrance/exits into the courtyard were to be blocked from further use.

The simulation was restarted based on the new situation. We interacted with the software by placing blockages at the requested areas from 7:27:27 until 7:28:19 (Figs. 11(a), 11(b)). At this point, the software began to recalculate the 5000 evacuee paths. At 7:29:08 calculations were completed and the reporting process rebuilt, including the scenario timeline and the temporal congestion and exit utilization charts. The police chief requested to see the simulation based on the new situation.

Evacuees are confirmed to be exiting the buildings away from the hazard, as seen in Figs. 11(c), 11(d). The simulated time to exit all buildings increased from 318 seconds to 687 seconds. There were large evacuee

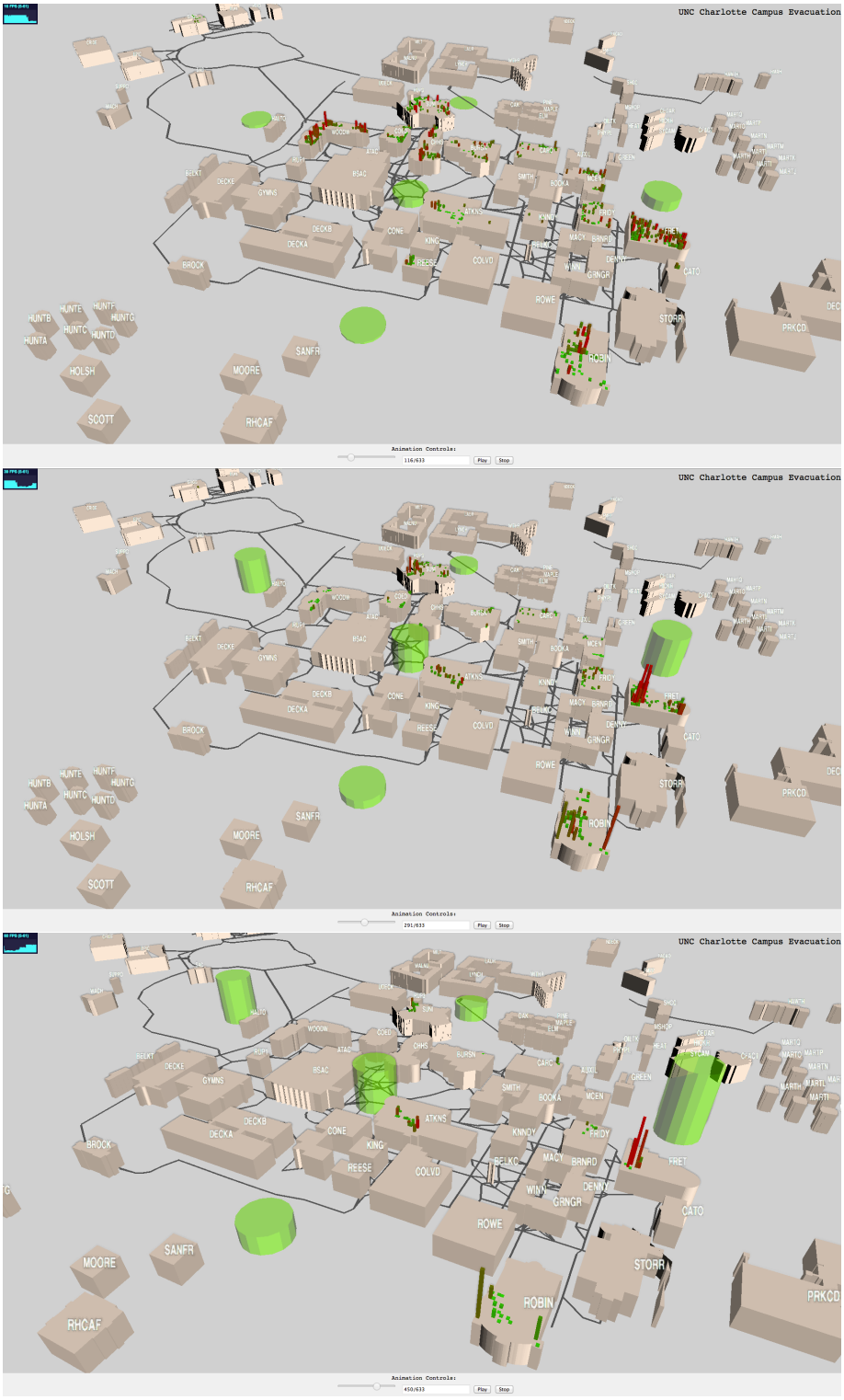
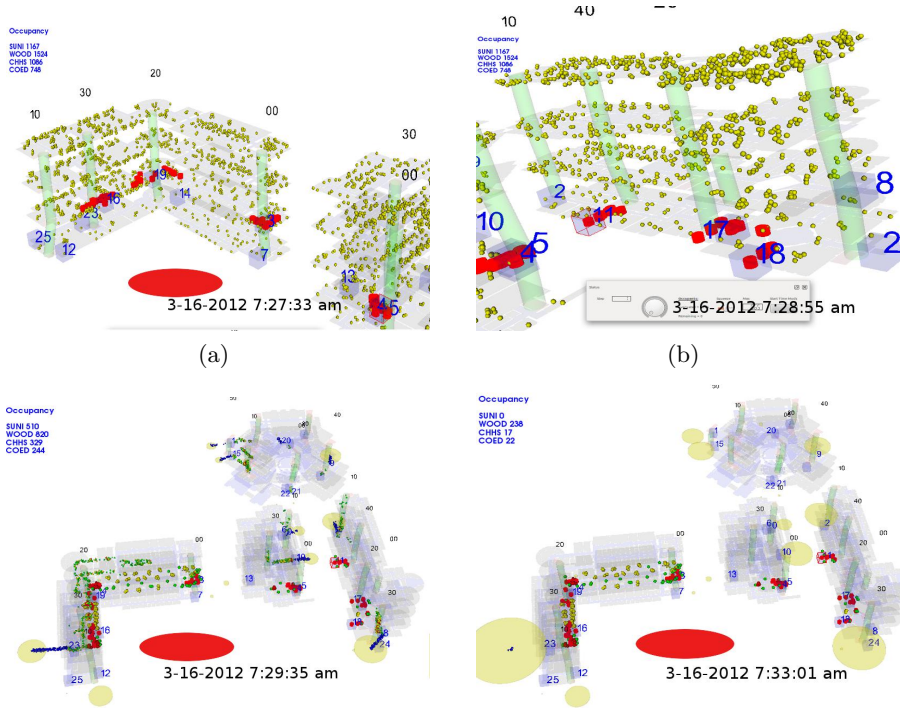


Figure 10. UNC Charlotte Campus Evacuation Simulation. A group of 22 academic, administrative and student resident buildings were involved in this evacuation simulation, comprising a total of about 15000 evacuees. Color coded bars above buildings indicate the number of occupants in that part of the building. Large green cylinders indicate staging areas(muster points) for evacuees to gather for further instructions. *(Top.)* Evacuation reaches a peak across most of the buildings, *(Middle.)* Evacuation is tapering off, and *(Bottom.)* Evacuation is almost complete except for a few buildings with higher populations (notice the green cylinders have reached their maximum size)



(a) (b)

(c) (d)

Figure 11. Tabletop Exercise: Gas Leak in Building. At approximately 7:20 a gas leak is reported. Lower quad campus possibly affecting four buildings. Evacuation simulation begins. At **Sim. Time: 7:26:38.**, leak confirmed near red ellipse. Simulation suspended, assets are deployed to prevent evacuation into the hazard. Views modified to simulate asset activities at building exits. (a)**Sim. Time; 7:27:33.** Blocking building exits into quad (b)**Sim. Time; 7:28:55.** Blocking building exits from adjacent buildings into quad complete. Signal sent for application to perform situationally aware rerouting, (c) **Sim Time; 7:29:35.** Visualization of new simulation, evacuation in progress, simulating responders interaction at exits to affected areas, (d) **Sim. Time; 7:33:01.** Evacuees are avoiding hazard and and exiting to safe zones, evacuee densities are indicated by areas of yellow circles.

populations in the areas of Woodward hall opposite the hazard and it was noted that due to the blockages in the second floor, some of the evacuees were trapped.

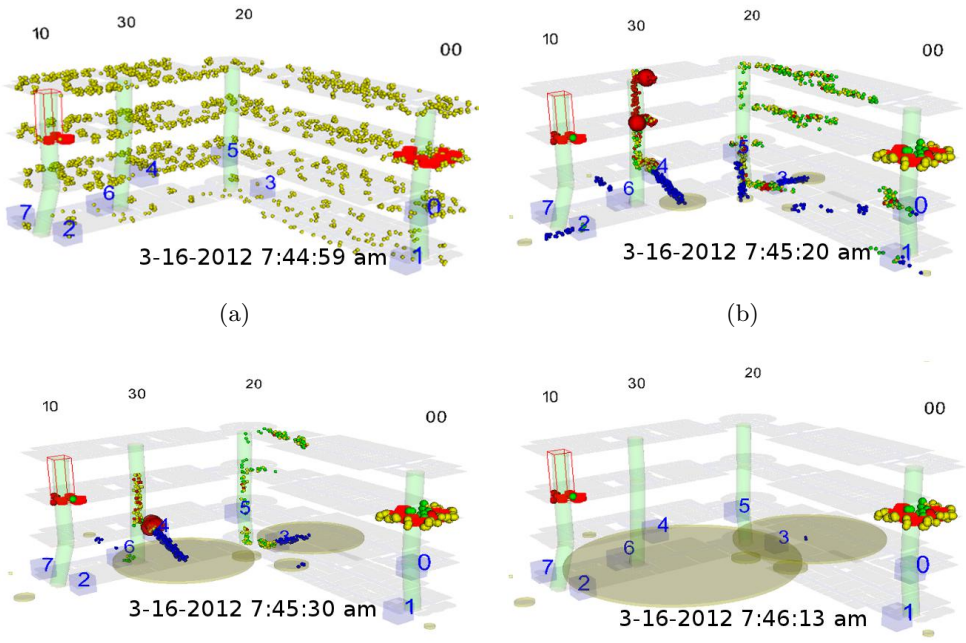
6.2 Scenario 2. Active Shooter in Building.

Figure 12 shows time sequenced snapshots of a simulated active shooter exercise in the Woodward hall. First, the police chief ordered a campus lock down and the building to be evacuated. At this point we switched from the base evacuation scenario of the lower quad (building cluster) to a base scenario of Woodward hall. We could have placed blockages in the locked down buildings but chose to open a single building scenario for the purposes of the exercise.

Some highlights in this exercise include: building rerouting and reporting occurs in 49 seconds (3 seconds for evacuee rerouting and 46 seconds for report generation). The total time here is similar to the multi-building evacuation because our base scenario included 3600 evacuees. This simulates a highly overloaded building to exercise the software for testing.

6.3 Scenario 3. Explosion in Utility Plant

An explosion in the RUP(regional utility plant) building created a scenario where the four building evacuation simulation of Figure 11 was also used. This scenario also found evacuees blocked in the upper floors and the explosion created a hazard in the building courtyard. As reports were received the building floors were blocked and the simulation was started. As more reports were received it became obvious that the personnel would



(a) (b)

(c) (d)

Figure 12. Tabletop Exercise:Active Shooter. At approximately 7:40 a shooter is reported at Woodward hall. Campus is locked down. Reports are received that 3rd floor stairwells are blocked at each end of the building. At 7:44:10 blockages are placed in Woodward by the operator and the simulation is recalculated. (a)**Sim. Time; 7:44:59.** Processing for new evacuation simulation is complete, commander orders visualization of new simulation, (b) **Sim. Time; 7:45:20.** Trapped evacuees noted at 3rd floor zone 00 and zone 10, (c)**Sim. Time; 7:45:30.** Extreme congestion noted at stairwells in floors 2, 3, and 4 at zone 30. (d)**Sim. Time; 7:46:13.** Simulated evacuation complete. Evacuee populations are indicated by approximate area occupied circles.

exit toward the hazard in the courtyard. The exit density circles alerted the police chief to this problem and emergency personnel were dispatched to redirect these evacuees. At this point the police chief requested the exits facing the courtyard to be blocked. The simulation was restarted and evacuation times and exit results were evaluated as in previous scenarios.

6.4 Analysis and System Assessment

We detail below both the observations from first responders as well as the important features and current limitations of our evacuation system, as noted from the table top exercise. Overall, the feedback from the chief of police (who played the role of incident commander) and his officers was positive and consisted of the following observations:

- The ability to see the 3D layout of the buildings and surrounding areas and get a sense of the current situation was considered the most valuable. The ability to see the evacuation unfold, the buildup at congestion points and the ability to direct evacuees away from a hazard were considered critically important.
- The near real-time responsiveness of the system and the ability to see the evacuation under blockages was valuable for assessment and taking appropriate action, such as dispatching first responders.
- In the gas leak exercise, a review of the evacuation helped the commander quickly size up the situation (number of evacuees, exit routes, etc) and order a building evacuation. Once the hazard was located, evacuees were routed away from it by injecting suitable blockages at key points in the building.
- In the active shooter scenario, the police chief noted that the exit utilization and congestion reports would be an invaluable tool for first responders to analyze the condition of a building and dispatch personnel.

- Additional work on the user interface will be needed to further minimize delays during a dynamically changing situation; for instance, blockages are specified one at a time; a 'lasso' style interface to specify multiple blockages was considered more intuitive and efficient.
- A limitation of the current system is its inability to localize blockages to the stairways or exits, trapping evacuees in the vicinity. This will be addressed by rerouting the evacuees in the blocked areas to other exits.
- A visualization issue that is common to visual analytic systems is visual clutter and the ability to unambiguously visualize critical information. As we extend our system to incorporate tens of buildings in evacuation scenario, these issues will require careful design and representation choices, with input from responders.

7. CONCLUSIONS

In this work we have presented an interactive visual analytic system for situationally aware evacuations of large urban structures. The goals of this work were to provide interactive visual analytic tools that can be used in real-world scenarios (large urban structures, dense collections of buildings) and more importantly, be able to run evacuation scenarios in the context of dynamically changing conditions. This is the most important contribution of this work and distinguishes it from previous work on evacuation planning/simulation. We have developed and used an LOD representation of building graphs that can be used as part of a visual analytic system for near real-time response. This in turn permits situational changes to be incorporated into the underlying models and evacuees rerouted. Also, our visual analytic system provides recommendations through the reporting functions that can be used for effective use of scarce resources in dispatching responders to areas of need during the emergency.

We extended our earlier work[2] to handle large networks of buildings by augmenting the building evacuation view with a campus level visualization, that permits the incident commander to obtain an overview of the evacuation, while permitting any individual building evacuation to be examined in more detail. We demonstrated our system with an evacuation scenario of 22 academic, administration and student residence buildings of UNC Charlotte. Depending on the total geometry rendered and the number of occupants, we are able to maintain a near real-time frame rate most of the time, and (ranging from 4-60 frames per second). Future work will require additional simplification of the visualization to handle large neighborhoods, as the system is scaled upto handle urban environments.

We evaluated our system with first responders, including the campus police chief, a senior police officer and public safety and business continuity/planning personnel. Input from these experienced personnel was invaluable. A tabletop exercise was performed with a smaller cluster of four buildings, with three different scenarios(gas leak, active shooter, and explosion) overseen by the police chief, acting as the situation commander. Overall, the system performed well, as evidenced by direct feedback from the first responders, with valuable suggestions to improve the system.

Extension of the evacuation to the larger campus building network was only recently completed, as well as the transition towards a web based implementation of the visualization. We are planning formal evaluation of the system with emergency responders. A full campus evacuation would also require routing occupants via the foot paths to parking lots and out of the campus. This work is only just beginning, as it would require control of the traffic flow, in a manner similar to the contra-flow work of Kim et. al.[24]

Experiments with our current system takes on the order of about 8 min of preprocessing time, a significant amount of time, given that the evacuation itself might be completed in a fairly short amount of time; the computation time is dependent on the building geometry, number of available exits, etc. Thus, future work has to focus on the ability to reduce the preprocessing time. There are two ways to significantly reduce the preprocessing time, (1) we can take advantage of known occupancies of urban structures; for campus buildings, it is possible to estimate occupancy based on the knowledge of class schedules, number of offices and their occupancy(staff, faculty, laboratory). These can be integrated into the evacuation system as part of scenario construction, and, (2) a large part of the route planner computation only depends on the building geometry

that very seldom undergoes significant change; thus routes within the building to the exits can be precomputed and stored for later retrieval. Beyond this, improved performance using more powerful computational resources, including more efficient and parallel implementation of key aspects of the route planner will improve the scalability of the system.

Finally, our goal for this evacuation system is for its use during an emergency. While more powerful computational resources are helpful, it is also useful to have knowledge of the number of occupants that remain in the building(s) as the evacuation progresses. In the work of Meguins et al.[15] RFID tags were attached to each occupant. This is infeasible in real-world situations, and raises privacy issues. Most current buildings (including the newer buildings on the UNC Charlotte campus) include motion and door sensors that can monitor passing traffic. These can be used to estimate the number of occupants entering or leaving a building via each exit, corridors, stairwells, etc. Such information, if available, can be acquired and integrated into the evacuation system. The accuracy of these estimates can be validated by manual collection of the data over the period of a normal week during a semester (for campus building scenarios, at least) and appropriate corrections applied to the estimates.

8. ACKNOWLEDGMENTS

This work was supported by a grant(award 2009-SQ-B9-K009) from the National Institute of Justice, Office of Justice Programs.

REFERENCES

[1] Lu, Q., George, B., and Shekhar., S., "Capacity constrained routing algorithms for evacuation planning: A summary of results.," *Springer-Verlag Berlin Heidelberg 2005* **3633**, 291–307 (Sept. 2005).

[2] Guest, J., Eaglin, T., Subramanian, K., and Ribarsky, W., "Visual analysis of situationally aware building evacuations," in [*IS&T/SPIE Electronic Imaging*], 86540G–86540G, International Society for Optics and Photonics (2013).

[3] Kamiyama, N., Katoh, N., and Takizawa, A., "An efficient algorithm for the evacuation problem in a certain class of networks with uniform path-lengths.," *Discrete Applied Mathematics* **157**, 3665–3677 (2009).

[4] Shekhar, S. and Yoo., J. S., "Processing in-route nearest neighbor queries: a comparison of alternative approaches," in [*Proceedings of the 11th ACM international symposium on Advances in geographic information systems*], (2003).

[5] Kim, S., Shekhar, S., and Min, M., "Contraflow transportation network reconfiguration for evacuation route planning," *IEEE Trans. on Knowl. and Data Eng.* **20**, 1115–1129 (August 2008).

[6] Lo, S. M., Liu, M., and Yuen, R. K. K., "An artificial neural-network based predictive model for pre-evacuation human response in domestic building fire," *Fire Technology* **45**, 431–449 (Sept. 2009).

[7] Fang, G., "Swarm interaction-based simulation of ocucupant evacuation.," in [*2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*], (2008).

[8] Guy, S. J., Chhugani, J., Kim, C., Satish, N., Lin, M., Manocha, D., and Dubey, P., "Clearpath: Highly parallel collision avoidance for multi-agent simulation," in [*Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2009)*], Grinspun, E. and Hodgins, J., eds. (2009).

[9] Castle, C. J. E. and Crooks., A. T., "Principles and concepts of agent-based modelling for developing geospatial simulations," in [*Centre for Advanced Spatial Analysis. University College London*], **110**, 1–52 (2007).

[10] Thomas, J. and Cook, K., [*Illuminating the Path: The Research and Development Agenda for Visual Analytics*], IEEE Press (2005).

[11] Andrienko, N., Andrienko, G., and Gatalsky., P., "Towards exploratory visualization of spatio-temporal data.," in [*3rd AGILE Conference on Geo-graphic Information Science*], (2000).

[12] Campbell, B. and Weaver, C., "Rimsim response hospital evacuation: Improving situation awareness and insight through serious games play and analysis," *Journal of Information Systems for Crisis Response and Management* **3**, 1–15 (Jul-Sept 2011).

[13] Kim, S., Maciejewski, R., Ostmo, K., Delp, E., Collins, T., and Ebert, D., "Mobile analytics for emergency response and training," *Information Visualization* **7**(1), 77–88 (2008).

[14] Kim, S., Yang, Y., Mellama, A., Ebert, D., and Collins, T., "Visual analytics on mobile devices for emergency response," in [*IEEE Symposium on Visual Analytics Science and Technology (VAST)*], 35–42 (2007).

[15] Meiguins, B. and Meiguins, A., "Multiple coordinated views supporting visual analytics," in [*Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*], 40–45, ACM, New York, NY, USA (2009).

[16] Ivanov, Y., Wren, C., Sorokin, A., and Kaur, I., "Visualizing the history of living spaces.," *IEEE Transactions on Visualization and Computer Graphics* **110**, 1153–1159 (November 2007).

[17] Andrienko, G., Andrienko, N., and Bartling, U., "Interactive visual interfaces for evacuation planning," in [*Working Conference on Advanced Visual Interfaces(AVI) 2008 Proceedings*], 472–473, ACM Press (2008).

[18] J.Liu, K.Lyons, Subramanian, K., and Ribarsky, W., "Semi-automated processing and routing within indoor structures for emergency response applications," in [*Proceedings of SPIE Conference on Defense, Security, and Sensing*], (Apr. 2010).

[19] Assembly, N. C. G., "North carolina state building code." `http://www.ncga.state.nc.us/EnactedLegislation/Statutes/HTML/BySection/Chapter_143/GS_143-138.html` (2012). [Online; accessed Aug-2013].

[20] CFPA-Europe, "European guideline:fire safety engineering concerning evacuation from buildings." `http://www.cfpa-e.eu/cfpa-e-guidelines/guidelines-fire-protection/CFPA_E_Guideline_No_19_2009.pdf` (2009). [Online; accessed Aug-2013].

[21] "WebGL:OpenGL ES 2.0 for the Web." http://www.khronos.org/webgl/.

[22] "OpenGL Shading Language." http://www.opengl.org/documentation/glsl.

[23] "three.js." http://threejs.org.

[24] Kim, S., Shekhar, S., and Min, M., "Contraflow transportation network reconfiguration for evacuation route planning," *IEEE Trans. on Knowl. and Data Eng.* **20**, 1115–1129 (August 2008).

Making Sense of the Operational Environment through Interactive, Exploratory Visual Analysis

Dr. William J. Tolone / Dr. Xiaoyu Wang / Dr. William Ribarsky
College of Computing and Informatics
University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223
UNITED STATES OF AMERICA

email: William.Tolone@uncc.edu / Xiaoyu.Wang@uncc.edu / ribarsky@uncc.edu

ABSTRACT

The success of defense and security operations depends on the ability to make sense the operational environment and to anticipate those factors that influence operations both negatively and positively. In this paper, we present a framework for interactive, exploratory visual analysis as a means for making sense of the operational environment. This framework is grounded in the sensemaking process and characterized by three essential dimensions: information retrieval/fusion, interactive visualizations and modeling and simulation. The framework reveals several important research gaps that must be filled to provide full benefit to the task of understanding the operational environment. The paper highlights two such gaps: identifying emerging topics and trends, and making sense of integrated modeling and simulation.

1.0 INTRODUCTION

It is widely recognized that effective interaction with local populations is essential to the success of defense and security operations. Effective interaction, however, depends on the ability to make sense the operational environment and to anticipate those factors that influence defense and security operations both negatively and positively. Unfortunately, the structure and behavior of the systems that commonly comprise these factors suggest that making sense of operational environments is a “wicked problem” [1].

Wicked problems are high-stakes, complex problems that are without definitive formulations; they are problems with open solution spaces where solutions have relative quality; and they are problems that are arguably unique in each instance. Given these characteristics, managing wicked problems presents both a difficult and daunting challenge. Fortunately, exploratory capabilities offer a promising approach to managing wicked problems as they provide the foundation for competitive analyses and the study of alternative hypotheses [2].

In this paper, we present a framework for interactive, exploratory visual analysis as a means for making sense of the operational environment. This framework is grounded in the sensemaking process [3] and characterized by three essential dimensions: information retrieval/fusion, interactive visualizations and modeling and simulation. While connecting these three dimensions, the framework reveals several important research gaps that must be filled in order to provide full benefit to the task of understanding the operational environment and those factors that influence the outcomes of defense and security operations. In this paper we describe two of these research gaps and illustrate two applications where sensemaking is enabled by interactive, exploratory visual analysis. We conclude this paper by connecting these applications to the framework and decision cycle by illustrating: i) how the identification of emerging topics and trends can inform modeling and simulation; and ii) how these same techniques can build further understanding when applied to data farming enabled by through modeling and simulation.

2.0 SENSEMAKING AND THE OPERATIONAL ENVIRONMENT

In order to appreciate the "wickedness" associated with the challenge of making sense of the operational environment, it is useful to highlight the characteristics that commonly comprise the operational environment and explore how those characteristics are aligned with military doctrine.

Within U.S. military doctrine, the operational environment is define to be "[a] composite of the conditions, circumstances, and influences that affect the employment of military forces and bear on the decisions of the unit commander" ([4], p. xi). Examining this definition more closely reveals several important insights. First, the operational environment exists in the context of a mission – i.e., the employment (or potential employment) of military forces. In other words, mission requirements are integral to defining the bounds of the operational environment. Second, the operational environment is both situated and dynamic. The operational environment is situated along multiple dimensions – e.g., temporal, geospatial, cultural, etc. – and these dimensions are what give meaning to actions or events within the environment. The operational environment is dynamic as a reflection of changes in conditions, circumstances and influences within the environment. Third, the operational environment is relational, suggesting that there are meaningful patterns and causalities that underlie and explain observable behaviors and changes within the environment.

Under recent U.S. military doctrine, the operational environment is frequently characterized as a combination of the political, military, economic, social, infrastructure, information, physical environment and time (PMESII-PT) factors (identified as operational variables), and their interdependencies, that affect military operations [5]. Each of these factors itself is a complex system exhibiting emergent, nonlinear behavior. In fact, understanding the structure and behavior of any one of these factors is arguably a wicked problem in its own right. Collectively, these dimensions challenge analysts and decision makers, and further stretch analytical thought.

Characterization of the operational environment is often a key element of military doctrine. The Joint Intelligence Preparation of the Operational Environment (JIPOE) [4] is a prime example. JIPOE is a four-step process designed to provide analytical support to decision-making in a joint operational context (see Figure 1). Table 1 summarizes the tasks associated with each step. It is easily observed that central to each step is the construction and maintenance of an understanding of the operational environment, initiated in Step 1 and explored in subsequent steps.

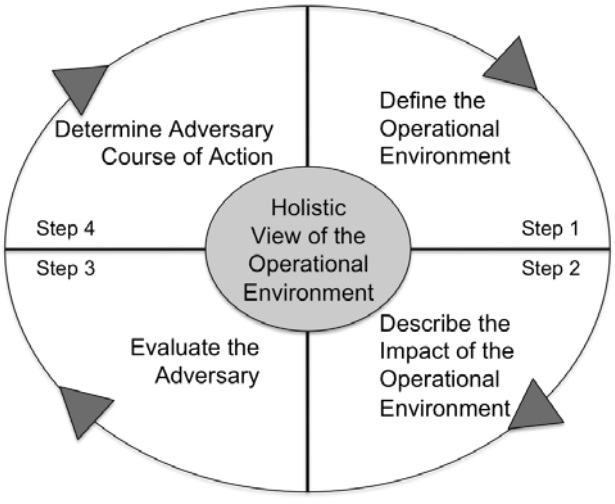


Figure 1 - Joint Intelligence Preparation for the Operational Environment [4]

Table 1 - JIPOE Tasks [4]

Step 1	Step 2
• Identify the joint force's operational area • Analyze the mission and joint force commander's intent • Determine the significant characteristics of the operational environment • Establish the limits of the joint force's areas of interest • Determine the level of detail required and feasible within the time available • Determine intelligence and information gaps, shortfalls, and priorities • Collect material and submit requests for information to support further analysis	• Develop a geospatial perspective of the operational environment • Develop a systems perspective of the operational environment • Describe the impact of the operational environment on adversary and friendly capabilities and broad courses of action
Step 3	**Step 4**
• Update or create adversary models • Determine the current adversary situation • Identify adversary capabilities and vulnerabilities • Identify adversary centers of gravity	• Identify the adversary's likely objectives and desired end states • Identify the full set of adversary courses of action • Evaluate and prioritize each course of action • Develop each course of action in the amount of detail time allows • Identify initial collection requirements

We argue that the process of building a proper understanding of the operational environment as characterized by the complex PMESII-PT variables and described in military doctrine such as JIPOE is largely a sensemaking process. Sensemaking has been described in numerous ways. Duffy [6] states that sensemaking is "how people make sense out of their experience in the world." The final report from the *2001 Sensemaking Symposium* [7] describes sensemaking "as the process of creating situation awareness in situations of uncertainty." Klein et al. [8] describe sensemaking as, "a motivated, continuous effort to understand connections…in order to anticipate their trajectories and act effectively."

Collectively, these descriptions of sensemaking share many common characteristics. First, each characterizes sensemaking as an iterative process with numerous feedback loops – c.f., [3]. Second, each argues that sensemaking involves several activities including foraging, encoding and reasoning. Central to these activities is the iterative construction and refinement of representations, i.e., models – a process that Klein et al. refer to as the framing process [8]. Russell et al. capture this characteristic in their notion of the Learning Loop Complex [9]. In the Learning Loop Complex, people search for a good representation; and, then, instantiate the representation – i.e., encode the data – based of the data available – i.e., data that have been foraged. Those data, called residual data, that do not "fit" the representation lead to the selection, construction or refinement of the representation – i.e., reframing. Third, sensemaking is largely a human-centric activity where judgment and critical thinking play essential roles. This suggests one must abandon the notion that outcomes (e.g., decisions or courses of action) are the output of computational tools; rather, tools should enable the exploration possible outcomes, facilitate human judgment and help to evaluate plausible futures.

Figure 2 is the sensemaking representation that we adopt for the work reported here. In this representation, the progression of data to information, information to knowledge and knowledge to understanding is clearly visible. Information, foraged from data and placed in analytical context, provides the foundational evidences to the analytical question(s). Knowledge, as representations encoded from information, emerges from relationships among concepts [10]. Understanding is synthesized from knowledge through reasoning and critical thought. This progression, however, is not necessary linear and often highly iterative.

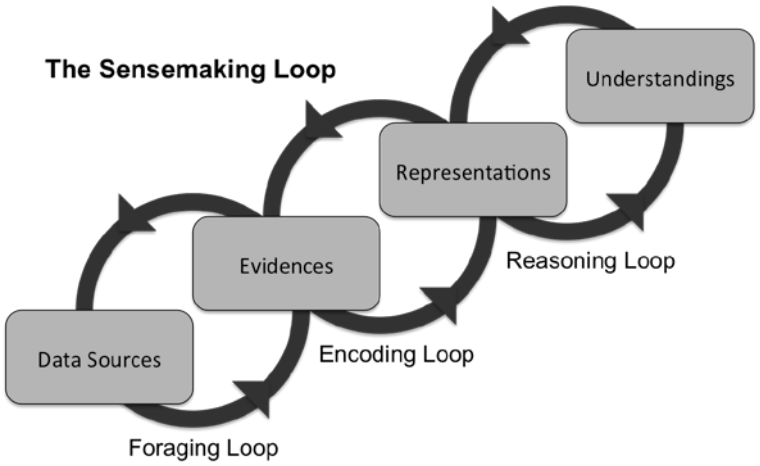


Figure 2 – The Sensemaking Loop

3.0 A FRAMEWORK FOR INTERACTIVE, EXPLORATORY VISUAL ANALYSIS

We contend that exploratory visual analysis offers important affordances to the sensemaking process. In support of this contention, we have developed a framework for interactive, exploratory visual analysis as a means for making sense of the operational environment. The sensemaking process is central to our framework (see Figure 3) with its foraging, encoding and reasoning loops serving as the core analytical activities. The utility of the framework is found in the direction it offers to tool and method research and development in relation to the ability to characterize operational environments quickly and accurately – relative to mission requirements. We believe that the volume, velocity and variety of data that describe the operational environment – e.g., political, economic, military, social, infrastructure, environmental systems, etc. – as well as the complexity of the systems they represent necessitate both knowledge-driven and data-driven approaches to analysis. This accounts for the trifold dimensions of information retrieval/fusion, interactive visualizations and modeling and simulation forming the foundation of the framework.

Knowledge-driven approaches emphasize the application and adaption of existing representations, both cognitive and digital, in support of the sensemaking process by enabling the ongoing reframing of representation as a reflection of current understanding. Modeling and simulation, particularly integrated modeling and simulation (i.e., coupling models representing the various PMESII-PT variables), offer an especially promising research direction as applied to the challenge of making sense of the operational environment,. Data-driven approaches, on the other hand, emphasize the construction of new representations from raw data and the situating of those representations within current understanding – revealing as new frames the hidden structure in large corpora of data. Here, techniques such as automated topic-modeling offer a promising research direction in support of the sensemaking challenge.

A further contribution of this framework, however, is its emphasis on the essential role of exploratory visual analysis in relation to both knowledge and data-driven approaches (as well as the dynamic interplay between them) as interactive visualizations provide the vocabulary for analytical dialogue between the human and the computer when facing wicked problems, in this case giving the human access: i) to enhance and direct information retrieval/fusion as well as modeling and simulation; and ii) to explore their results and other relevant data.

Unifying all three dimensions of the framework are the processes associated with operational planning. The embedded representation of operational planning in our framework suggests that the analytical methods and tools be applied not only to mission data and knowledge, but also to data and knowledge that are both a representation and product of the planning process (indicated by thickening red arrows). In other words, operational planning is an exploratory process that produces both data and knowledge. We contend that there is hidden structure and meaning in these products that can offer valuable insights to the decision making process. For example, these products can reveal: i) the limitations of the employed tools and methods; ii) the provenance of the underlying inputs to the characterization of the operational environment; and iii) the processes (and their embedded intuitions as well as biases) that led to the constructed understanding of the operational environment.

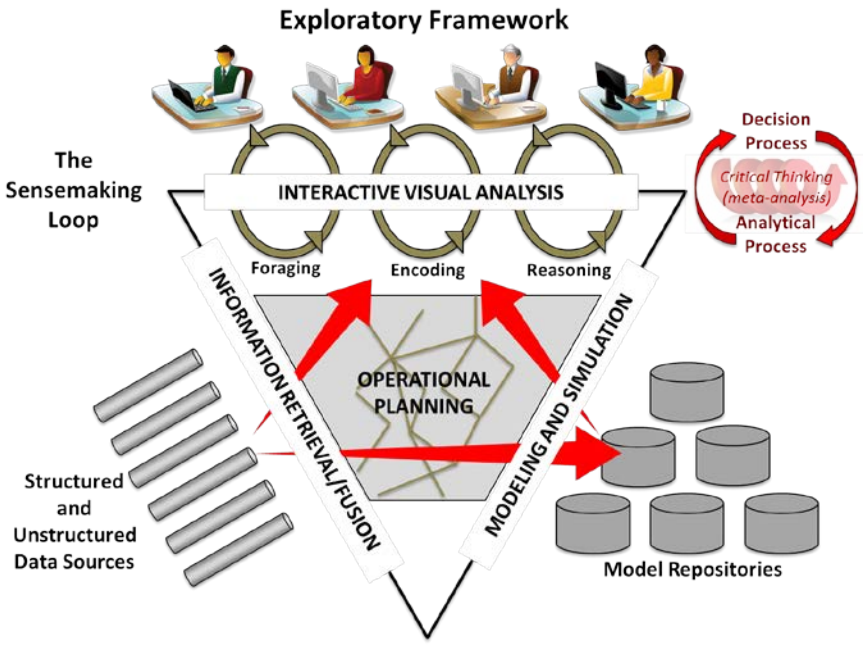


Figure 3 – Exploratory Framework for Interactive Visual Analysis

There are several research gaps suggested by our framework that must be filled in order to provide full benefit to exploratory analysis of wicked problems – in particular, to the process of making sense of operational environments and the factors that influence the outcomes of defense and security operations. The first research gap highlights the challenge of revealing and leveraging hidden structure within large corpora of data as a means of making sense of those data. The second research gap highlights the challenge of understanding the structure and behavior of integrated models, and simulations they produce, as reflections of the systems they represent. With both gaps, we emphasize the value of exploratory visual analysis as an essential affordance to meeting these challenges.

4.0 GAP 1: IDENTIFYING EMERGING TOPICS AND TRENDS

The first gap and application illustrates data-driven methods for identifying emerging topics and trends from structured and unstructured data sources. Digital textual content is being generated at a daunting scale, much larger than we can ever comprehend. Vast amounts of content are accumulated from various sources, diverse populations, and different times and locations. For example, 1.35 million scholarly articles were published in 2006 alone. With an average annual growth rate of 2.5%, research articles are currently being published at the pace of approximately 4400 titles per day. In the social media world, people are contributing to the accumulation at an even faster pace. By June 2012, Twitter is seeing 400 million tweets per day. Meanwhile, 900 million active Facebook users have been busy sending 1 million messages every 20 minutes. It is generally agreed in government and industry that valuable but latent information is hidden in the vast amount of digital textual content, information that can provide insights into proper characterization of operational environments. For instance for emergency response agencies, sifting through massive amount of social media data could help them monitor and track the development of and response to natural disasters, as illustrated in the use of Twitter to reach victims from hurricanes.

With all this activity, a large and growing problem is how analysts and decision makers can gain an understanding of the ideas and connections being expressed in media, and the trends, relationships, events, and social connections indicated or implied in this activity. Once a particular event or activity is identified, one can use search and organizational capabilities to gain more information about it, but this leaves out the majority of events that may be of interest but have not fully bloomed yet and, thus, are not known. In this case, one is overwhelmed with the noise of unrelated events and activity; even in the situation where analysts have an idea what they are looking for, they still are faced with actually examining too large a corpus of data messages in order to get a good understanding of how topics develop, intertwine and change.

To meet these challenges, we have developed interactive visual analytics system that aims to provide sophisticated visual interfaces and verifiable analytics results to augment the analytics capability to detect and validate events from heterogeneous, unstructured data sources. We illustrate our research results through its application to the detection and validation of social events from heterogeneous social media sources. Though a specific example, one can easily see how these results can benefit sensemaking process focused on the operational environment.

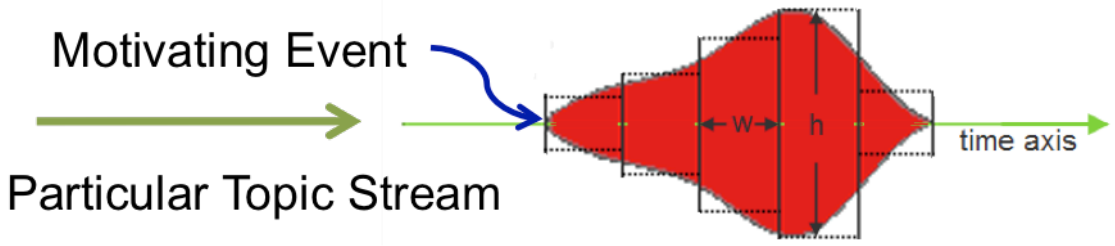


Figure 4--- Bursty Event Structure.
H indicates the volume of documents (e.g., tweets) associated in this event. W suggests the duration of the event.

The fundamental component of our visual analytics system is the **Event**, which we define as a "meaningful occurrence in space and time." Events are bursts of activity over a relatively short time period, the time scale depending on the category of the temporal data. For example, with streaming Twitter data, a typical single event burst lasts one to two days; major events can be longer lasting, but they usually can be divided into sub-events. In this paper, events are associated with a particular topic (as shown in **Figure 4**) so that an event occurs for a

particular topic, time, and set of extracted entities (e.g., location, indicated past or future times, names of people, etc. extracted from the social media texts). Thus, in the case of the interactive interface we have developed for Twitter data, a selection of an event chooses only those tweets for the given topic and for the part of the event burst time range selected. As discussed below, events provide a great focus and together make up an interpretable narrative; thus this selection is a powerful filtering tool.

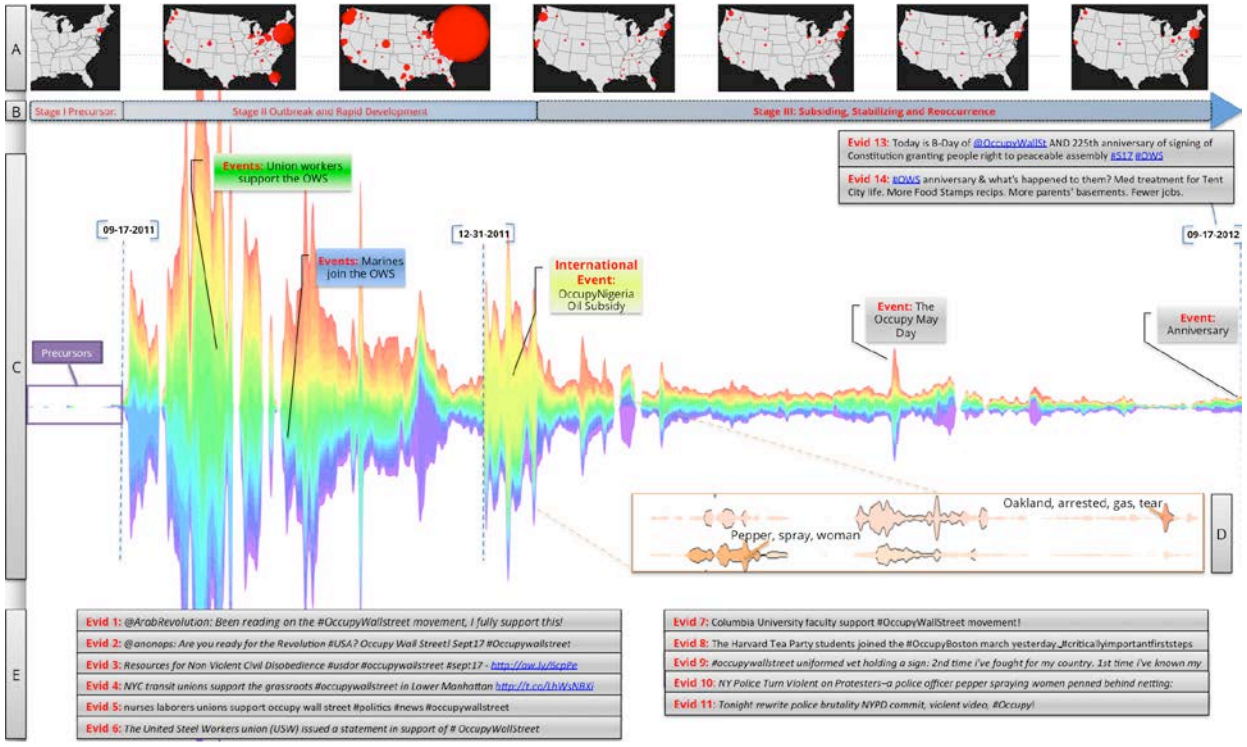


Figure 5 --- Overview of the Occupy Wallstreet Movement (OWM) in our VA System.
A: Occupy hotspots over time. *B*: Three stages of the OWM divided based on the rise and fall of the overall activities. *C*: Visual summary of the Occupy activities. The x-axis represents time; each color-coded ribbon represents a topic extracted from the tweets. Event detection is performed on individual topic to identify bursts as indicators of events. *D*: Sample events labeled with corresponding keywords. *E*: Evidence [Evid.X] mentioned in the paper. For more detail refer to [11].

To identify emerging topics and trends, we perform one more analysis step on our event structure. We label as events only those bursty structures that have a motivating event (see **Figure 4**). A motivating event is an occurrence, either described in the event burst tweets themselves (usually at the beginning) or external to this set of tweets that has motivated the bursty response. Most if not all event bursts of this type are responses to the initial motivating event. For example, the main topical events on September 17, 2011 were clearly associated with the launch of Occupy Wall Street (OWS) on that date at Zucotti Park in New York City, but most of the associated tweets, from individuals and from online news, were in response to this event. In fact, OWS was large enough that there were several topics with their associated events on that date.

We have found that by just analyzing the shape, size, and duration of the burst, we can automatically identify events that will have clear motivating events. Thus, we have a mechanism for automatically identifying meaningful events that we have tested successfully on multiple categories of data, not just streaming social media. This is not to say that there are no other, unmarked bursts that are meaningful. Nor is it to say that the meaning is immediately clear from this analysis. Input from a human-in-the-loop is necessary to resolve these

questions. But this identification of meaningful events is still a boon for exploratory analysis since we have found it identifies most of the major events and also directs the user's attention. We have applied our visual analytics system to tell the complete story of OWS from precursor discussion before the launch till now. As shown in **Figure 5**, this shows how a comprehensive, rich narrative can be built efficiently [11].

5.0 GAP 2: MAKING SENSE OF INTEGRATED MODELING AND SIMULATION

The second research gap highlights the challenge of understanding the structure and behavior of integrated models, and the simulations they produce, as a means for making sense of the operational environment. As defined by U.S. military doctrine, the operational environment is a "composite of the conditions, circumstances, and influences that affect the employment of military forces and bear on the decisions of the unit commander" [4]. These "conditions, circumstances, and influences" are understood in terms of operational variables (e.g., PMESII-PT) that characterize the operational environment. Integrated modeling and simulation offers a promising approach to this sensemaking challenge as models representing each of the operational variables are composed, or coupled, to represent an interdependent system of systems within the operational environment.

To illustrate, consider the *Vu* architecture depicted in Figure 6. A product of our previous research results [12],[13],[15], *Vu* is a knowledge-driven approach to integrated modeling and simulation of complex systems of systems. As a knowledge-driven approach, the architecture supports the adaptation, or reframing, of models to produce an integrated representation of specified operational environment. It also enables the exploration of alternative hypotheses, a useful strategy to managing wicked problems (c.f., Roberts [2]).

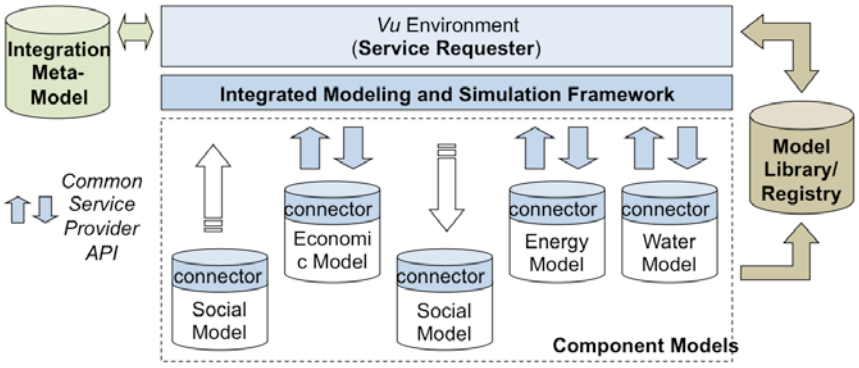


Figure 6 – Architecture for Integrated Modeling and Simulation

While the architecture [14] presents a novel approach to integrated modeling and simulation, its utility depends directly on the essential role of exploratory visual analysis. The *Vu* user experience (see Figure 7) consists of temporally situated, geospatially-oriented, interactive visualizations of interdependent models as they respond to disruption or reconstitution events with cascading effects. More specifically, the *Vu* environment provides numerous interactive visualizations that enable users to identify and understand quickly: i) emerging, cascading effects; ii) chains of causality and their underlying interdependencies; and iii) plausible futures of potential courses of action. Furthermore, the exploratory interfaces help users to identify key events in sequence in order to turn these identified events, their behaviors and relationships, into actionable plans or courses of action, augmented with alternative scenarios. The interactive visualizations also play an essential function in model verification and validation by increasing transparency of model projections.

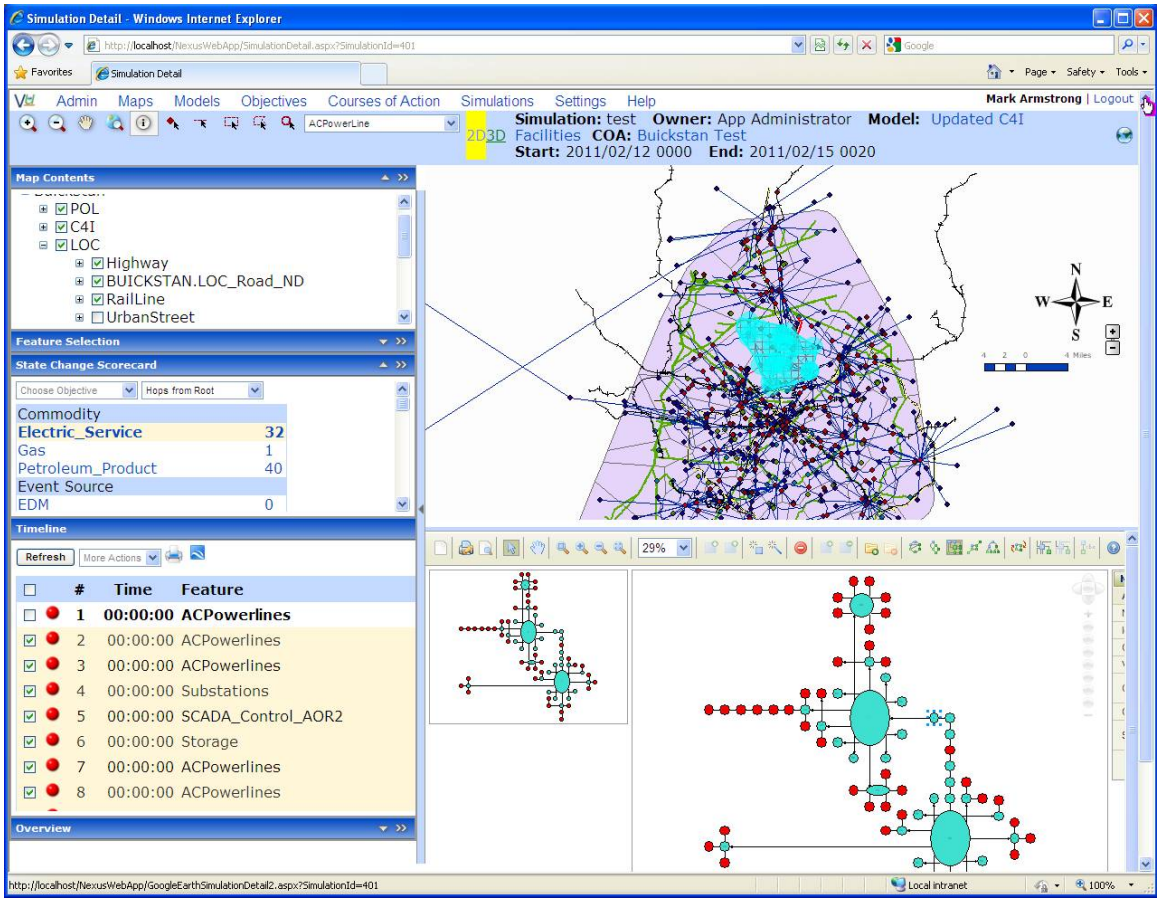


Figure 7 – The *Vu* User Experience

The second gap and application also illustrate data-driven and knowledge-driven methods working in concert to understand better the properties of knowledge representations. More specifically, the planning process produces numerous knowledge artifacts including models of the operational environment, potential courses of action and simulation traces representing plausible futures. The application of data-driven methods supported by exploratory interfaces to these knowledge artifacts can reveal important insights about the representations themselves and the complex space of simulations they produce. Figure 8 offers two illustrations of this affordance. Both allow visual exploration of a suite of *Vu* simulations produced using data farming techniques. The visualization on the left is aggregate sustainability analysis of simulation outcomes for an integrated urban model. The visualization on the right is aggregate temporal analysis of integrated infrastructure resiliency in the face of majorly disruptive weather events (e.g., hurricanes). Each visualization reveals hidden structure within the integrated modeling and simulation data – thus, providing further insight about the respective operational environment.

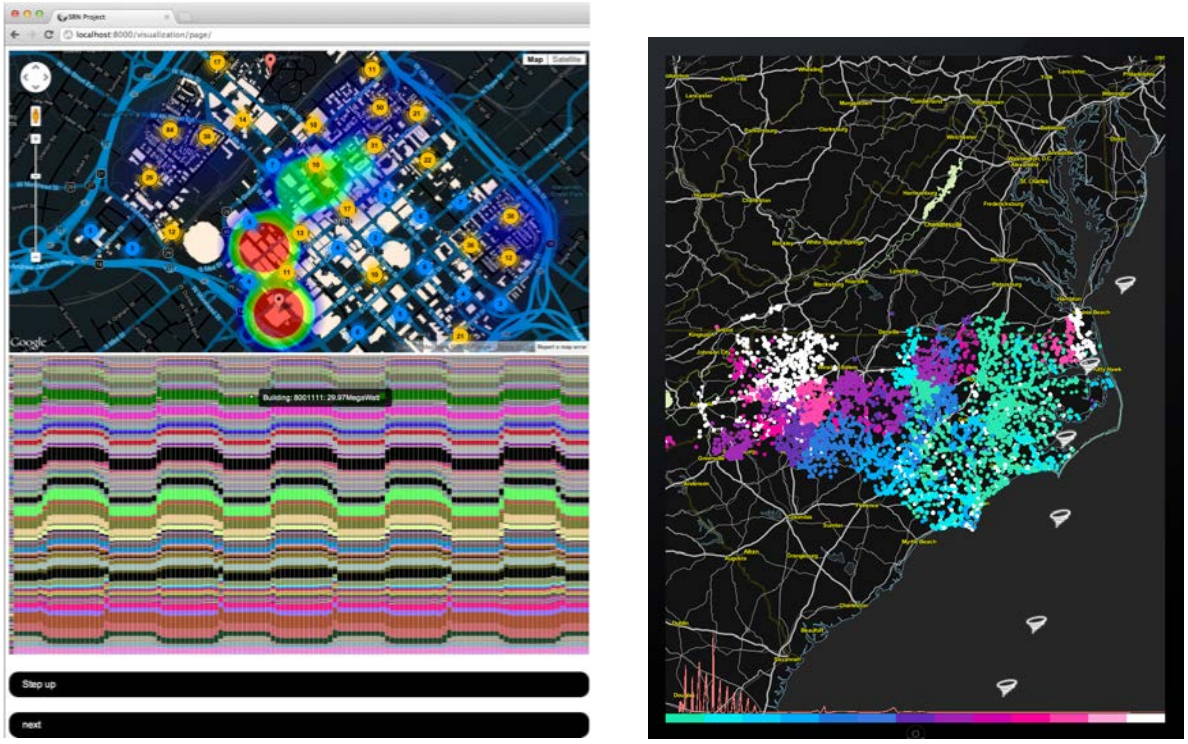


Figure 8 – Visual Analysis of the Space of Simulations.

6.0 CLOSING THE LOOP

We conclude this paper by connecting framework dimensions – i.e., information retrieval/fusion, interactive visualizations and modeling and simulation – to the sensemaking process in general, and the process of making sense of the operational environment more specifically.

As demonstrated by our discussion of the illustrative gaps, exploratory visual analysis can benefit all aspects of the sensemaking process – i.e., foraging, encoding and reasoning. In particular, exploratory visual analysis can facilitate the efficiency and efficacy of the foraging loop, including the organization of collected data and the identifications of key evidence. As illustrated by Gap 1, data-driven tools and methods offer important utility to this process, as there is often hidden structure and meaning within collected data. Connecting exploratory visual analysis capabilities to these data-driven tools and methods also enables the encoding loop as evidences are organized into relevant representations. These activities become the initial, though powerful, forays into the reasoning loop.

As illustrated by Gap 2, extending exploratory visual analysis capabilities to knowledge-driven tools and methods can further facilitate the reasoning loop. Exploratory visual analysis enables users not only to understand better the modeled phenomenon, but also the limitations of the models as reflected in identified uncertainties, biases, inaccuracies and/or missing evidences. Exploratory visual analysis also can help users understand better the range of potential representations and plausible futures in the face of recognized uncertainty. Paul and Elder stress the importance of such exploration in their discussion of the elements of reason and the articulation of assumptions, evidences, inferences and consequences [16]. Exploring the range of plausible futures (rather than just the predicted future) is especially important in the face of wicked problems,

such as making sense of the operational environment. In particular, Roberts highlights the value of competing analyses as a promising method for managing wicked problems [2].

Connecting data-driven approaches with knowledge-driven approaches, however, can further empower the sensemaking process – not only through collective support but also through synergy between the approaches. In particular, data-driven approaches (associated with information retrieval/fusion) can be connected to knowledge-driven approaches (associated with modeling and simulation) through the encoding loop. Here, military doctrine regarding the desired characterization of the operational environment (e.g., PMESII-PT) frames the encoding process. Exploratory visual interfaces are essential to this activity as they help modelers identify which evidences to encode and which representations to employ. As illustrated in our discussion of Gap 2, these capabilities in particular can help modelers develop new models, adapt existing models and configure models for exploration and analysis.

7.0 SUMMARY

Making sense of the operational environment is arguably a wicked problem. We believe that interactive exploratory visual analysis can offer important affordances to the sensemaking process in this context. To that end, this paper presents a framework for interactive, exploratory visual analysis. This framework is grounded in the sensemaking process and characterized by three essential dimensions: information retrieval/fusion, interactive visualizations and modeling and simulation. The utility of the framework is found in the direction it offers to tool and method research and development in relation to the ability to characterize operational environments quickly and accurately – relative to mission requirements. Furthermore, we believe that the volume, velocity and variety of data that describe the operational environment as well as the complexity of the systems they represent necessitate both knowledge-driven and data-driven approaches to analysis. To support our contentions, we present to research gaps and associated applications. The first gap and associated application illustrates the benefit of exploratory visual analysis to a data-driven approach to the identification of emerging topics and trends within large corpora of unstructured data. Such support can offer important affordances to characterizing operational environments. The second gap and associated application illustrates the benefit of exploratory visual analysis to integrated modeling and simulation. Integrated modeling and simulation can provide multi-dimensional (e.g., PMESII-PT) representations of the operational environment in support of the sensemaking process. The second gap and application also illustrate data-driven and knowledge-driven methods working in concert to understand better the properties of these integrated representations.

8.0 REFERENCES

[1] H. Rittel, M. Webber. “Dilemmas in a General Theory of Planning,” *Policy Sciences* 4: 155-159, 1973.

[2] N. Roberts. “Wicked Problems and Network Approaches to Resolution,” *The International Public Management Review*, 1.1: 1-19, 2000.

[3] P. Pirolli, S.K. Card. “The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis, *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[4] Joint Publication 2-01.3. *Joint Intelligence Preparation of the Operational Environment*, June 2009.

[5] Field Manual 3-0. *Operations*. Department of the Army, February 2008.

[6] M. Duffy. "Sensemaking in Classroom Conversations," *Openness in Research: The Tension between Self and Other*, ed. by I. Maso et al.,Van Gorcu, 1995.

[7] D.K. Leedom, *Sensemaking Symposium: Final Report*, Command and Control Research Program, Office of the Assistant Secretary of Defense for Command, Control, Communications and Intelligence, 2001.

[8] G. Klein, B. Moon, R.R. Hoffman. "Making Sense of Sensemaking 1: Alternative Perspectives," *IEEE Intelligent Systems*, July/August 2006.

[9] D.M. Russell, M.J. Stefik, P. Pirolli, S.K. Card. "The Cost Structure of Sensemaking," *Proceedings of InterCHI*, 1993.

[10] J. Locke. *An Essay Concerning Human Understanding*, 1690.

[11] X. Wang, W. Dou, Z. Ma, W. Ribarsky. "Discover Diamonds-in-the-rough Using Interactive Visual Analytics System: Tweets as a Collective Diary of the Occupy Movement." In the 7th *International AAAI Conference on Weblogs and Social Media 2013 (ICWSM 2013)*.

[12] W.J. Tolone. "Making Sense of Complex Systems through Integrated Modeling and Simulation." *Advances in Information and Intelligent Systems*. Ras, Z. and Ribarsky, W., eds., pp. 21-40, *Studies in Computational Intelligence*, Volume 251, 2009.

[13] W.J. Tolone, M. Armstrong. "Integrated Analytics: Understanding Critical Infrastructure Behaviors for Resilience Analysis." *The Homeland Security Review*. Vol. 5(3): 241-258, California University of Pennsylvania, 2011.

[14] W.J. Tolone, E.W. Johnson, S.W. Lee, W.N. Xiang, L. Marsh, C. Yeager, J. Blackwell. "Enabling System of Systems Analysis of Critical Infrastructure Behaviors." In *Proceedings of the 3rd International Workshop on Critical Information Infrastructure Security (CRITIS 2008). Lecture Notes in Computer Science #5508*, Setola, R., Geretshuber, S., eds., pp. 24 – 35, Springer, 2009.

[15] W.J. Tolone, S.W. Lee, W.N. Xiang, J. Blackwell, C. Yeager, A. Schumpert, E.W. Johnson. "An Integrated Methodology for Critical Infrastructure Modeling and Simulation." In IFIP International Federation for Information Processing, Volume 290; *Critical Infrastructure Protection II*, 2nd Edition, Goetz, E. and Shenoi, S. eds., pp. 257 – 268, Springer, 2008.

[16] R. Paul, L. Elder. *The Miniature Guide to Critical Thinking: Concepts and Tools*. The Foundation for Critical Thinking, 2006.

Social Media Analytics for Competitive Advantage

William Ribarsky Wenwen Dou
Xiaoyu Wang

Abstract

Big Data Analytics is getting a great deal of attention in the business and government communities. If it lives up to its name, visual analytics will be a prime path by which visualization competes successfully in this arena. This paper discusses some fundamental work we have done in this area through integration of interactive visualization and automated analysis methods and the applications that have resulted.

1. Introduction

As we know, big data analytics has become more than a term, it is now a movement. Whether or not "big data" is a buzzword or not, there is reason to believe that interest in the broader issues surrounding not just scalably large data but even more so data that are complex and can be associated with problems that require complex reasoning will not abate soon and could even grow stronger. There are many reasons for this. More companies (and government agencies, too) are building comprehensive and long-term databases. But there is a growing realization that traditional database techniques, though they are in principle scalable and useful for many things, can't tell you basic things about what the database contains and what the important relations and trends are [1]. (For example, we have worked with a bank that has a very large, comprehensive transactional database yet struggles to use all the value it contains.) In addition, the advent of social media and online sources show that useful data can come from anywhere, inside or outside the company. Due to the availability of all these data, there is a growing push to increase data-driven decision-making, but this is hampered, as inferred above, by not knowing what actionable information the data contains. Finally, business studies indicate that timely, effective use of data-derived knowledge is a competitive advantage and that not using this knowledge effectively is a competitive disadvantage [2-4]. Companies who do not marshal their data resources will be losers in the long term. In fact, those that find new uses for their or other data will be the biggest winners.

Another reason that big data analytics will have staying power is that a robust infrastructure is being built. Based on surveys in the U.S. alone, McKinsey estimates that there will be a deficit of nearly 200K professionals with deep analytics skills by 2018 and the need to retrain 1.4M managers so that they understand the value of data and know the right questions to ask [2]. Gartner estimates that 1.9M big data jobs will be created in the U.S. by 2015 [4]. It will be hard to fill even a fraction of this need at the current rate of production for data analytics professionals.

The premise of this paper is that visualization and especially visual analytics is ideally positioned to take advantage of this opportunity. By its nature, visual analytics supports exploration, discovery, and complex reasoning about data and data-driven problems. Statistical, data mining, machine learning, signal processing, and other deep analytics methods are tightly integrated with interactive visualizations. Visual analytics aims to put the human in the loop at just the right point to discover key insights, develop deep understanding, make decisions, and take effective action. On both the visualization side and the analytics side, visual analytics is positioned to effectively manage and support understanding of scalably large data, and this promise is being made concrete as new techniques are developed [7]. Because of the central role of visual analytics, there is also the opportunity to imbue the massive influx of new data analytics professionals with knowledge of visual analytics through courses and training. This should definitely be a key part of the visual analytics academic agenda; it will result in a whole generation of analysts, engineers, and managers who appreciate and know how to use these tools.

In this paper, we will address how competitive advantage can be derived from the analysis of unstructured data, especially social media data

(though the techniques used can be applied to a broader range of unstructured data). We will use mostly examples from our own work, though there is a range of other work.

2. The Nature of the Data

We focus on streaming Twitter data in this paper. We have been collecting a 1% random sample of these data for nearly 1.75 years. This results in a large number of tweets (now in excess of 20B). We have used this Twitter collection as a testbed for several recent studies with diverse subjects [7-9] including the ones reported here. We are now building a collection of texts from Facebook posts, which will permit us to explore a different demographic range than that for Facebook.

Text messages from Twitter, Facebook, and several other social media services have general attributes such as unstructured content and intrinsic uncertainty as to the validity of the messages. In addition, these data have the attributes of *data physicality* and *data sociality*. The messages are often intimately connected with particular times and locations (either locating where and when the message was sent or by mention of places and dates, either past, present, or future in the message bodyOf course, social medial messages are sent, received, or re-sent by people, so there can be a rich social connectivity revealed. The availability of such information, often minute-by-minute or over the whole length of a story that may take months to unfold, is a new and very powerful aspect of social media analyses.

3. Topic Modeling and Entity Extraction

To provide meaning and organization to the unstructured data, we use Latent Dirichlet Allocation (LDA) [5,6], which reveals latent topics from large text collections, which are then described by coherent sets of keywords (with the leading words being the most meaningful for the topic, as illustrated in Figure 2). To this we add named entity recognition, based on a customized dictionary and the use of LingPipe with statistical chunking. This permits the identification of people, locations, buildings, times, dates, etc. from within the text messages. We have extended the traditional LDA approach to handle temporal features and structure (in particular events, as described further below). We have also developed scalable capabilities for efficiently generating topics even for very large text collections [7]. We have successfully applied these techniques to a range of text collections including project abstracts, reports, research papers, streaming Twitter data, and recently patent descriptions. This set of approaches provides the ability to attack unstructured data both inside and outside the company, conferring competitive advantage [10].

4. Events and Time Structuring

The fundamental component of our time structuring is the *event*, which we define as a "meaningful occurrence in space and time". Events are bursts of activity over a relatively short time period, the time scale depending on the category of the temporal data. For example, with streaming Twitter data, a typical single event burst lasts one to two days; major events can be longer lasting, but they usually can be divided into sub-events. In this paper events are associated with a particular topic (as shown in Figure 1) so that an event occurs for a particular topic, time, and set of extracted entities (e.g., location, indicated past or future times, names of people, etc. extracted from the social media texts). Thus in the case of the interactive interface we have developed for Twitter data, a selection of an event chooses only those tweets for the given topic and for the part of the event burst time range selected. As discussed below, events provide a great focus and together make up an interpretable narrative; thus this selection is a powerful filtering tool.

We perform one more analysis step on our event structure. We label as events only those bursty structures that have a *motivating event* (see Figure 1). A motivating event is an occurrence, either described in the event burst tweets themselves (usually at the beginning) or external to this set of tweets, that has motivated the bursty response. Most if not all event bursts of this type are responses to the initial motivating event. For example, the main topical events on September 17, 2011 were clearly associated with the launch of Occupy Wall Street (OWS) on that date at Zucotti Park in New York City, but most of the associated tweets, from individuals and from online news, were in response to this event. In fact, OWS was large enough that there were several topics with their associated events on that date. We have found that by just analyzing the

shape, size, and duration of the burst, we can automatically identify events that will have clear motivating events [11]. (These are the bursts with dark outlines in Figure 2 and subsequent figures.) Thus we have a mechanism for automatically identifying meaningful events that we have tested successfully on multiple categories of data, not just streaming social media. This is not to say that there are not other, unmarked bursts that are meaningful. Nor is it to say that the meaning is immediately clear from this analysis. Input from a human-in-the-loop is necessary to resolve these questions. But this identification of meaningful events is still a boon for exploratory analysis since we have found it identifies most of the major events and also directs the user's attention. We have applied all the techniques in Secs. 3 and 4 to tell the complete story of OWS from precursor discussion before the launch till now. This shows how a comprehensive, rich narrative can be built efficiently [9].

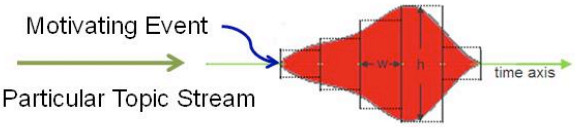


Figure 1. Bursty structure for an event.

5. Deriving Competitive Advantage

In the rest of this paper, we discuss some business cases that show how competitive advantage might be derived.
We first did a study of department stores in the Charlotte region (Belk, Macy's, Dillards, Nieman Marcus, and Saks Fifth Avenue) by using messages with hash tags or entities naming the stores and then bringing in related tweets through topic modeling. The study was over a 10 month period starting in Fall, 2011. About 15 main topics resulted for this time period. They immediately revealed some general information. Twitter response often had to do with marketing around celebrations, charitable campaigns, and holiday events (e.g., Macy's Thanksgivings Day parade). There were often tie-ins to women's cosmetics and beauty products. Macy's had the largest Twitter presence during the period followed by Nieman Marcus. Belk had a growing presence later in the period. Dillards never had much of a presence.

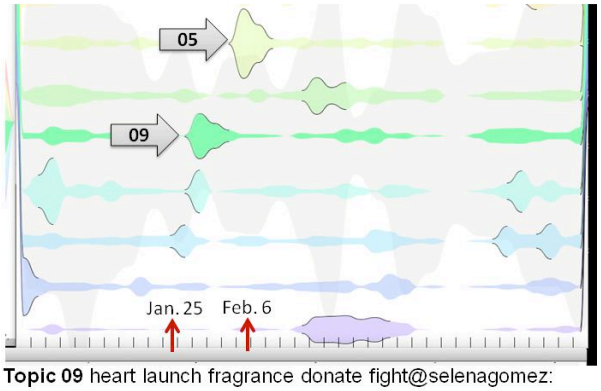


Figure 2. Event view of Macy's women's heart donation campaign (Topic 09) and Belk/ Nieman Marcus beauty products promotion (Topic 09).

Figures 2 and 3 show specific events illustrating how detailed analyses over time can be done (in this example restricted to a 3 month period at the beginning of 2012). The presented events plus other events that fill in the Twitter story for these stores were generated automatically and then quickly studied in more detail by selecting associated tweets for each topic and event. Figure 2 shows response to a marketing effort associated with the gored for women heart campaign sponsored by Macy's (Topic 09)It also shows merged responses to two similar marketing efforts launched by Nieman Marcus and Belk on beauty products that started during Super Bowl week and then continued for another week or two. In both cases, the events can easily be identified by looking at the lead words in the topic lists and some of the associated tweets.

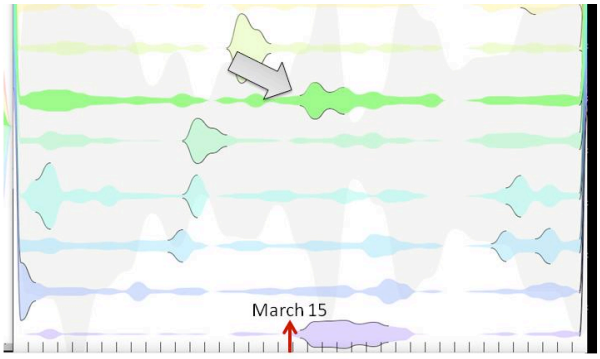


Figure 3. Burst of negative responses to Rush Limbaugh's comments (associated with Macy online).

Figure 3 shows another type of event with a different temporal structure. The event is

conservative talk show host Rush Limbaugh's diatribe against a Georgetown University student because of her stand in favor of access to birth control, and the response to his comments. The motivating event caused a firestorm of comments against Limbaugh including a burst of activity having to do with the StopRush Twitter movement and its campaign to boycott sponsors of Limbaugh's show. Macy's became embroiled because of its sponsorship. Although the event has to do with the specific outburst about the student, other bursts and events in the same topic stream show other public complaints against Limbaugh both before and after the event in Figure 3. This stream of related activity goes over a period of months and shows that negative events and associations, even if inadvertent, can have a long effect.

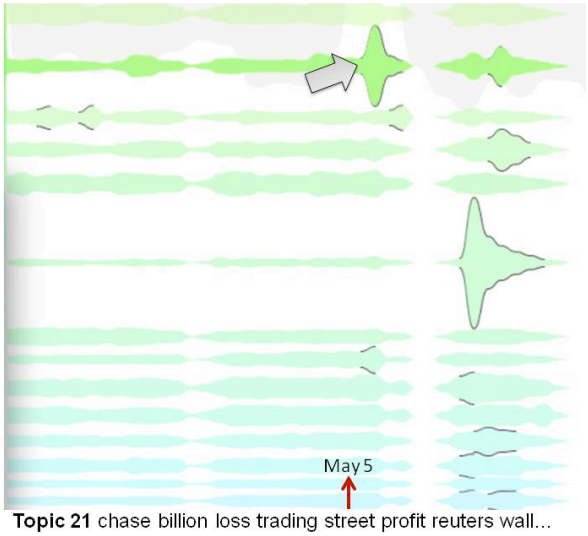


Figure 4 Event associated with the revelation of JP Morgan Chase multibillion dollar trading loss (the London Whale).

Figures 4 and 5 show details from a set of topical analyses having to do with several banks, including those headquartered in the Charlotte region (such as Bank of America), over a period of several months. The topic modeling analysis was set up in a similar fashion as the department store analysis; the set of topics is larger. In Figure 4 the indicated event was motivated by the initial exposure of the multibillion dollar trading loss at JP Morgan Chase. The sharp burst begins at the day of this disclosure; there was also an associated stock market drop on that day. In addition, there is a series of events for the same topic before and especially after the disclosure that tell the unfolding story of this scandal.

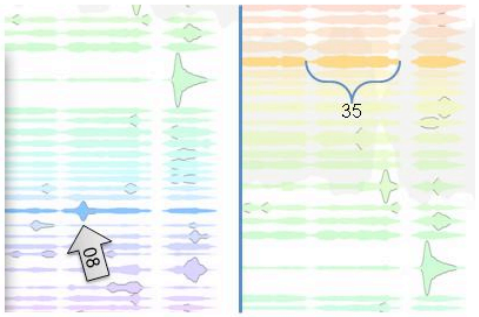


Figure 5 Banking and mobile app credit card surge (Topic 08) and bank new talent acquisition (Topic 35).

Although most of the events in the banking analysis have to do with disclosures like this, ongoing fallout from the mortgage and banking collapse of 2008, and response to proposed government regulations, there are still events and topics that discuss other aspects. Figure 5 gives a couple of examples. Topic 08 has to do with the development of new mobile debit and credit card apps. The event indicated is response to actions by Walmart and Target, among others, and reported in the Wall Street Journal and Forbes, to develop their own mobile payment systems. This would eliminate the middle men (the banks) and would be of great interest and possible risk to them. These analyses show that even in this case where the events are dominated by negative news about banks, the exploratory visual analytics tools still find events that tell banks about their competitive environment.

6. Conclusions

We have presented the work described here plus addition analyses to a set of business partners from retail and banking. This has generated considerable interest and feedback. The ability to analyze competitor strategies is considered important. There is a desire to know about the demographics of the people generating messages for selected events and also how retweeting spreads a message. (We are working on both these things.) There is a desire to do targeted marketing based on real-time streaming tweet analysis, and we have developed a capability in this area. More generally, companies see the opportunity to analyze the response to marketing and advertising campaigns as they unfold and also investigate what affects the public view of

their brand image (which can be affected by external circumstances, as we have seen). Banks want to do emerging risk analysis using both internal and external sources. We also expect that the set of methods described in this paper will be applied in the future to internal company data.

7. References

[1] Andrew Oliver. Big data woes: Which database should I use? InfoWorld (August 3, 2012).

[2] J. Manyika, M. Chui, B. Brown, et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute (May, 2011).

[3] S. LaValle, E. Lesser, et al. Big Data, Analytics and the Path From Insights to Value. MIT Sloan Management Review Vol. 52 (2) (Winter, 2011).

[4] Patrick Thibodeau. Big Data to Create 1.9M IT Jobs in U.S. by 2015. Computerworld (August, 2012).

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. "*Latent dirichlet allocation*". J. Mach. Learn. Res., 3:993–1022, March 2003

[6] Wenwen Dou, Remco Chang, Xiaoyu Wang, and William Ribarsky. ParallelTopics: A Probabilistic Approach to Exploring Document Collections. Accepted for publication, IEEE VAST 2011.

[7] Wang, X., Dou, W., Ma, Z., Villalobos, J., Chen, Y., Kraft, T. and Ribarsky, W. I-SI: Scalable architecture of analyzing latent topical-level information from social media data. *EuroVis 2012 Computer Graphics Forum*.

[8] Xiaoyu Wang, Wenwen Dou, Zhiqiang Ma, Li Yu, and William Ribarsky. Hierarchical Topics: Visually Exploring Large Text Collections Using Topic Hierarchies. Preliminary acceptance for publication, IEEE VAST 2013.

[9] Xiaoyu Wang, Zhiqiang Ma, Wenwen Dou, and William Ribarsky. Discover Diamonds-in-the-Rough using Interactive Visual Analytics System: Tweets as a Collective Diary of the Occupy Movement. Accepted for publication. IEEE SocMedVis 2013.

[10] R. Nair and A. Narayanan. Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage. Booz & Company (2012). www.booz.com/media/file/BoozCo_Benefitting-from-Big-Data.pdf

[11] D. Luo, J. Yang, M. Krstajic, J. Fan, W. Ribarsky, and D. Keim. EventRiver: Interactive visual exploration of constantly evolving text collections. *IEEE Trans. On Visualization and Computer Graphics* (October, 2010), doi.ieeecomputersociety.org/10.1109/TVCG.2010.225.

Towards a Visual Analytics Framework for Handling Complex Business Processes

William Ribarsky
Department of Computer Science
University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC, 28262
ribarsky@uncc.edu

Derek Xiaoyu Wang
Department of Computer Science
University of North Carolina at Charlotte
xiaoyu.wang@uncc.edu

Wenwen Dou
Department of Computer Science
University of North Carolina at Charlotte
wdou1@uncc.edu

William J. Tolone
Department of Software and Information Systems
University of North Carolina at Charlotte
William.Tolone@uncc.edu

Abstract

Organizing data that can come from anywhere in the complex business process in a variety of types is a challenging task. To tackle the challenge, we introduce the concepts of virtual sensors and process events. In addition, a visual interface is presented in this paper to aid deploying the virtual sensors and analyzing process events information. The virtual sensors permit collection from the streams of data at any point in the process and transmission of the data in a form ready to be analyzed by the central analytics engine. Process events provide a uniform expression of data of different types in a form that can be automatically prioritized and that is readily meaningful to the users. Through the visual interface, the user can place the virtual sensors, interact with and group the process events, and delve into the details of the process at any point. The visual interface provides a multiview investigative environment for sensemaking and decisive action by the user.

1. Introduction

The idea of "just in time (JIT) manufacturing" has been around for some time and has been implemented in large scale production environments. Lately there has been the need for dynamic just in time manufacturing where the distribution and even types of products produced may change fairly quickly, often in response to a previously unforeseen need. This is the case for some large government entities that must supply materiel in support of changing missions and for some large manufacturers.

In this paper we describe a framework we have developed in cooperation with a partner engaged in dynamic manufacturing. A major issue is that the business process (in this case manufacture and delivery of a variety of systems, which could be complex themselves or components of even larger systems) is complex and changing. It can be across several parts of a large manufacturing and delivery organization and across many different suppliers. A disruption at any point in the organization or among the suppliers could result in a failure to deliver full capabilities on time, which would be a serious problem.

A main issue that can arise is shown in the following example, illustrated in Figure 1. Suppose the assembly of a system that must be deployed by a certain date is dependent on the availability of a component that is provided by an external supplier. Further the component is not installed till late in the assembly process. Since this is a dynamic manufacturing process where different systems with different externally supplied components are required to be produced at different times, the organizational structure cannot be closely aligned with the manufacturing process. In this case the procurement and assembly components are in different part of the organizational structure with different reporting paths. In the illustration, a clue to the problem is given by unusual communication being opened up between the parts acquisition/preparation department and the quality assurance manager. (This isn't the only possible path; the communication between the production supervisor and the technical manager could possibly be due to the same or a related problem.) There may also be unusual activity within a department even without external communication. In Figure 1, this is picked up by a "virtual sensor" (dis-

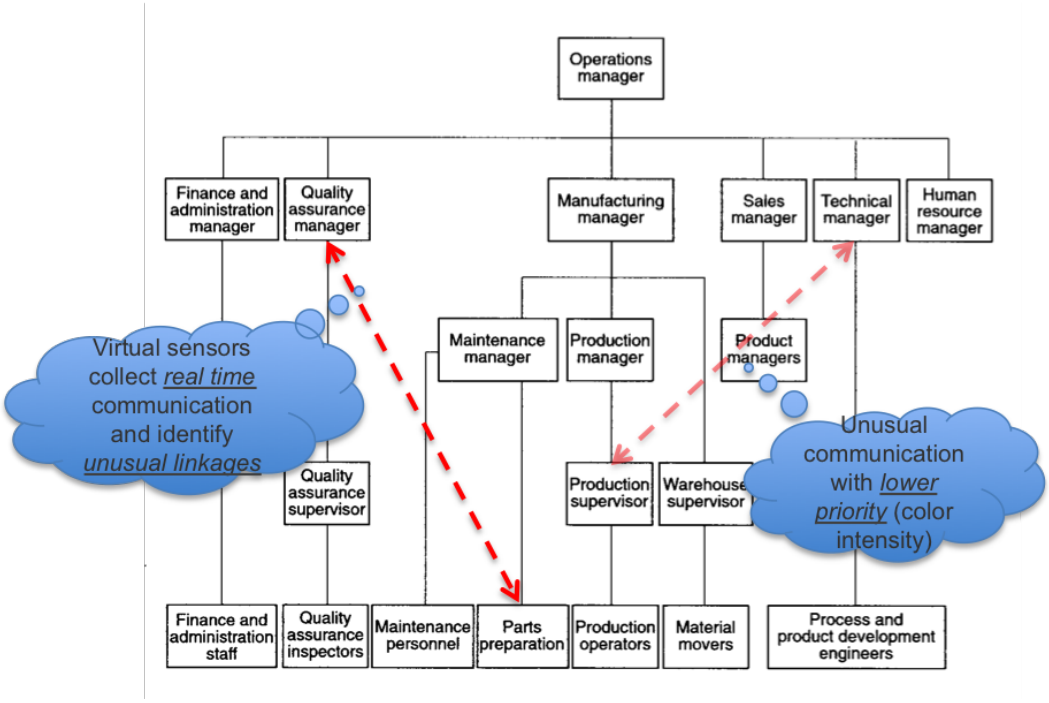


Figure 1. An organization structure for a production operation with unusual communication paths.

cussed further below and in Sec. 5.) Unless there is appropriate and timely communication across the structure and in particular to the operations manager (and above), the manufacturing process will break down. This is to be contrasted with a non-dynamic JIT process. In this case, the steps of the process are more fixed, permitting better organizational alignment from the start and better communication. Furthermore, non-dynamic JIT permits predictive planning where a forecast model based on past production and other factors is imposed to predict inventory needs, among other needs [15]. Such predictive methods will be much harder to develop in dynamic JIT manufacturing. Finally, most JIT planning approaches have been applied to much more uniform manufacturing processes (e.g., the production of a certain car model) than are the case here, where quite different systems are produced at different times.

The main contribution of this paper is the development of a framework to attack the above problem by applying a visual analytics approach. Although there has been work in process-aware information systems (see Related Work), these approaches haven't been applied to manufacturing processes nor, more importantly, are they as comprehensively data-driven as our approach. To enable our approach, we have introduced two new concepts: virtual sensors and process events. Virtual sensors are software collector/transmitters that can be placed at will anywhere in the data flow of the manufacturing or production process. They are meant to deal with a range of data from quantitative process information to communications among operators or managers. Since the virtual sensors monitor the stream of data over time, the visual analytics framework can identify process events, meaningful occurrences in time with respect to the process. We have shown how general event-based approaches can organize complex data histories in a semi-automatic way [14, 19]. What this means is that the visual analytics analysis can reveal the data process and through it the underlying business process. If, for example, the data process does not conform with the organizational model of the manufacturing business, this can signal problems that should be addressed.

2. Related Work

There is a history of business process visualization in the business industry [3, 13, 16, 20]. In the area of customer relations management, Azvine et al. presented a configurable business process analytics tool that constantly monitors the performance of a decision model on both overall and individual levels with the goals of customer satisfaction and operational excellence [1]. Noting that existing business analytic applications are usually closed-loop decision making systems which only present output to operational managers, Azvine et al. incorporated visualization components into an intelligent business analytics system in order to put users in the loop and adjust business operations based on users'

analyses in real time.

In the domain of Process-Aware Information System (PAIS), which pertains to both administrative processes and cross-organizational processes, visualization techniques have been used to aid in the understanding of process schemas and their run-time behavior through simulation [7] and process mining [11]. For example, to improve business process models, Alast et al. combined process mining with visual analytics to incorporate human judgment into otherwise static business process models [11]. On the one hand, process mining supports automatic discovery of a business model and checking how a process model conforms to actual process executions. On the other hand, visual analytics combines automated analysis with interactive visualizations to allow decision-makers to apply their flexibility, creativity, and domain knowledge to come to an effective understanding of situations in the context of large data sets. The authors noted that insights obtained from visualizations could be used to improve processes by removing inefficiencies and addressing non-compliance. In addition to process mining, visualization techniques have also been applied to visualizing work items for an overall business process [12]. Leoni et al. incorporated visualizations to support work assignment in process-aware information systems. The visualizations enable users to better select work items to ensure the performance of the overall business process [12].

3. Probing Dynamic Business Processes

In collaboration with our partner, we have developed the following set of questions that must be effectively addressed in order to achieve a successful dynamic business process:

- What are overall trends? Is the overall process on schedule?
- What are the detailed trends for individual programs within the manufacturing and production process? (A program is the production of a particular system, usually distinct from other systems that may also be produced during the production process.)
- What?s the cost-schedule performance over time?
- What is the capacity of each part of the operation, including excess capacity?
- Is something going wrong and where (including things that have not been fully recognized yet)?
- Is resource re-allocation needed, and where?
- If there is a problem, what expertise should be deployed to most quickly solve it?

To have any chance of answering these questions, the production executives must have continuously available the latest information on the manufacturing process at the program and component levels. To this end, our partner has developed a "dashboard" that accepts data from all stages of the process for all programs. The dashboard system has both an interface and a data collector that provides a summary analysis to determine where the process, at any stage, stands with respect to a few simple benchmarks. Importantly, production managers can provide annotations and comments at any point in the process. They can also communicate with each other or the executives via the dashboard. The dashboard displays comparative information on the production process for all components and programs. It shows some information on trends and when a component is falling behind schedule according to a simple milestone analysis. It has some drill-down capability so that the executive can get more information about why a trend is occurring.

Although this sort of analysis is necessary, our partner has found it is not sufficient for enabling a dynamic business process without interruptions or missed deadlines. In particular, the dashboard does not have a way to effectively organize and use all the information it collects, especially the annotations and comments that are critical to understanding the production process. In addition, disruptions in different parts of the organization may affect downstream production in unexpected ways. Further, there may be no direct line of communication between the parts of the organization that will be affected (although there could be informal communication). Neither of these aspects are readily apparent to the manager in the dashboard setup. Finally, since unexpected disruptions will arise, it is only possible to know of them after the fact and often with incomplete information as to cause. Thus it is difficult to put into place predictive or even prospective monitoring and response, nor to pursue some of the questions above such as those having to do with re-allocating resources and expertise as an event is unfolding. In the next sections, we discuss how visual analytics can be applied to address these issues and produce more successful outcomes for dynamic business processes.

4. Visual Analytics Framework

Visual analytics is the science of analytic reasoning facilitated by interactive visual interfaces [9]. It is meant to support exploratory analysis leading to discoveries since it is frequently applied to complex real world problems with large amounts of data. These problems are often open-ended with no clear path to solutions [9], and, as a result, it is often unclear what pertinent knowledge the data may contain [9]. Thus the analysis needs to be exploratory to support discovery of hidden relations, patterns, and trends.

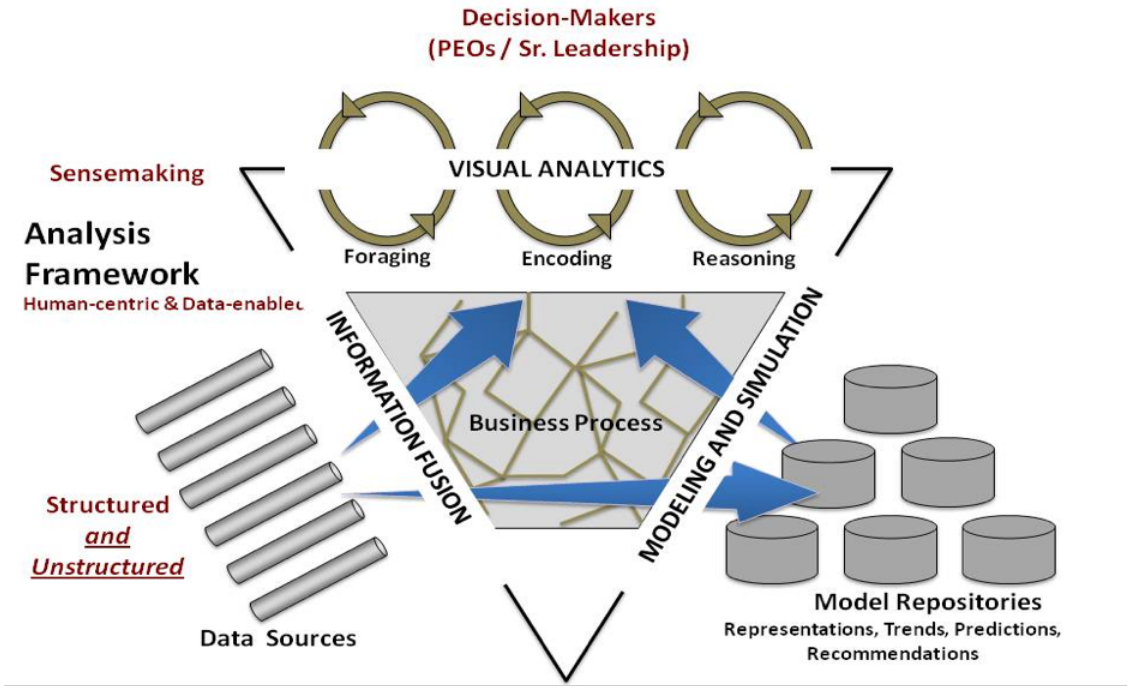


Figure 1. Overall visual analytics framework for complex business processes

Figure 2. Overall visual analytics framework for complex business processes.

Once discovered, these aspects should be investigated in detail, by referring back to original data and gathering additional evidence to confirm or refute hypotheses that are formed. A comprehensive visual analytics approach would support all these aspects. In this paper we apply such a comprehensive approach, employing visual analytics methods for both automated analyses and user-led exploration through the interactive visual interface.

Our first step is to embed the dashboard system in a visual analytics framework. This is depicted schematically in Figure 2. The repository of data sources (lower left) will inevitably contain both structured and unstructured data. Since a range of sources will be involved, the repository will be heterogeneous. Since useful information could be found anywhere (including perhaps outside the business) and, once found, should be used, some of these sources could be incomplete or fragmentary. Analytic tools must thus be flexible, able to handle different kinds of data, including incomplete data, yet producing results that contribute to a "common picture" that will be meaningful to the user.

Inevitably some of the data in a business process will be textual, containing comments, annotations, descriptions, reports, human communications, and so on. These texts will tend to be partly or mostly unstructured. Therefore we apply methods we have developed for extracting topics and topic-based events over time from unstructured texts [9]. These results are combined with named entity extraction for people's names, dates including future dates, and locations (if desired). The analysis can be applied, with approximations, either to streaming data [9] with results in a couple of minutes or less, or more slowly and accurately to histories over a period of time. These methods are quite flexible and have been applied to many different types of texts including reports, research papers, patent data, twitter streams, online news, and customer messages [15, 19, 18]. To these textual analyses, we add methods for exploring categorical data and numerical data, including finding trends over time. The categorical methods, for example, reveal relationships between multiple categories at selected times or time ranges, which are related to specific events in the topical event analysis. For all these methods plus the embedded dashboard, we develop a new interactive visual interface, as described in Sec. 6.

The analytic methods apply not only to the data repository but also to the business processes embedded in the middle of Figure 2 and, indeed, even to the model repository at the lower right. The business processes produce new information themselves and thus must be included to produce a most useful and comprehensive common picture. (This is indicated by thickening red lines that cross the middle of the figure.) In fact, as we have seen above, it is quite important to analyze information created during the business process since that will produce direct knowledge of the ca-

pabilities and capacities of different parts of the process as well as signal when things are going wrong and what parts of the process are affected.

A general definition of a physical event is "a meaningful occurrence in space and time" [14]. For the purposes of this framework, we modify this definition to focus on "process events", which we define as "meaningful occurrences in time with respect to the process". In both cases, time is central. For this framework, bursts of activity over a relatively short time scale, unexpected trends or relationships (that appear in a short time span), or outliers could all qualify as process events. We have shown that events can be automatically found and organized (including putting them in a hierarchical structure) for social media such as Twitter [16]. This work, involving topic modeling, also demonstrates how users can quickly attach meanings to the events. The hierarchical structure becomes significant when one deals with complex data for which one can have many events. As data grows in size and comprehensiveness, it will tend to get complex in this way. Certainly the complex manufacturing and production processes considered here will benefit from hierarchical structuring, which will give them high level meaning and make the overall processes easier to understand.

5. Virtual Sensors

We now have the mechanisms we need to build the effective framework illustrated in Figure 2. The main idea is to instrument the whole business process (acquisition, production, deployment, etc.) with "virtual sensors". These sensors are flexible so that they can access any data stream (e.g., unstructured text, categorical data, numerical data, model outputs) anywhere, such as in the data repository, in the business processes themselves, or in the models. Typically sensors are placed at a set of general points that are effective for monitoring many different types of production. Then they are placed at additional points that are useful for a specific type of business process. Finally they can quickly be placed at other points by the manager, often to understand some abnormality in the process. This set of networked sensors then report to the visual analytics layer at the top of Figure 2. This layer provides a set of "sensemaking" capabilities to explore, understand, reason with, and test out hypotheses w.r.t. the sensor results. Since these results are expressed in a common language of process events, they can be displayed together and manipulated in an interactive visualization, though some details with respect to the events will be different depending on the underlying data or process. At any point, the manager can select one or more events and get at the underlying data.

The structure of the virtual sensor is depicted in Figure 3. The bottom layer grabs data directly from the business process stream. In the case of the business partner mentioned above, the dashboard already collection points that can be used to insert the sensors at several places. The middle layer converts these data to a standard internal format recognized by the analysis framework in Figure 2. It also attaches whatever metadata or labels are available (time range and time stamps, data descriptors, data units, etc.). The top layer transmits the data and metadata to the analysis framework.

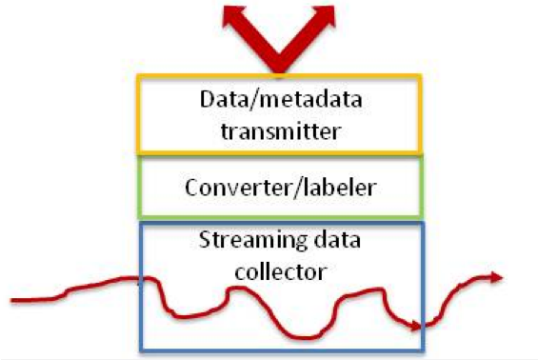


Figure 3. Virtual sensor structure.

Full analysis is performed by an analytics engine in the analysis framework, depending on the type of data. For instance, the analytics engine applies topic modeling and topic-based event extraction to unstructured text messages [19]. It applies temporal signal processing and anomaly detection to streams of numerical data to pull out both events and anomalies. It performs categorical and multidimensional analysis to categorical or attribute data. It performs relational and other statistical analyses to connect events across different types of data and to find similar signatures. These analyses are automated and all results are in a common event-based format regardless of the original input format (using the process event definition in Sec. 4.). Thus, for example, alignments of events from the text messages with events from categorical or numerical data streams would be readily spotted in the embedded visual interface described in the next section and then delved for deeper relationships. The common event-based format has the capacity for depicting additional "decorations" of the events stemming from details of the different types of original streaming data.

6. The Embedded Visual Interface

The visual interface is designed to support sensemaking in the business processes. More specifically, the visual interface incorporates 4 different views that highlight temporal, geospatial, and topical summaries using both the information available a priori (such as knowledge of the organization hierarchy) and information collected in real time by

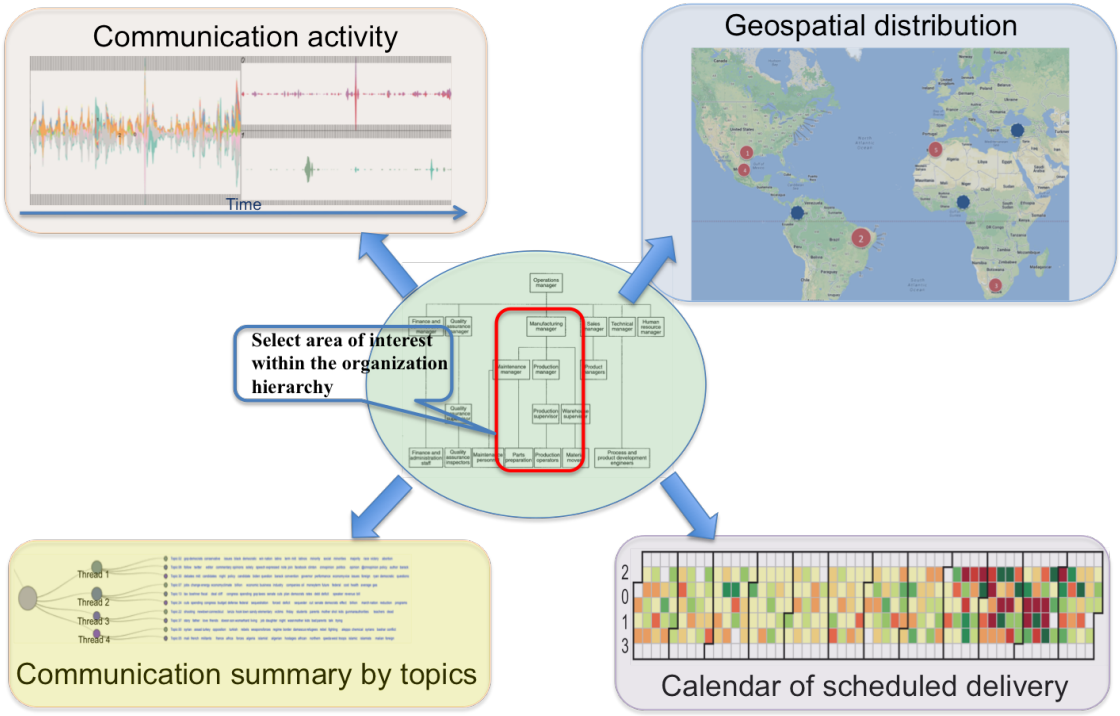


Figure 4. Embedded visual interface with labels to indicate functions.

the virtual sensors.

At the heart of the visual interface is the embedding into the external automatic and real-time sensory data repository and analyzer. The visual interface is set up to run on a single user's computer with intimate access to the external repository; it captures information flows to and from that user around office documents, calendars, emails, organizational charts, and Web pages, etc. Depending on the user analysis needs, the visual interface utilizes various meta-information that comes along with the virtual sensor results. In particular, it collects information about which documents, emails, and web pages were read for that user, and how long they were open. It also collects document metadata, such as information about the senders and recipients of email messages.

All this heterogeneous data is stored in a unified NoSQL external data repository for future analysis. On the fact collection level, tools like UpLib [8], can be utilized to extract information from and about each data input, including its title and authors, its text, the people and other entities that it mentions, its paragraphs and its images. All of the captured information is indexed and grouped with its related documents. On the event level, our signal processing then performs event analysis, time series analysis, clustering, and narrative reconstruction based on the collected organizational facts. All this is then interactively presented to the user through the embedded visualization interface.

6.1. Visual Components

Instead of presenting the diverse analyzed event structures through a keyword search interface, our framework embeds the investigative retrieval cues (e.g., who, when, where, what) into a coordinated multilevel visualization system.

At a high level, our visual interface encodes the four cues with a set of four visualizations, each of which presents a particular aspect of the organizational and business process activity. To provide a lower level detailed view, our interface also presents a visualization that integrates related activity information for a single worker or a sensor. Using this multi-level structure, the interface helps users to cohesively find the specific details they need.

Organizational Hierarchy View To give an overview of the complex operational environment, we designed an Organizational Hierarchy view that gives a cohesive overview while encapsulating organizational knowledge. Therefore, a user can start the analysis process by recalling event activities beginning from any retrieval cue that they remember (e.g., the department that mostly likely to initiated the process), or they can focus on events or cross-organization communication links that the automated analysis has identified as important. (See the application example in Sec. 7 for an illustration of the latter.) Interaction in any view cause

updates in coordinated views so that the maximum associated information is provided permitting the user to converge on the desired conclusion quickly.

The Where: Geospatial view In the geospatial view, locations involved in the overall business process are highlighted in an interactive map. The importance of the location is represented by the size of the circles and could be determined by user-selected criteria. The numbering on the circle indicates the sequence of the manufacturing or delivering process. The color of the circles is used to denote main/alternative supply or manufacturing locations. The map supports standard user interactions such as zooming and panning. The map is coordinated with other views in that selections within the map will filter to events related to the chosen locations.

The When: Activity Heatmap View To facilitate multiscale temporal analysis, we developed an Activity Heatmap view that allows analysts to monitor events in both retrospective and future time frames. This view shows how a business process unfolds over time. It presents both the number of organizational activities (e.g. message exchange) that have interacted with business process, and the types of that business and its key personnel. This view is created as an interactive calendar for ease of interaction and it shows the temporal trends and patterns of organizational activities. Each cell in Figure 4C represents aggregated mentions of that date in all the existing business processes; dates that have been mentioned more frequently appear in a darker shade of blue. For example, we can quickly see that there are multiple delivery events that happen on the July 30th of this year.

Using this view, a manager can highlight a time range to select a subset of data (e.g., an hour to a day or months) to be analyzed retrospectively. Besides showing general trends and patterns, the temporal view also allows the user to drill down into time periods. When the user selects a time period on the horizontal axis in the center of the view, our visual interface will zoom into that period of time and present the most relevant business information.

The Who: Communication Activities between participating parties To help corporate managers and decision makers efficiently retrieve specific events of interest, we designed the Communication activities view to aggregate both the documents and the people that a user has interacted with during a particular period of time. Like Lee et al. [10], our activities view allows the user to filter and sort information based on automatically-extracted data facets, including different communication types and format. The user can visually depict the relationships between the extracted organizational facts with selected business projects. For example, the user can choose to see or hide activities with email, with office documents, with Web pages, or with people. Each of these facets can be turned on or off by pressing an associated button.

In order to fit the activity information into a reasonable guided exploration, our visual interface sorts events by importance, and displays the most important documents at the top and with the most salient presentation by computing the importance value centered around the relevance to its core business process. As described in Sec. **??**, the organizational hierarchy view is also important in highlighting which events should be followed. In addition, to enable fast exploration, a summarized information panel (see Figure 4(D) right) is shown when the mouse hovers over a visual element representing an activity. Like the Document Card [17], this panel includes a readable thumbnail and aggregated information about that visual element. If the user needs more details, the user can double click on the visual element to bring up a specific detail view.

The What: Communication Summary by Topics To support the analysis of communications among different teams and parties during the overall business process, we designed the Communication Summary view, which provides an overview and permits analysis based on streaming or accumulated communication data. The analysis enables users to discover pressing issues that need to be addressed in order to keep the entire business process on track.

We assume most of the communication data are unstructured texts sent between communicating parties or attached as annotation to the stages of the business process. The virtual sensors collect all communications and put all the received communications into a database. The attribute in the database contains sender, receiver, time stamp, communication content, and other available information. All collected texts are then summarized into semantically meaningful topics using state-of-the-art topic models [2] as shown in Figure 4B. The topic summary presents individual topics to summarize what the communications are about. Coupling with the temporal trend of the topics shown in Figure 4A, one can quickly discover when a burst of communications has occurred and what issues were discussed. Other forms of data that are collected and analyzed are discussed in Sec. 7.

We now have complete support for the sensemaking process depicted at the top of Figure 2. The foraging stage is supported by the distribution of virtual sensors and their outputs. The encoding is supported by the conversion to a common event-based format in the analytics engine. The reasoning (including hypothesis-building and testing) is supported by the exploratory visual interface, which includes placement of virtual sensors where needed, and review of original data. All these stages are iterative and inter-

connected through the visual interface. In the next section we will discuss how this sensemaking can be applied.

7. An Application Example

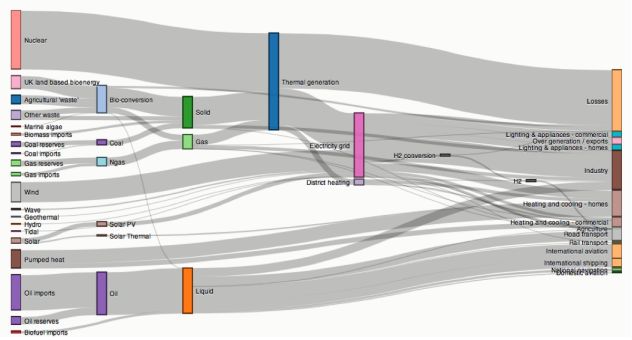


Figure 5. Illustration of virtual sensors deployed in manufacturing process. Four virtual sensors are labeled V.S. in the diagram.

In this section, we will build on the example given in the Introduction and illustrated in Figure 1. A real world manufacturing operation of any size will have a bigger organizational structure than depicted in Figure 1 and will have a layer of senior management (C-level: CEO, CIO, CTO, CFO, etc.).above the production manager. Because of complexity, the changing nature of dynamic JIT manufacturing and production, and the fact that signals of interest can come from unexpected places, our approach must be exploratory, bringing information from disparate sources into a common picture for evaluation and action. However, it is also data-driven, and since the initial data analytics especially are automated, there will be some tension between automated and user-directed processes. The visual analytics framework and the interactive visual interface are designed to minimize this tension and to make these processes work together in support of effective user actions.

A main aspect of the data-driven analysis is to identify process events (defined in Sec. 4) and to arrange them so that they tell a meaningful story when interpreted by the user. In the case of the dashboard used by our business partner, process events can be derived from numerical and categorical data collected at each stage of the dynamic manufacturing process, status reports indicating whether each stage or component of the process is on, behind, or ahead of schedule at that point, and individual annotation including comments by managers or engineers for that stage. There is also a mechanism for communication among managers, engineers, operators, and supervisors across the organization. The dashboard thus provides comprehensive data that we use in the above implementation of the visual analytics framework. However, though it provides comprehensive data, we have heard from the executive in charge that the dashboard is not sufficient to fulfill the need for on-time dynamic processing with minimum disruptiontoo many real or potentials points for breakdown can be overlooked; likewise too many opportunities for improvement and efficiency can be missed.

Thus we instrument the dashboard data and annotation collection points throughout the manufacturing process with virtual sensors, as illustrated in Figure 5. Process events are then identified in multiple ways. One way of doing so is by identifying bursts of activity over relatively short time scales in the topics derived from the textual annotations and communications. There could also be a trend in the numerical data that is unexpected compared to the normal manufacturing process. In the example here of not having a particular part when needed farther along in the manufacturing process, the burst of activity centers around one or more topics derived from both the annotations and the communications among workers indicating the part, the component system it goes into, and the supplier. An appropriate scale for this burst of activity is a few days. (In other words, text messages on the same topic accumulate in a burst over a few days.) Named entity extraction is also applied to the texts to indicate what sections of the organization (and individuals) are communicating, the names of parts, systems, and suppliers, and dates mentioned in the texts (either past or future). In addition, analysis of the inventory signal can indicate a drop in the supply of the crucial part over time that becomes significant enough to generate another event. The importance of the process events from the topic analysis is further raised because communication occurs across organizational boundaries between the parts manager, the quality assurance manager, and the technical manager (and their teams). The organizational knowledge of what is and is not unusual communication across the organizational hierarchy is encoded in the visual analytics framework. We have shown that such temporal bursts of topical activity can be automatically identified and that they can be connected to real events with high probability. We have successfully done this event analysis for streaming Twitter data (e.g., events during the development of the Occupy Wall Street movement over a year's time) [5] and collections of research proposal abstracts and full-text research papers [6] where the events are often connected with new programs started by funding agencies. These quite different text collections and quite different time scales demonstrate that the topical event-identification methods are general.

Up to this point the analysis is all automatic, but the process event results based on both statistical techniques (e.g., signal processing) and process and organization knowledge permits highlighting of events, anywhere in the process or

organizational structure, that are likely to be of interest to the high-level manager. The manager starts to interact with these results via the visual interface in Figure 4. Through the central organization hierarchy view, he sees highlighted links between the parts group and the quality manager and technical management groups that he knows are unusual. A check of the communication activity view shows when these bursts of communication occurred. The problem has not yet shown up in the calendar view (i.e., the affected process components are not yet falling behind schedule). He can then dig down to the actual texts of topical messages and can filter for geographical distributions of both suppliers and the finished systems, which will be displayed in the geographical view. He can set a new virtual sensor to watch the inventory stream (and to collect historical data). If there is a more complex problem, there may be several relevant topics. In this case, the hierarchical topic view (communication summary view) is useful. The view shows which topics are related and can be grouped and also permits regrouping by the manager. The data associated with the topics, including text messages and other data, are reorganized as well. In summary, the manager can quickly find issues that should concern him and then effectively analyze them, moving towards making key decisions.

The visual analytics framework has other useful attributes. The process of investigation by the manager that leads to a conclusion and a plan of action is itself captured in terms of selected process events, topics, the manager's own annotations, and interactions with the views in the visual interface. We have found in our studies of bank fraud analysts that an interface such as this can reveal high level strategies and its own meaningful story of how an expert reaches a decision [4]. This story can be shared with other managers, giving a rich argument for a course of action. In addition, like the fraud analysts, the managers in this dynamic manufacturing process have expertise developed through experience that is not easy to share. However in the case of our partner, the problem is exacerbated because, due to the nature of the business, there is some flow of managers into and out of their positions and thus a knowledge and experience gap. Use of the visual interface and underlying analytics by experts for a variety of situations can be captured and made available as training modules for these new managers, giving them insights that would be hard to obtain without long experience.

8. Conclusion

We have presented a visual analytics framework for handling complex business processes. It is applied to dynamic just in time manufacturing, but it is applicable to a range of agile business processes.

In order to organize data that can come from anywhere in the complex business process and can be in a variety of types, we have introduced the concepts of virtual sensors and process events. The virtual sensors permit collection from the streams of data at any point in the process and transmission of the data in a form ready to be analyzed by the central analytics engine. Process events provide a uniform expression of data of different types in a form that can be automatically prioritized and that is readily meaningful to the users. Through the visual interface, the user can place the virtual sensors, interact with and group the process events, and delve into the details of the process at any point. The visual interface provides a multiview investigative environment for sensemaking and decisive action by the user.

We have shown that with this visual analytics environment, the user can answer the major questions that must be addressed in order to achieve a successful dynamic production environment. These include finding overall trends and then detailed specific trends in the production process; determining, as early as possible, what may be going wrong and where in the process; determining what the capacities of various parts of the process are and what resource reallocation is needed; and determining what expertise should be deployed to most quickly solve a problem.

We are now working on detailed set up and evaluation of this framework. The response to the initial design and implementation by our partner is quite positive. The event- and process-based analysis presented here is of particular interest because it will bring data from multiple streams (especially textual annotations and comments) into a common view for analysis and decision-making. These are new capabilities for the partner. We are now embarking on a full project to use these capabilities in detail. This will lead to improvements and to the reporting of concrete case studies.

References

[1] B. Azvine, D. D. Nauck, C. Ho, K. Broszat, and J. Lim. Intelligent process analytics for crm. *BT Technology Journal*, 24(1):60–69, Jan. 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] R. Bobrik, M. Reichert, and T. Bauer. View-based process visualization. In *Proceedings of the 5th international conference on Business process management*, BPM'07, pages 88–95, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, pages 155–162, Washington, DC, USA, 2007. IEEE Computer Society.

[5] W. Dou, D. Wang, Z. Ma, and W. Ribarsky. Discover diamonds-in-the-rough using interactive visual analytics

system-tweets as a collective diary of the occupy movement, 2013.

[6] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231–240, 2011.

[7] G. Hansen. *Automated Business Process Reengineering: Using the Power of Visual Simulation Strategies to Improve Performance and Profit*. Prentice-Hall, Englewood Cliffs, 1997.

[8] W. C. Janssen. The uplib personal digital library system. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 410–410, New York, NY, USA, 2005. ACM.

[9] T. Kraft, X. Wang, J. Delawder, D. W., L. Yu, and W. Ribarsky. Less after the fact: Investigative visual analysis of events from streaming twitter. In *IEEE symposium on large data analysis and visualization*, 2013.

[10] B. Lee, G. Smith, G. G. Robertson, M. Czerwinski, and D. S. Tan. Facetlens: exposing trends and relationships to support sensemaking within faceted datasets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1293–1302, New York, NY, USA, 2009. ACM.

[11] A. Leoni, M. de, W. van der, and A. t. Hofstede. Process mining and visual analytics: Breathing life into business process models. *BPM Center Report BPM-11-16, 31pp*, 2011.

[12] M. Leoni, W. Aalst, and A. Hofstede. Visual support for work assignment in process-aware information systems. In M. Dumas, M. Reichert, and M.-C. Shan, editors, *Business Process Management*, volume 5240 of *Lecture Notes in Computer Science*, pages 67–83. Springer Berlin Heidelberg, 2008.

[13] P. Oude Luttighuis, M. Lankhorst, R. Van De Wetering, R. Bal, and H. Van Den Berg. Visualising business processes. *Comput. Lang.*, 27(1-3):39–59, Apr. 2001.

[14] W. Ribarsky, E. Sauda, J. Balmer, and Z. Wartell. The whole story: Building the computer history of a place. In *Hawaii International Conference on Systems Science (HICSS 2012*, 2012.

[15] S. A. Ruffa. Going lean: How the best companies apply lean manufacturing principles to shatter uncertainty, drive innovation, and maximize profits. *AMACOM (American Management Association)*, 2008.

[16] A. Streit, B. Pham, and R. Brown. Visualization support for managing large business process specifications. In W. Aalst, B. Benatallah, F. Casati, and F. Curbera, editors, *Business Process Management*, volume 3649 of *Lecture Notes in Computer Science*, pages 205–219. Springer Berlin Heidelberg, 2005.

[17] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA, 2004. ACM.

[18] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-si: Scalable architecture for analyzing latent topical-level information from social media data. *Comp. Graph. Forum*, 31(3pt4):1275–1284, June 2012.

[19] X. Wang, W. Dou, W. Ribarsky, D. Skau, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 93–102, Washington, DC, USA, 2012. IEEE Computer Society.

[20] W. Wright. Business visualization adds value. *Computer Graphics and Applications, IEEE*, 18(4):39–, 1998.

VASA: Interactive Computational Steering of Large Asynchronous Simulation Pipelines for Societal Infrastructure

Sungahn Ko, Jieqiong Zhao, Jing Xia, *Student Member, IEEE*, Shehzad Afzal, Xiaoyu Wang, *Member, IEEE*, Greg Abram, Niklas Elmqvist, *Senior Member, IEEE*, Len Kne, David Van Riper, Kelly Gaither, Shaun Kennedy, William Tolone, William Ribarsky, David S. Ebert, *Fellow, IEEE*

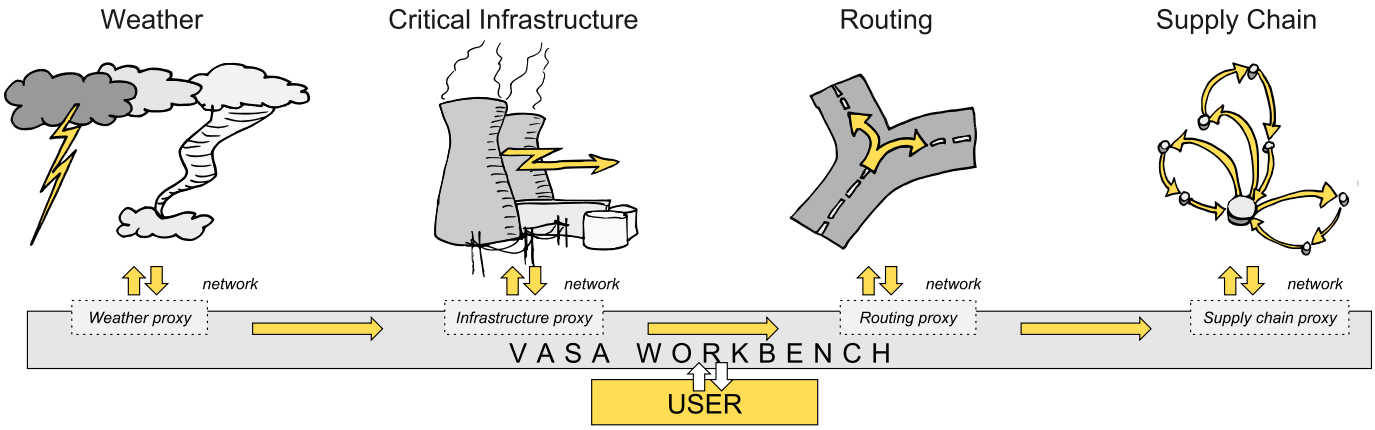


Fig. 1. Conceptual overview of the VASA system, including four simulation components for weather, critical infrastructure, road network routing, and supply chains, as well as the VASA Workbench binding them together.

Abstract—We present VASA, a visual analytics platform consisting of a desktop application, a component model, and a suite of distributed simulation components for modeling the impact of societal threats such as weather, food contamination, and traffic on critical infrastructure such as supply chains, road networks, and power grids. Each component encapsulates a high-fidelity simulation model that together form an asynchronous simulation pipeline: a system of systems of individual simulations with a common data and parameter exchange format. At the heart of VASA is the Workbench, a visual analytics application providing three distinct features: (1) low-fidelity approximations of the distributed simulation components using local simulation proxies to enable analysts to interactively configure a simulation run; (2) computational steering mechanisms to manage the execution of individual simulation components; and (3) spatiotemporal and interactive methods to explore the combined results of a simulation run. We showcase the utility of the platform using examples involving supply chains during a hurricane as well as food contamination in a fast food restaurant chain.

Index Terms—Computational steering, visual analytics, critical infrastructure, homeland security.

1 Introduction

Highways, interstates, and county roads; water mains, power grids, and telecom networks; offices, restaurants, and grocery stores; sewage, landfills, and garbage disposal. All of these are critical components of our societal infrastructure that help run our world. However, the complex and potentially fragile interrelationships connecting these components also mean that this critical infrastructure is vulnerable to both natural and man-made threats: twisters, hurricanes, and flash floods; traffic, road blocks, and pile-up collisions; disease, food poisoning, and major pandemics; crime, riots, and terrorist attacks. How can a modern society protect its critical infrastructure against such a diverse range of threats? How can we design for resilience and preparedness when perturbation in one seemingly minor aspect of our infrastructure may have vast and far-reaching impacts across society as a whole?

Simulation, where a real-world process is modeled and studied over time, has long been a standard tool for analysts and policymakers to answer these very questions (e.g., applications for modeling the real world [10]). Using complex simulations of critical infrastructure components, expert users have been able to create "what-if" scenarios, calculate the impact of a threat depending on its severity, and—last but not least—study optimal mitigation measures to address them. In fact, analysts have gone so far as to name "simulation as the new innovation" [35]: instead of endeavoring to produce the perfect solution once and for all, this new school of thought is to create a whole range of possible solutions and determine the optimal one using modeling and simulation. For example, during the Obama reelection campaign, it was reported that Organizing for Action data analysts ran a total of 62,000 simulations to determine voter behavior based on data from social media, political advertisements, and polling [43]. Basically, the philosophy with big data analytics driven by simulation is not to get the answer perfectly right, but to be less wrong over time [34]. Put differently, while it would be inappropriate to state—as others have done [2]—that big data will never somehow overtake theory, it is clear

- *Sungahn Ko, Jieqiong Zhao, Shehzad Afzal, Niklas Elmqvist, and David S. Ebert are with Purdue University in West Lafayette, IN, USA. E-mail: {ko, zhao413, safzal, elm, ebertd}@purdue.edu.*
- *Jing Xia is with Zhejiang University in Hangzhou, China. E-mail: xiajing@zjucadcg.cn.*
- *Xiaoyu Wang, William Tolone, and William Ribarsky are with University of North Carolina at Charlotte in Charlotte, NC, USA. E-mail: {xiaoyu.wang, ribarsky}@uncc.edu.*
- *David Van Riper, Len Kne and Shaun Kennedy are with University of Minnesota in Minneapolis, MN, USA. E-mail: {vanriper, lenkne, kenne108}@umn.edu.*
- *Greg Abram and Kelly Gaither are with University of Texas at Austin in Austin, TX, USA. E-mail: {gda, kelly}@tacc.utexas.edu*

Submitted to IEEE VAST 2014. Do not redistribute.

that large-scale simulation is a new and powerful tool in our arsenal for making sense of the world we live in.

Applying simulation to the scope of entire critical infrastructures—such as transportation, supply chains, and power grids—as well as the factors impacting them—such as weather, traffic, and man-made threats—requires constructing large *asynchronous simulation pipelines*, where the output of one or more simulation models becomes the input for one or more other simulations arranged in a sequence with feedback. Such a *system-of-systems* [12, 30] (SoS) will enable leveraging existing high-fidelity simulation models without having to create new ones from scratch. However, this approach is still plagued by several major challenges that all arise from the complexity of chaining together multiple simulations in this way: (C1) *monolithic simulations* that are designed to be used in isolation, (C2) *complex configurations* for each model, (C3) *non-standard data exchange* for passing data between them, and (C4) *long execution times* for each individual simulation that are not amenable to interactive visual analytics.

To address these challenges, we present **VASA** (Visual Analytics for Simulation-based Action), a visual analytics platform for interactive decision making and computational steering of these types of large-scale simulation pipelines based on a visual analytics approach. The VASA Workbench application itself is an interactive desktop application that binds together a configurable pipeline of distributed simulation components. It eables the analyst to visually integrate, explore, and compare the inter-related and cascading effects of systems of systems components and potential final alternative outcomes. This is acheive by visualizing both intermediate and final results from the simulation components using a main spatiotemporal view as well as multiple secondary views. The tool provides an interface for the analyst to navigate in time, including stepping backwards and forwards, playing back an event sequence, jumping to a particular point in time, adding events and threats to the timeline, and initiating mitigation measures. Moreover, it allows them to select between or combine different ensemble outputs from one simulation to be fed to other SoS components and explore consequences. Using this interface, an analyst could for example add a weather event (e.g., either an existing hurricane from a historical database, the union of several ensemble output paths, or simulation of a new one) to a particular time, and then step forward a week to see its impact on roads, the power grid, food distribution, and total economic impact in southern United States.

The simulation components provide the main functionality to the VASA platform. Each simulation component communicates with the Workbench using a representational state transfer (REST) API that standardizes the data and parameter exchange. The data flows and parameters passed in the pipeline can be configured using the Workbench application using a graphical interface. Furthermore, the Workbench also includes a local *simulation proxy* for each remote simulation component that provides real-time approximations of each simulation model to enable using them for interactive visual discourse. This feature also provides the computational steering functionality of the Workbench: after configuring a simulation run in an interactive fashion, the analyst can launch the (possibly lengthy) execution from the Workbench. The Workbench then provides tools to manage the simulation pipeline, for example to prematurely shut down a simulation component to accept a partial result, or to skip a particular run.

Our work on the VASA project has been driven by stakeholders interested in supply chain management of food systems, with an initial working example of a food production to restaurant system. For this reason, other than the VASA Workbench application and the protocols and interfaces making up the platform, we have also created VASA components for simulating weather (including storms, hurricanes, and flooding), the power grid, supply chains, transportation, and food poisoning. We describe these individual components and then present an example of how the VASA platform can be used to explore a what-if scenario involving a major hurricane sweeping North Carolina and knocking out a large portion of the road networks and power grid. We also illustrate how the tool can be used to simulate food contamination outbreaks and how this information can be used to track back the contaminated products to the original distribution centers.

2 Background

Visual analytics [38], can be a powerful mechanism to harness simulation for understanding the world. Below we review the literature in visual analytics for simulation and computational steering, as well as appropriate visual representations for such spatiotemporal data.

2.1 Simulation Models

The potential for applying visual analytics to simulation involves not only efficiently presenting the results of a simulation to the analyst, but also building and validating large-scale and complex simulation models. For example, Matkovic et al. [27, 28] show that visual analytics can reduce the number of simulation runs by enabling users to concentrate on interesting aspects of the data. Maciejewski et al. [23] apply visual analytics techniques to support exploration of spatiotemporal models with kernel density estimation and cumulative summation. This work was extended to a visual environment for epidemic modeling and decision impact evaluation [1]. Similarly, Andrienko et al. [5] propose a comprehensive visual analytics environment that includes interactive visual interfaces for modeling libraries and supports selection, adjustment, and evaluation of such modeling methods. Our work is different from this prior art in that our approach combines multiple components in a simulation pipeline, where each stage in the pipeline produces visualization for analysis.

Supply chain management is also a multi-decisional context where what-if analyses are often conducted to capture provenance and processes of supplies. Simulation is recognized as a great benefit to improve supply chain management, providing analysis and evaluation of operational decisions in the supply process in advance [37]. With the IBM Supply Chain Simulator (SCS) [9] and enterprise resource planning (ERP), IBM is able to visualize and optimize nodes as well as relations in the supply chain [20]. Perez also developed a supply chain model snapshot [31] with Tableau. However, existing visualizations of supply chain are mostly limited to either local supply nodes or a metric model rather than managing the overall supply process.

2.2 Computational Steering

Computational steering refers to provoding user control over running computations, such as simulations. Mulder et al. [29] classify uses of computational steering as model exploration, algorithm experimentation, and performance optimization. Applications include computational fluid dynamics (CFD) [13], program and resource steering systems [40], and high performance computing (HPC) platforms [7].

For all of the above applications, the user interface is a crucial component that interprets user manipulation for reconfiguration of data, algorithms, and parameters. Controlling, configuring, and visualizing such computational steering mechanisms is an active research area. Waser et al. proposed World Lines [41], Nodes on Ropes [42], and Visdom [33] as well as an integrated steering environment [33] to help users to manage *ensemble simulations*—multiple runs of the same or related simulation models with slightly perturbed inputs—of complex scenarios such as flood simulations. In the business domain, Broeksema et al. [8] propose the Decision Exploration Lab (DEL) to help users explore decisions generated from combined textual and visual analysis of decision models rooted in articifical intelligence.

2.3 Spatiotemporal Data

Spatiotemporal visual analytics systems enable users to investigate data features over time using a visual display based on geographic maps [3]. In these systems, color, position, and glyphs display features of different regions by directly overlaying the data on the map.

Many approaches to visual analytics for spatiotemporal data exist. Inspired partly by a survey by Anselin [6], we review the most relevant ones below. Andrienko and Andrienko [4] use value flow maps to visualize variations in spatiotemporal datasets by drawing silhouette graphs on the map to represent the temporal aspect of a data variable. Hadlak et al. [16] visualize attributed hierarchical structures that change over time in a geospatial context. Fuchs and Schumann [15] integrate ThemeRiver [17] and TimeWheel [39] into a map to visualize spatiotemporal data. Ho et al. [18] present a geovisual analytics

framework for large spatiotemporal and multivariate statistical flow data analysis using bidirectional flow arrows coordinated and linked with a choropleth map, histogram, or parallel coordinates plot. Our approach is different from those in that our system provide a visual analytics environment for managing and analyzing the results from multiple types of simulations.

Some approaches enable analysis of spatially-distributed incident data. Maciejewski et al. propose a system for visualizing syndromic hotspots [24, 22] while Malik et al. [25] develop a visualization toolkit utilizing KDE (Kernel Density Estimation) to help police better analyze the geo-coded crime data. The latter system is extended to a visualization system [26]where historic response operations and assessment of potential risks in the maritime environment can be analyzed. In our work we also employ KDE for visualizing spatial distrubution of ill people who consumed contaminated food in a supply chain.

3 Design Space: Steering System-of-System Simulations for Modeling Societal Infrastructure

Computational steering is defined as user intervention in an autonomous process to change its outcome. This approach is commonly utilized in visual analytics [38] to introduce a human analyst into the computation loop for the purpose of creating synergies between the analyst and computational methods. In our work, the autonomous processes we are studying are simulation models (often based on discrete event models) that are chained together into asynchronous simulation pipelines where the output of one or several simulations becomes the input to one or several other simulations. Such a simulation pipeline is also a *system-of-systems* [12, 30] (SoS): multiple heterogeneous systems that are combined into a unified, more complex system whose sum is greater than its constituent parts. Synthesizing all these components yields the concept of visual analytics for *steering system-of-system simulations*: the use of visual interfaces to guide composite simulation pipelines for supporting sensemaking and decisionmaking. In this work, we apply this idea to modeling societal infrastructure, such as transportation, power, computer networks, and supply chains.

In this section, we explore the design space of this concept, including problem domains, users, tasks, and challenges. We then derive preliminary guidelines for designing methods supporting the concept.

3.1 Domain Analysis

A wide array of problem domains may be interested in creating large-scale system-of-system simulation pipelines for studying impacts on societal infrastructure. Our particular domain is for business intelligence for supply chain logistics in the fast-food business, but we see multiple potential applications (each with a specific example):

- **Supply chain logistics:** Impact of large-scale weather events on the distribution of goods (particularly perishables, e.g., food).
- **Public safety:** Crime, riots, and terrorist attacks on critical infrastructure, such as on roads, bridges, or the power grid.
- **Food safety:** Incidence, spread, and causes of food contamination, often due to weather (power outage) or transport delays.
- **Cybersecurity:** Societal impact of cybersecurity attacks, such as on power stations, phone switches, and data centers.

3.2 User Analysis

The intended audience for computational steering of simulation models using visual analytics are what we call "casual experts": users with deep expertise in a particular application domain, such as transportation, supply chain, or homeland security, but with limited knowledge of simulation, data analysis, and statistics. Their specific background depends on the problem domain; for example, they may be business or logistics analysts for supply chain applications, police officers for public safety, and homeland security officials for food safety and cybersecurity. Because of this "casual" approach—a term we borrow from Pousman et al.'s work on casual information visualization [32]—our intended users are motivated by solving concrete problems in their application domain, but are not necessarily interested in configuring complex simulation models and navigating massive simulation results.

Even if our primary user audience is these casual experts, it is very likely that the outcome of a simulation steering analysis will be disseminated to managers, stakeholders, or even the general public [38]. Thus, a secondary user group for consuming our analysis products is laypersons with an even more limited knowledge in mathematics, statistics, and data graphics.

3.3 Task Analysis

Based on our review of the literature (Section 2) as well as feedback from domain experts, we identify a preliminary list of high-level tasks for steering system-of-system simulations for societal infrastructure:

- Increasing *preparedness* for potential scenarios;
- Improving the *resilience* of an organization; and
- Planning for *mitigation and response* to a situation.

3.4 Challenges

Modeling the real world is a tremendously difficult and error-prone process. However, we leave concerns about the fidelity, accuracy, and quality of a simulation to research within the simulation design. Rather, in this subsection we concern ourselves with the challenges intrinsic to connecting multiple individual simulation models into large-scale pipelines. In the context of simulation steering for such pipelines, we identify the following main challenges:

C1 **Monolithic simulations:** While individual high-fidelity simulation models exist for all of the above components and threats, these models are monolithic and not designed to work together.

C2 **Complex relationships:** Each high-fidelity simulation model consists of a plethora of parameters and controls that require expertise and training, which is exacerbated when several such models are combined into a single model.

C3 **Non-standard data:** No standardized data exchange formats exist for passing the output of one simulation model, such as for weather, as input to another model, such as supply chain routing.

C4 **Long execution times:** Most state-of-the-art, high-fidelity simulation models require a non-trivial execution time, often on the order of minutes, if not hours. Such time frames are not amenable for real-time updates and interactive exploration.

C5 **Uncertainty and fidelity:** Chaining together multiple simulations into a pipeline may yield systematically increasing errors as uncertain output from one model is used as input to another. This is compounded by the fact that heterogeneous simulation models may have different levels of fidelity and accuracy.

3.5 Design Guidelines

Based on our review of the problem domain, users, and tasks above, as well as the challenges that these generate, we formulate the following tentative guidelines for designing visual analytics methods for steering system-of-system simulation pipelines:

G1 *Simulations as standardized network services:* Distributing simulation models as network services avoids the trouble of integrating a monolithic design with another system (C1) and automatically provides a data exchange format (C3). The simulations also become decoupled, which means they can be parallelized and/or distributed in the cloud to manage long execution times (C4).

G2 *Simulation proxies for interactive response:* Meaningful sensemaking in pursuit of one of the high-level tasks in Section 3.3 requires real-time response to all interactive queries. This means that long execution times (C4) of simulation models in the pipeline should be hidden from the user. We propose the concept of a *simulation proxy* as an approximation of a remote simulation service that is local and capable of providing real-time response at the cost of reduced (often significantly) accuracy.

G3 *Visual and configurable relationships:* The interactive visual interfaces routinely employed in visual analytics may help to simplify and expose the complex configurations necessary for many high-fidelity simulation models (C2), even for non-expert users.

G4 *Partial and interruptible computational steering:* Once an analyst has configured a simulation run using simulation proxies (G2) and visual mappings (G3), the full simulation pipeline must be invoked to calculate an accurate result. A full-fledged simulation run may take minutes, sometimes hours, to complete. The computational steering mechanisms provided by the software should provide methods for continually returning partial results [14] as well as interrupting a run halfway through.

G5 *Visual representations of both intermediate and final results:* To fully leverage the power of visual analytics, we suggest using interactive visual representations of simulation results. Such visualizations should be used for both intermediate data generated by a simulation component anywhere in the pipeline—which would support partial results and interrupting a run at any time—as well as for the final results. All visual representations should be designed with uncertainty in mind (C6), and providing intermediate visualizations should also help in exposing propagation of increasing error. Finally, it may also be useful to use visual representations for the approximations created by simulation proxies (G2), but these should be clearly indicated as such.

4 VASA: Overview

As previously described, our VASA system is a distributed component-based framework for steering system-of-system simulations for societal infrastructure. Figure 1 gives a conceptual model of the system architecture. At the center of the system is the VASA Workbench (Figure 2), a user-driven desktop tool for configuring, steering, and exploring simulation models, impacts, and courses of action. The workbench provides a visual analytics dashboard based on multiple coordinated views, an event configuration view, and a computational steering view. The workflow of the workbench revolves around initiating, controlling, analyzing, exploring, and handling events from the remote simulation components as well as the local simulation proxies.

Within the dashboard, events are displayed in a selectable calendar view (a) where each event's name, dates and a user-selected representative attribute (e.g., storm's maximum wind speed) are shown. The selected events from (a) are listed based chronologically in the event viewer (b) where a user can select times for investigation. In (b-1), various options are provided, including initiating simulations (e.g., cyberattack, storm simulations, distribution re-routing), selecting combinations of events (union, intersection, difference), selecting event visualization modes (polygons, contours), and chronological playback.

Users can fix a time within an event for comparison (right-clicking on a event's black rectangle) and a red mark is shown in the upper right corner of the associated rectangle(b-2), and the impact is shown in the main geospatial view (d-1). We provide a legend window (c) for selected properties (e.g., distribution centers, restaurants, power plants and other infrastructures) and the geographical view (d) provides the simulation results including event evolution, routing paths, and impacts on critical infrastructures. A food delivery schedule to each store within a supply chain is provided in (e) where the x-axis presents corresponds to different restaurants while the y-axis represents different food processing centers or different types of foods. Here, the darker the red, the larger the quantity of the delivered food. The quantity information is provided in a tooltip that helps a user to estimate possible losses. This view enables traceback analysis (e.g., which type of food was contaminated from which processing centers, how much contaminated food was delivered to which store) for food contamination incidents.

5 VASA: Components

Our current VASA suite consists of four simulation components that implement the VASA interface: components for weather, critical infrastructure, routing, and supply chains. We review each of these next.

5.1 Weather Component

In order to provide clients with a one-stop source for weather data, we implement a server that asynchronously amasses data from various online sources and presents it to clients through a RESTful web interface. This provides access to various data through a singly authenticated service that provides consistent and convenient APIs for data acquired from many sources.

5.1.1 Simulation Model

For example, a collaboration of several research centers runs the ADCIRC model during hurricane season off the east and gulf coasts of the U.S. When storms are present, these models are run every four hours, producing ADCIRC-formatted datasets at fixed intervals forward from the initial times. These results are made publicly available using THREDDS and OPeNDAP for cataloging, discovery and data access. When this data appears, we import it onto a VASA server, and provide a simple RESTful API to access the data in convenient multiresolution formats. Similarly, NOAA produces wind-speed probabilities along the tracks of storms as contours at 34, 50, and 64-knot levels. This data is also imported asynchronously onto the VASA service and provided through the VASA RESTful API.

5.1.2 Simulation Proxy

The proxy in this component has two roles. The first role is to prepare all event data sets from the remote event server. Therefore, the system first checks for new updates from the server. If there is a new update, it retrieves the data and saves it on the local workbench for faster loading. The second role is to visualize new status of an event on the date that a user selected and notify the status change of the event to other proxies. An example status change is a user changing the start date of a hurricane in the event viewer. When this happens, the proxy visualizes a new status of the hurricane on the date and notifies this change to other components, which initiates each proxy's work (e.g., estimating an area without power and impassable roads).

A user can select the hurricane visualization type either as polygons or contours for estimation by clicking a button as shown in Figure 2 (b-2, the last button). In the polygon mode, two probability models (blue with two different opacities) are projected as shown in the magnification view in Figure 2. Here, the smaller polygon means an expected path with high probability, and a larger one presents an expected path with low probability. When a user fixes a hurricane, the hurricane turns red for comparison to other paths (of other hurricanes). For example, in Figure 2 the path of Hurricane Irene on August 24, 2011 is projected (blue) and the path of Hurricane Sandy in October 27, 2012 is presented in red for comparison.

In the contour mode, hurricanes are drawn using three different sizes of contours, each of which represents mean areas in different wind speeds (e.g., Hurricane Irene in our simulation model has 64 knot highest wind speed at the innermost contour, and 34 knot lowest wind speed at the outermost contour as shown in Figure 6). To utilize different wind speeds in simulation steering, a user can set up a threshold for infrastructures (e.g., a power generation unit is disabled if the wind hitting the plant has speed higher than 34 knot). In addition, a user can apply one of the contours for a time. For example, Figure 6 (top-right) presents which power generation units are affected when a contour with 34 knot hits the area. Here red circles represent affected restaurants and red circles present the impacted power generation units supplying electricity to those restaurants.

5.1.3 Implementation Notes and Performance

From the client's point of view, the VASA API consists of URLs that encode procedures and parameters that, when issued, return JSON objects containing the results. This provides a very simple interface for use both by browser-based visualization UIs that use AJAX to issue requests asynchronously, and other native platforms that provide equivalent access through language-specific interfaces.

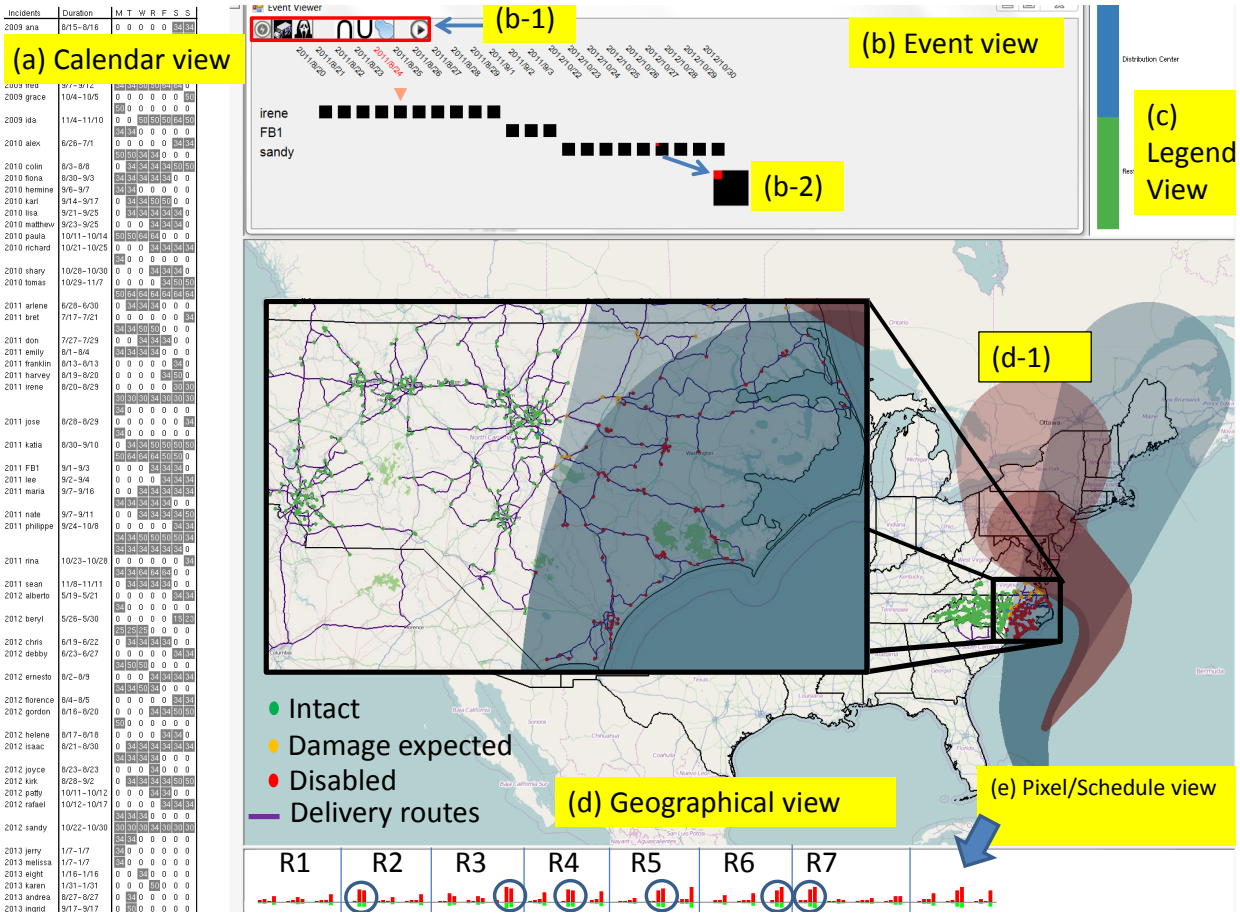


Fig. 2. Multiple coordinated views in the VASA Workbench. (a) Calendar view with available events (e.g., weather, food poisoning, cyberattack, etc). (b) Event timeline for configuring events. (b-1) Event buttons. (b-2) Fixed event. (c) Map legend. (d) Geographical map. (d-1) Hurricane (red). (e) Pixel/Schedule view showing food deliveries. Each area divided by a blue line means a route that visits 3–4 restaruns, 3 time a week. This view also can be used for pixel-based visualization.

5.2 Critical Infrastructure Component

Widespread emergencies such as hurricanes, flooding, or cyberattacks will often affect multiple societal infrastructures. High winds and flooding from a hurricane, for example, could knock out parts of the power grid, the effect of which would cascade to traffic signals, the communications network, the water system, and other infrastructures. The flooding might simultaneously make parts of the road network impassable. These breakdowns would affect critical facilities such as schools, hospitals, and government buildings. For longer-lived disasters, food distribution might break down due to power outage, route disruption, or other cascading effects. The purpose of VASA's critical infrastructure component is to simulate how such external emergencies, modeled in other components, will impact critical infrastructure.

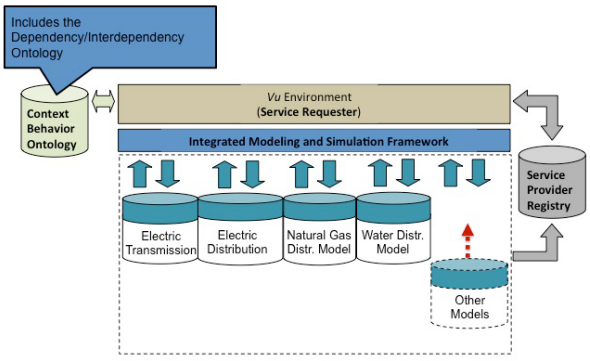


Fig. 3. Vu environment showing the modular structure where different simulation submodels can be inserted.

5.2.1 Simulation Model

To capture these complex, multifarious, and dynamic effects, we developed a simulation model that takes into account the interrelationships between critical infrastructure systems. The simulation is built within the Vu environment (Figure 3), which provides a rule-based framework for integrating multiple infrastructure components at a high level. This results in an interdependency ontology. Thus, for example, a breakdown of a power substation would immediately cascade to power loss at points on its distribution network. If a school were a node in the distribution network, it would be switched to backup power that, after a given time, would also shut down. Likewise, telecommunication nodes would switch to backup power that might also shut down after its prescribed duration. There could also be outages due to power load imbalances at other points in the grid.

These interlaced critical infrastructures are captured in a set of networks, with each node having a set of properties according to its category and the edges providing a dependency rule according to the category and state of the connected nodes. Relations between networks is captured by edges between nodes in the two networks. The timings of interdependencies and state changes are set according to a universal clock, so that any simulation of cascading effects evolves over time and space (since nodes are geographically located). The rules for networks and interdependencies are set in consultation with experts (in the case of the power grid, for example) or through consultation of the appropriate literature for an infrastructure. However, some of the interdependencies are not directly known, even by experts, since measures or simulations linking some infrastructures have never been done or validated. In this case, we define plausible rules that produce outcomes consistent with experience. This is in fact an advantage of the

Vu approach in that it permits investigating interdependencies at a high level with a quick adjustment of the model and turnaround of results. These interdependencies can then be studied in depth as needed.

The Vu approach will not capture detailed interdependency effects, transients, or complex interactions. However, it has been shown to be quite successful as a high-level simulator for a range of applications. To achieve this, it is modular, allowing new modules to be dropped in (Figure 3) and, as soon as the interdependency ontology is set up, simulations to be run. This makes it easy to configure simulations for several different infrastructures and even involving other types of simulation models, such as population and economic models.

5.2.2 Data

Our prototype system currently uses data from the state of North Carolina, and the data collection and organization process involves locating and identifying components of the various infrastructures for the state. We use publicly available data sources, in some cases identifying infrastructure components by indirect means. For example, comprehensive information about the electrical grid is closely held by the utility companies. However, we have shown our results to utility company officials and received confirmation as to their high level accuracy.

The infrastructures we currently model include the electric grid; the communications network including TV stations, radio stations, cellular switch controls, and cell towers; transportation facilities including airports, bus terminals, rail lines and terminals, bridges, tunnels, and ports; the road network including main and secondary roads; natural gas pipelines and pumping stations; critical facilities including fire stations, police stations, schools, hospitals, emergency care facilities, manufacturing locations, government buildings, and hazmat facilities.

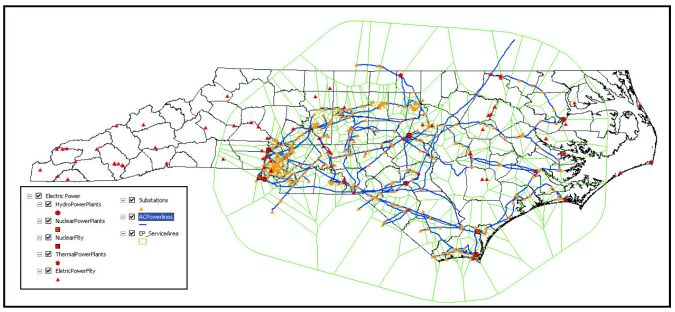


Fig. 4. Power transmission grid with parts of the distribution network.

Figure 4 shows the electric grid, which includes the complete transmission network down to substations for both North and South Carolina. Parts of the distribution network are also included, especially for critical installations. Figure 5 shows the transportation network for North Carolina, including roads, airports, rail lines, etc. For the purposes of VASA, we have also added store and distribution center locations for a large food chain in North and South Carolina. These facilities are linked to the power grid and road networks.

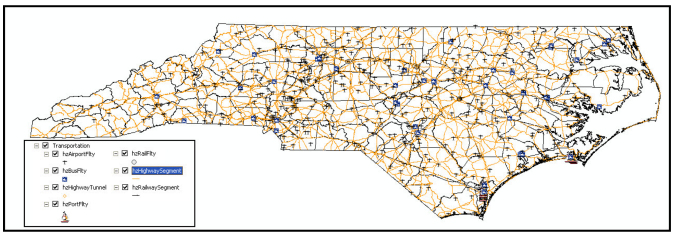


Fig. 5. Transportation network including transportation facilities.

5.2.3 Simulation Proxy

The proxy for the critical infrastructure component maintains a simplified connectivity network of critical infrastructure. In this graph, restaurants are connected to the nearest plant for impact approximation due to an event. When the proxy receives a signal of a new event (e.g., storm path change, new day for approximation), it computes which infrastructures are affected by the event. For example, when a user moves the hurricane simulation forward to a new day, our proxy checks which infrastructures are newly affected and produces an estimate and its corresponding visualization (e.g., condition color changes for restaurants affected by power plant disruptions).

5.2.4 Implementation Notes and Performance

As for all the other VASA simulation components, we use a web service that can accept requests from the VASA Workbench and send either complete or approximate simulation results ready to be presented in the user interface (see Figure 1). The critical infrastructure server itself has two components. One contains a searchable database of the pre-computed ensemble of simulation runs. The other accepts current storm path and other inputs from the weather server, converts them into courses of action, and computes a fresh set of cascading infrastructure disruption results. When a request is issued via the user interface, the Simulation Proxy determines the weather inputs to send to the ensemble database component that immediately selects the closest ensemble simulation for use in the visual analysis. This proxy is then replaced by the more accurate result based on the current weather simulation as soon as it is available. Therefore an emergency response manager can make initial decisions based on the proxy and then refine them, if desired, once the up-to-date result is available.

5.3 Supply Chain Component

Most food systems involve a number of firms from on-farm production of inputs through processing, distribution and retail sales. For the fast food system in VASA, three different firms have collaborated to provide the data on normal system performance: a vertically integrated poultry firm (hatchery to processed chicken), a warehouse and distribution firm, and a fast food restaurant firm. Each firm contributed data from their portion of the supply chain to enable modeling of product movement from farm to restaurant. The type of data provided includes geospatial information on the facilities involved (e.g., feed mills, hatcheries, poultry farms, poultry processing facilities, distribution centers and restaurants), normal transportation routes and scheduled times from each facility to the next facility in the system and details on actual shipment quantities on average (hatchery through processing) or actual shipment records for a limited time frame (distribution centers to restaurants). As an illustration of the amount of data that drives these systems, one week of data on product delivered from the two distribution centers to the nearly five hundred individual restaurants alone is over 120,000 individual records.

Hurricanes pose significant risks for normal supply chain operation from impassable roads, power outages and floodings disrupting facility operation and distribution of products throughout the system. Understanding which routes and locations are likely to be at risk from a storm would enable a firm to develop contingency plans in advance of a storm, thus reducing operational losses immediately after a storm. Given that daily sales at larger fast food restaurants can be $4,300-$7,400, losses can mount up quickly. If in the case of an impending storm or immediate aftermath, near real time rerouting could enable firms to most efficiently maintain their distribution systems for both maintaining product distribution and retrieving of food from restaurants without power to minimize spoilage losses.

5.3.1 Simulation Model

Food contamination can occur both intentionally or as a malicious act at any point in the supply chain and can result in significant public health consequences, from morbidity to mortality. While firms are required to have information one step forward and one step back in their supply chain, they often have difficulty gaining visibility beyond that. By gathering data from each step in the supply chain, it is possible to trace product from farm through to restaurant and from restaurant back to farm. Using data on actual lot sizes from the firms involved, two illustrative contamination scenarios were constructed to illustrate how differently seemingly similar contamination scenarios would transpire. This system also illustrated a common problem of “hidden nodes” in

the system, i.e., facilities that one firm in the system does not realize are part of its supply chain. One of the poultry slaughter and processing facilities ships raw poultry to a further processing facility that then ships the resulting product to the distribution centers. If there were a contamination at the "blind" facility, neither the distribution firm for the restaurant firm would initially know that it was part of their supply chain. A contamination scenario builder is now under development that would enable users to model a wide range of contamination events and see how they would propagate through the supply chain.

Our simulation model can generate food-borne illness data based on an approach similar to the Sydovat [21] system. There are two major components of the model for generating synthetic illness data: temporal and spatial data. A time series is contructed from its individual components (day-of-week, interannual, interseasonal, and remainder) similar to seasonal trend decomposition. To generate the time series of food-borne illnesses for a user-injected restaurant location, the user defines the mean daily count of illnesses along with seasonal and day of week components. If the historical food-borne illnesses data is available then seasonal and day of week components can be randomly selected from this historical data. Spatial locations for temporal data are generated based on the density distribition that approximate the population in that area. Additionally, users can customize the grid size and density distributions.

5.3.2 Simulation Proxy

Our simulation proxy for the supply chain component maintains a low-fidelity representation of the transport network. This is used together with the weather polygons to approximate when a distribution center and store must shut down. For food-poisoning data, this inherently contains spatially-distributed points of ill people simulated based on the simulation model (Section 5.3.1). To visualize the spatial distribution and the hotspots of the poisoned people, the proxy in this component uses a modified variable kernel density estimation technique with varying scales of the parameter of estimation based upon the distance from a patient location to the k_{th} nearest neighbor [36]. The model used for estimating the number of people poisoned is the same model utilized in Maciejewski et al. [1, 22], but we adjust parameters to consider different population densities in different regions.

5.3.3 Implementation Notes and Performance

The supply chain component is built in ArcGIS and Arc Network Modeler so that storm impacts can model solutions accounting for restaurants out of service (power, flooding) and impassable roadways.

5.4 Routing Component

The purpose of the routing component is to provide a mechanism for other VASA components to find appropriate routes from one facility to another given a dynamically changing world model, where roads may become impassable due to weather or other widespread emergencies.

5.4.1 Data Model

We obtained the addresses of two distribution centers and 505 fast-food outlets, as well as the route information that links the centers to the outlets. We geocoded the addresses using the Environmental Systems Research Institute (Esri) ArcGIS 10.2 Server with the Network Analyst extension, and StreetMap Premium for ArcGIS (TomTom North America data) Geodatabase. We then calculate N shortest path routes, where N is the number of routes specified in the input data, using Esri's Network Analyst Route tool and the StreetMap Premium road network. The road network has a long list of attributes used to determine the shortest route, including road class, speed limit, number of lanes, and weight restrictions.

5.4.2 Simulation Model

The input to the routing component is a GeoJSON polygon representing an area impacted by severe weather (such as a hurricane). The component ingests the GeoJSON object as a polygon barrier in the road network. Attributes of the road network are weighted to create a friction surface which iterates through routing options to determine the optimal route. The model does not currently include current traffic conditions or construction activity, but these factors could be added in the future. Each route minimizes the travel time between the distribution center and the first store or between stores. This set of routes represented the baseline scenario—how delivery trucks would travel under normal circumstances. Since delivery trucks can no longer reach outlets covered by the weather barrier, the routing service recomputes the routes with the barrier in place and returns new routes which avoid the outlets and roads covered by the barrier. If the barrier covers a distribution center, no deliveries will be made to outlets serviced by the center. The routes are output as a set of large GeoJSON objects and sent back to the caller.

5.4.3 Simulation Proxy

The main focus of the proxy in the routing component is on approximating the number of routes that will be replaced if a complete simulation result exists. The proxy investigates which nodes in routes are expected to be disabled when there is an event. Then, after the investigation, it builds a polygon by connecting outer-most nodes and visualizes the polygon. This gives awareness to a user that the routes in the polygon are likely to be changed after a complete simulation. A user can initiate the simulation by clicking the "run" button (Figure 9).

5.4.4 Implementation Notes and Performance

The goal was to use as much Commercial Off-The-Shelf (COTS) software as possible when implementing the routing model. The Esri suite of Geographical Information System (GIS) tools is widely used in a variety of industries and provides a robust set of tools and data. Specifically, we used ArcGIS Server 10.2 with the Network Analyst extension. The server provides web-based services through REST endpoints and provides a robust API accessed with HTTPS GET or POST requests. The VASA workbench initiates a request to the routing service by providing a GeoJSON representation of the affected area. The affected area polygon is input to Network Analyst Service to recalculate the route to traverse around the affected area. The response is two large GeoJSON objects containing a list of outlets no longer reachable, incremental travel time between stops, and the new route. Currently, the route processing requires 2-3 minutes to complete; this can be significantly improved when a production server is commissioned.

6 Examples

We showcase the utility of the VASA Workbench and our current simulation components using three examples: the impact of weather on macro-scale supply chains, foodborne illness contamination and spread, and a simplified cyber-attack on the power grid infrastructure.

6.1 Supply Chains in Hurricane Season

Our first example is the potential impact of hurricanes on North Carolina's critical infrastructure, especially our food distribution network, in North Carolina (NC). Our exploration begins by selecting appropriate historial hurricanes for examination using the calendar view as shown in Figure 2, where each hurricane name, duration, and selected summary attribute (e.g., maximum hurricane wind speed) are provided. While we investigate the paths of these historical hurricanes, we see that Irene in 2011 and Sandy in 2012 passed over NC. Because Sandy passed over only a small area in upper NC (Fig 2 (d), red hurricane polygon), we choose to focus on Irene for further investigation.

One interesting date is August 27, 2011 when Irene passed directly over eastern NC, an area with many power generation facilities, as shown in Figure 6 (top-right, purple circles). After we set up the wind tolerance value for these facilities to be 34 knot, our hurricane proxy instantly estimates which restaurants will be impacted based on the relationships between the units and the restaurants and colors the impacted restaurants red. Here, we also initiated a complete simulation for power outages and transportation network damage. Next, a polygon is shown representing an area where restaurants are disabled and which roads are blocked (bottom-left in Figure 6). To efficiently manage distribution, this impact requires the food provider to change its

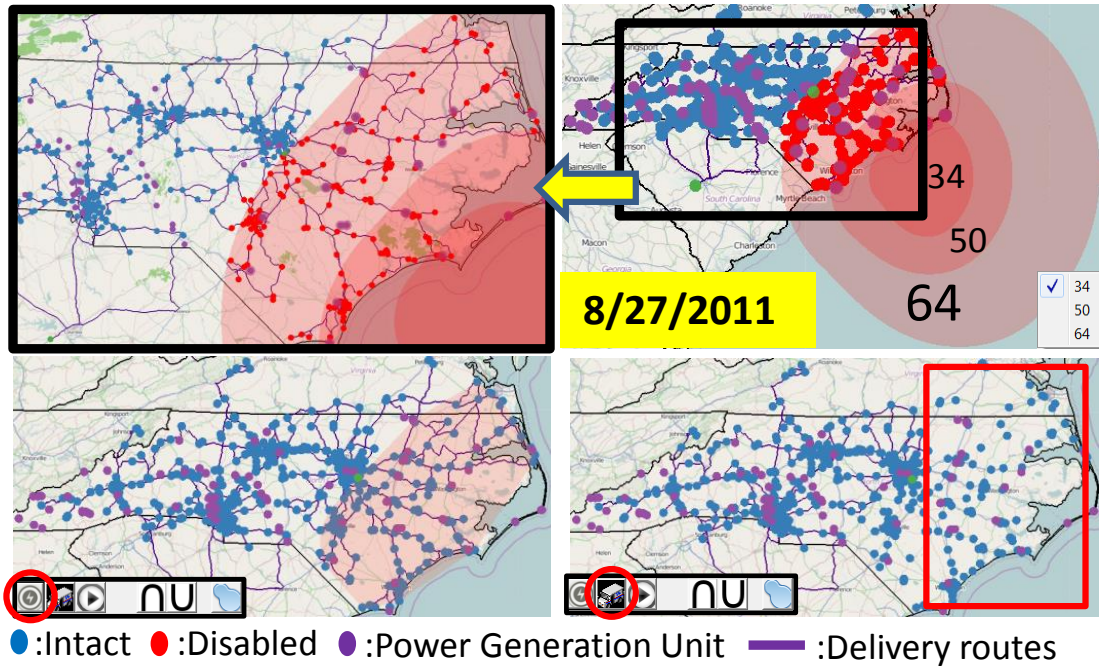


Fig. 6. In this simulation, power generation units were hit by up to 34 knot during Hurricane Irene on August 27, 2011. Our hurricane proxy instantly estimates the impacted restaurants (right-top, left-top). Note that one distribution center (green) is outside the hurricane. After a complete power-grid simulation run is finished (by clicking the circled lightning button), a polygon representing the power outage area is shown. Next, this polygon is sent for use in computing new food delivery paths. Note that food is not delivered to the power outage area (right-bottom, red box).

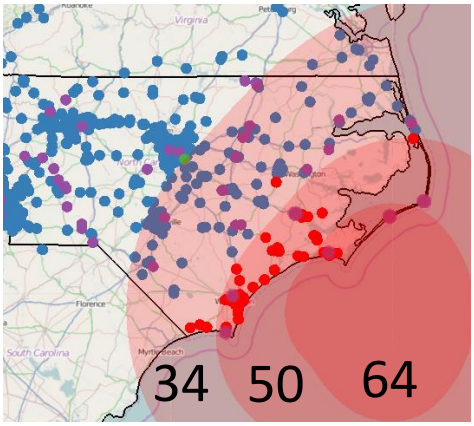


Fig. 7. If the power generation units could have resisted up to 50 knot wind, the number of impacted restaurants will be much smaller.

delivery schedule, and this new routing is computed based on the impacted restaurant polygon and road conditions (e.g., blocked by flooding). After a simulation to compute the new routes (by clicking the truck button in a red circle, right-bottom Figure 6), we see that the updated delivery paths do not include the affected restaurants. The economic loss caused by this event is estimated based on the model in Section 5.3 as being up to $1.13 million. Another possible what-if question is "How different would the result be if the power generation units can resist winds up to 50 knot?" Figure 7 shows the first step of the analysis where we see many fewer restaurants affected compared to Figure 6 top-right (units are resilient to 34 knots). In this case, the estimated losses are less than $333,000.

6.2 Fast Food Contamination

Food poisoning is an illness caused by eating contaminated food containing viruses, bacteria and germ-generated toxins. There are many possible causes of food contamination including storage at inappropriate temperatures [19], improper food handling, and cross-contamination during processing or packaging. As unfortunately experienced several times per year, tracing back the cause of the contamination is a very difficult lengthy process. In this example, we explore a hypothetical scenario demonstrating how VASA can be used to trace-back the root causes of an incident of foodborne illness.

To create the distribution of the ill population, we simulate the distribution of contaminated food to stores, then simulate the illnesses in the neighboring areas using the simulation model discussed in Section 5.3. This creates the common base scenario of reports of people who are ill, their date of illness and their location to create the food contamination scenario for the trace-back investigation.

For example purposes, we simulated these illnesses occurring during a three day span (September 1, 2011 to September 3, 2011) as shown in Figure 8. Since this is almost one week after Hurricane Irene, one may assume that power outages during the storm could be the possible reason behind the contamination. To confirm this hypothesis, we looked at the hot spots in Figure 8 and identified the stores closest to these hot spots. On cross comparison, we can identify the common products/lots in those stores, their distribution center, as well as their delivery mechanisms. As shown in Figure 8 bottom matrices, the rows represent 3 food processing centers and 4 types of food, and there is a column for each restaurant. Each cell is colored such that the darker the red color, the higher the amount of each product provided. Here, the restaurants in the affected area that are selected in the box in the top-left are highlighted with light green boxes. For stores S9 and S12, only one food processing center provided products, while other processing centers supplied most of the food throughout the network. Upon further inspection, one can determining that 3rd and 4th row product lots are common in most of the restaurants where individuals are. Some example routes are shown in Fig. 2 (e) where each route supplies 3-4 restaurants. A red bar means the supplied food and the green bar means the food consumed at a restaurant. Here, we see that a large amount of the third and fourth foods (blue circles in Fig. 2 (e)) are delivered and will all be consumed within a few days. Therefore, these two product lots are good candidates for further inspection in tracing back the contaminated food item.

6.3 Cyberattack on Critical Infrastructure

Part of the mission of the VASA project is to study the impact and mitigation of man-made attacks on societal infrastructure. Cybersecurity is becoming an increasingly important threat to modern society [11]

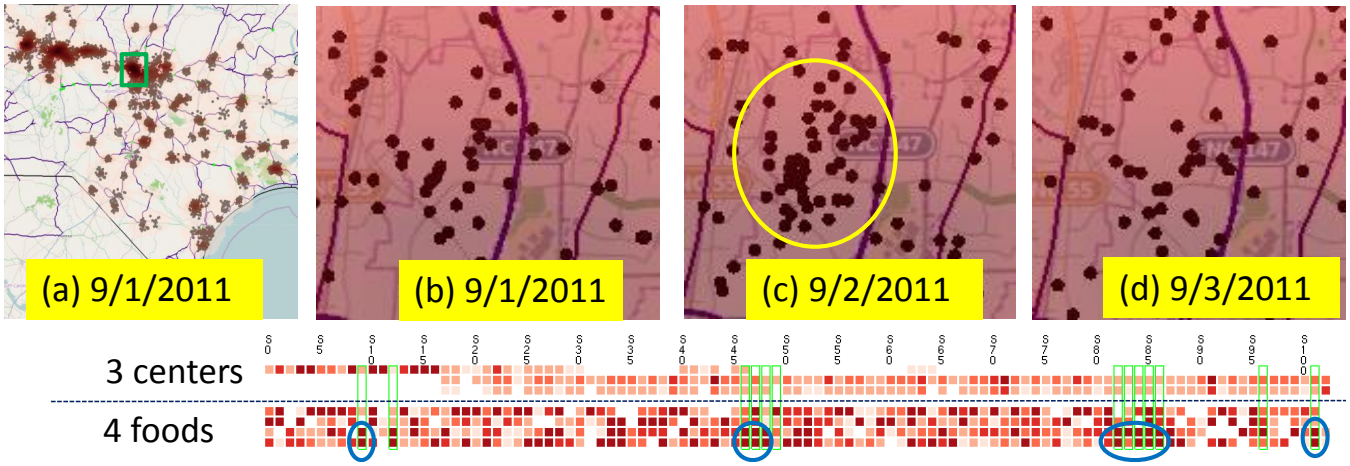


Fig. 8. Ill people caused by contaminated food is presented using a KDE hotspot visualization. In (a), the darker location has a larger number of poisoned people. Brown points mean ill people in the reported location. The locations highlighted by a green box in (a) is magnified in (b), (c) and (d) on different dates. As the timeline shows, the number of ill people increased until 9/2/2011, then started decreasing on 9/3/2011. The bottom matrices show which food processing centers (1–3) were involved and which foods (1–4) were delivered to which store in 8/30/2011, two days before the illness. Here, the restaurants in the light green boxes are the those selected by the thicker green box in (a). We see that a large quantity (darkest red pixels in blue circles) of two foods (third and fourth rows) are commonly provided to restaurants in the area.

Fig. 9. An example of the cyberattack simulation. (left) A user selects to disable power transmission by a cyberattack in the option menu and selects the region shown in the red rectangle. One power plant (purple dot) is included within this rectangle (shown in the dark red circle). (second-left) The infrastructure proxy instantly estimates the affected restaurants (red dots), and a full simulation is initiated. (second-right) Power outage regions are presented by the polygon, and new distribution routes are computed. (right) The new routes are shown as paths and do not include the affected restaurants but, unlike the hurricane scenario, all roads are available for food distribution. For comparison, see the path radiating from the polygon that was not allowed in the hurricane scenario.

and may have a significant effect on an increasingly connected society where power plants and substations are all controlled from afar.

While we do not yet provide a cyberattack module for VASA, many of the simulation components provide direct access to changing the state of particular infrastructure components through VASA. This enables us to simulate a cyberattack by, for example, shutting down a particular or several specific critical infrastructure component even if it is not affected by weather or other natural threats. Figure 9 shows a screenshot of an analyst studying precisely such a scenario where a power plant has been disabled by a cyberterrorist. Here the analyst simulates that the terrorist shuts down the power transmission grid for one of the two power plants in the town by drawing a red rectangle (left) around it. Then, the infrastructure proxy instantly estimates possible affected restaurants and a full simulation is initiated for more accuracy (second-left). The simulation result is presented by a polygon and new route computations can be initiated (second-right). The new routes with impacted restaurants are visualized. Note that this routing example is different from the hurricane case because roads are still passable: purple paths are still shown within the polygon (right).

7 Conclusion and Future Work

We have introduced the notion of visual analytics for simulation steering within the context of societal infrastructure. To our knowledge, ours is the first to study visual analytics for simulation from a *systems-of-systems* [12] perspective, where multiple heterogeneous—often physically distributed—systems are combined into a unified, more complex system in which the linkages between components provide a sum greater than its constituent parts. This notion transcends individual simulation models and instead chains together multiple high-fidelity simulations into large-scale asynchronous pipelines. The VASA system we presented as a practical example of such an approach is a distributed application framework consisting of a central Workbench controlled by an analyst and a set of loosely coupled simulation components implemented as distributed network services.

Big data simulation is a powerful new tool for data science, and while our work on applying visual analytics to this domain is conceptually complete, it really only scratches the surface of what is possible. Future work on the VASA system will involve integrating even more advanced and detailed simulation components, such as high-fidelity power grid models, gas pipelines, and power plants for energy infrastructure; bridges, tunnels, and causeways for transporation networks; and hospitals, police stations, and fire stations for societal infrastructure. In doing so, we envision designing additional novel visual representations and interactions for configuring these components as well as visualizing their proxy, intermediate, and final results.

Acknowledgments

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under award no. 2009-ST-061-CI0002.

REFERENCES

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 191–200, 2011.

[2] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, June 2008.

[3] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Verlag, 2006.

[4] N. V. Andrienko and G. L. Andrienko. Interactive visual tools to explore spatio-temporal variation. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, pages 417–420, 2004.

[5] N. V. Andrienko and G. L. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.

[6] L. Anselin. From SpaceStat to CyberGIS: Twenty years of spatial data analysis software. *International Regional Science Review*, 35(2):131–157, 2012.

[7] J. Biddiscombe, J. Soumagne, G. Oger, D. Guibert, and J.-G. Piccinali. Parallel computational steering and analysis for HPC applications using a ParaView interface and the HDF5 DSM virtual file driver. In *Proceedings of the Eurographics Conference on Parallel Graphics and Visualization*, pages 91–100, 2011.

[8] B. Broeksema, T. Baudel, A. G. Telea, and P. Crisafulli. Decision exploration lab: A visual analytics solution for decision management. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1972–1981, 2013.

[9] S. Buckley and C. An. Supply chain simulation. In *Supply Chain Management on Demand*, pages 17–35. Springer, 2005.

[10] L. Costa, O. Oliveira, G. Travieso, F. Rodrigues, P. Boas, L. Antiqueira, M. Viana, and L. Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 3(60):319–412, 2011.

[11] R. Deibert. Towards a cyber security strategy for global civil society? Technical report, The Canada Centre for Global Security Studies, 2011.

[12] D. DeLaurentis and R. K. Callaway. A system-of-systems perspective for public policy decisions. *Review of Policy Research*, 21(6):829–837, 2004.

[13] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara. Interactive focus+context analysis of large, time-dependent flow simulation data. *Simulation*, 82(12):851–865, 2006.

[14] D. Fisher, I. O. Popov, S. M. Drucker, and m. c. schraefel. Trust me, I'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1673–1682, 2012.

[15] G. Fuchs and H. Schumann. Visualizing abstract data on maps. In *Proceedings of the International Conference on Information Visualization*, pages 139–144, 2004.

[16] S. Hadlak, C. Tominski, H.-J. Schulz, and H. Schumann. Visualization of attributed hierarchical structures in a spatiotemporal context. *International Journal of Geographical Information Science*, 24(10):1497–1513, 2010.

[17] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 115–124, 2000.

[18] Q. Ho, P. H. Nguyen, T. Åström, and M. Jern. Implementation of a flow map demonstrator for analyzing commuting and migration flow statistics data. *Procedia - Social and Behavioral Sciences*, 21:157–166, 2011.

[19] B. C. Hobbs. *Food poisoning and food hygiene*. Edward Arnold and Co., London, United Kingdom, 1953.

[20] A. Kamran and S. U. Haq. Visualizations and analytics for supply chains. Technical report, IBM, February 2013.

[21] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. Cleveland, S. Grannis, and D. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *Computer Graphics and Applications, IEEE*, 29(3):18–28, May 2009.

[22] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics and Applications*, 29(3):18–28, 2009.

[23] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010.

[24] R. Maciejewski, S. Rudolph, R. Hafen, A. M. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, M. Wade, and D. S. Ebert. Understanding syndromic hotspots - a visual analytics approach. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technologoy*, pages 35–42, 2008.

[25] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert. Visual analytics law enforcement toolkit. In *Proceedings of the IEEE Conference on Technologies for Homeland Security*, pages 222–228, 2010.

[26] A. Malik, R. Maciejewski, B. Maule, and D. S. Ebert. A visual analytics process for maritime resource allocation and risk assessment. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 221–230, 2011.

[27] K. Matkovic, D. Gracanin, M. Jelovic, A. Ammer, A. Lez, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1449–1457, 2010.

[28] K. Matkovic, D. Gracanin, M. Jelovic, and Y. Cao. Adaptive interactive multi-resolution computational steering for complex engineering systems. In *Proceedings of the EuroVis Workshop on Visual Analytics*, pages 45–48, 2011.

[29] J. D. Mulder, J. J. van Wijk, and R. van Liere. A survey of computational steering environments. *Future Generation Computer Systems*, 15(1):119–129, 1999.

[30] C. Ncube. On the engineering of systems of systems: key challenges for the requirements engineering community. In *Proceedings of the IEEE Workshop on Requirements Engineering for Systems, Services and Systems-of-Systems*, pages 70–73, 2011.

[31] R. Perez. Supply chain model, April 2011.

[32] Z. Pousman, J. T. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[33] H. Ribicic, J. Waser, R. Fuchs, G. Bloschl, and E. Gröller. Visual analysis and steering of flooding simulations. *IEEE Transactions on Visualization and Computer Graphics*, 19(6):1062–1075, 2013.

[34] G. Satell. Why our numbers are always wrong. *Digital Tonto*, October 2012.

[35] G. Satell. Why the future of innovation is simulation. *Forbes*, July 2013.

[36] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[37] S. Terzi and S. Cavalieri. Simulation in the supply chain context: a survey. *Computers in industry*, 53(1):3–16, 2004.

[38] J. J. Thomas and K. A. Cook. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[39] C. Tominski, J. Abello, and H. Schumann. Axes-based visualizations with radial layouts. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1242–1247, 2004.

[40] J. S. Vetter and K. Schwan. Progress: A toolkit for interactive program steering. Technical Report GIT-CC-95-16, Georgia Institute of Technology, 1995.

[41] J. Waser, R. Fuchs, H. Ribicic, B. Schindler, G. Bloschl, and E. Gröller. World lines. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1458–1467, 2010.

[42] J. Waser, H. Ribicic, R. Fuchs, C. Hirsch, B. Schindler, G. Bloschl, and E. Gröller. Nodes on ropes: A comprehensive data and control flow for steering ensemble simulations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1872–1881, 2011.

Analyzing High-dimensional Multivariate Network Links with Integrated Anomaly Detection, Highlighting and Exploration

Sungahnn Ko* Purdue University
Shehzad Afzal* Purdue University
Simon Walton† Oxford University
Yang Yang* Purdue University
Junghoon Chae* Purdue University
Abish Malik* Purdue University
Yun Jang‡ Sejong University
Min Chen† Oxford University
David Ebert* Purdue University

ABSTRACT

This paper focuses on the integration of a family of visual analytics techniques for analyzing high-dimensional, multivariate network data that features spatial and temporal information, network connections, and a variety of other categorical and numerical data types. Such data types are commonly encountered in transportation, shipping, and logistics industries. Due to the scale and complexity of the data, it is essential to integrate techniques for data analysis, visualization, and exploration. We present new visual representations, *Petal* and *Thread*, to effectively present many-to-many network data including multi-attribute vectors. In addition, we deploy an information-theoretic model for anomaly detection across varying dimensions, displaying highlighted anomalies in a visually consistent manner, as well as supporting a managed process of exploration. Lastly, we evaluate the proposed methodology through data exploration and an empirical study.

Index Terms: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques; I.3.8 [Computer Graphics]: Applications—Visual Analytics

1 INTRODUCTION

The recent trend of increasing size, complexity, and variety in datasets (e.g., spatial, temporal, quantitative, qualitative, network data) makes analysis and decisions from these data more challenging, often called the *big data* problem [24, 34, 40]. One very challenging type of big data is multivariate network data, especially when there are multivariate values for both nodes and links. For example, transportation, shipping, logistics, commerce, trading, electricity and communication industries [8, 46] have many connected operational locations where multiple variables describe each location's operations. With flight delay network data, various multivariate operational aspects are considered simultaneously: types of delay, patterns based on airport location, trends in time, and relationships among the airports. To reduce the analysts' information overload and to enable effective planning, analysis and decision making, an interactive visual exploration and analysis environment is needed as traditional machine learning and big data analytics alone can be insufficient [10].

While various systems and techniques for network visualization have been proposed [22], few support analyzing both multivariate network data (e.g., [43] and [28]) and map-based spatial network data (e.g., [19] and [8]). There still remains a gap in effective multivariate spatial network data exploration and analysis to efficiently answer challenging questions such as the following: What are the patterns in multivariate variables on a node or among node-node pairs? Are the patterns relevant to specific regions and times? Is there any seasonality in the patterns? Can we verify the patterns on a map? Which network nodes and links could be anomalous?

In this work, we fill this gap by integrating a family of visual analytics techniques for exploring and analyzing such complex data. We employ multiple linked views [33] (see Fig. 1), two new multivariate visualization techniques, *petals* and *threads*, and an information-theoretic analytical backend engine for aggregate-level and detail-level network analysis.

Petals and *threads* efficiently present a simplified representation of many-to-many networks where multi-attribute vectors represent the size of attributes in different directions. Specifically, *petals* represent an aggregated summary view of directional data (Fig. 3) and *threads* encode multiple variables of links (Fig. 2). An information-theoretic model provides our analytical engine the ability to highlight anomalies in the data. The anomaly detection can be dynamically configured based on new contextual requirements that usually result from user-generated hypotheses stimulated from visualization and exploration of data. The analytical method provides the visualization with additional warning signals and enables users to prioritize their exploration strategy.

The contributions of our work in the multivariate spatiotemporal network visualization and analysis domain are 1) designing *petals* and *threads* for high-dimensional multivariate network link analysis, 2) evaluating *petals* and *threads* with a user study, 3) designing and implementing a visual analytics system using multiple coordinated views, 4) integrating an information-theoretic anomaly detection method in the interactive visualization analysis process, and 5) exploring complex data (e.g., flight delay network) to illustrate the use and potential of our designs in the multiple-coordinated views.

Our system can be applied to exploration of any multivariate spatiotemporal, network link data generated in transportation, shipping, logistics, commerce, trading, and communication industries (e.g., AT&T communication network data [8] and electric power grid data [46]).

2 RELATED WORK

While the research topics in network visualization are as numerous as the visualizations themselves [22, 38], in this work, we consider network visualization techniques and tools that are pertinent to multivariate geospatial network data. For multivariate network visualization research, Wattenberg [43] has designed PivotGraph, a software tool focusing on the relationships between node attributes and connections of multivariate graphs on a grid layout. Ploceus [28] enables multi-dimensional and multi-level network-based visual analysis on tabular data while Honeycomb [42] focuses on scalability (e.g., millions of connections) using a matrix representation that is also incorporated in our matrix view. Shneiderman et al. [38] visualize networks by semantic substrates and Selassie et al. [36] present an edge bundling technique for directed networks.

For geospatial network visualization, Guo [19] has developed an integrated, interactive visualization framework that visualizes

*e-mail: {ko|safzal|yang260|jchae|amalik|ebertd}@purdue.edu
†e-mail: {simon.walton|min.chen}@oerc.ox.ac.uk
‡e-mail:jangy@sejong.edu

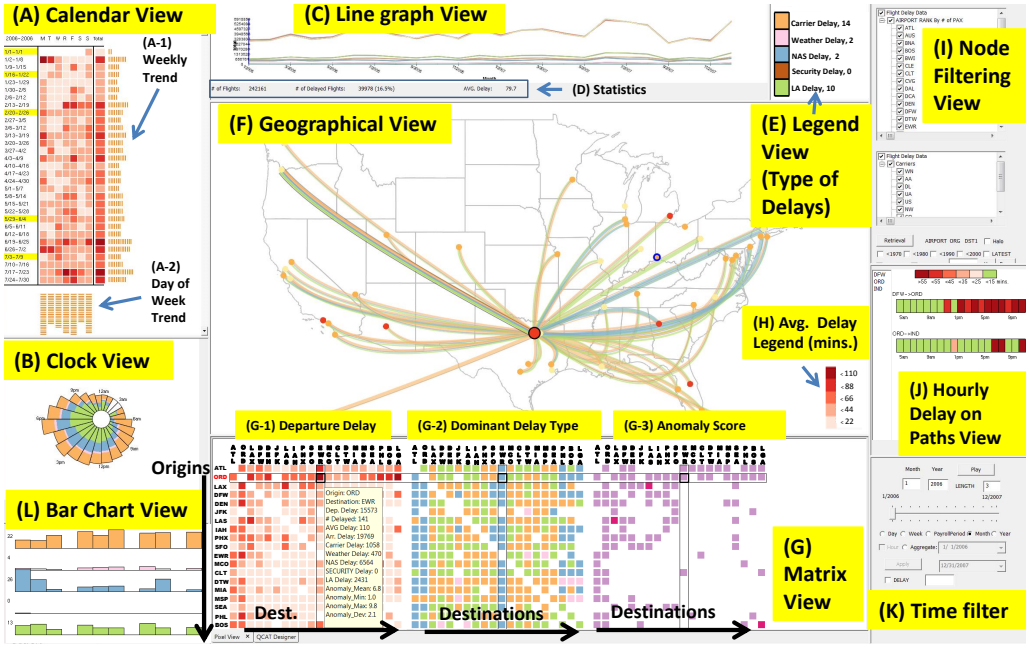


Figure 1: Our system consists of multiple coordinated and linked views: (A) Calendar view, (B) Clock view, (C) Line graph view, (D) Statistics, (E) Legend view for displaying types of delays, (F) Geographical view, (G) Matrix view, (H) Legend view for delay type and time, (I) Node filter, (J) Pattern on itinerary view, (K) Time and aggregation filter, and (L) Twitter tag cloud view. In the (H) legend, the darker the red, the longer the average delay is. A route from Dallas (DFW) to Portland (PDX) is specified in (F), and the top 20 airports in terms of delays are visualized in (G) for explanation. In (G-3), the red links have the highest level of Z-scores, while the purple links have the second largest level of Z-scores.

major flow structures and multivariate relations at the same time. SeeNet [8] visualizes geospatial network data in a communication industry; however, its visualization focuses on univariate data. In contrast to the previous work, our system allows users to analyze all combinations of spatial, temporal, multivariate, and network characteristics simultaneously. Herman et al. [22] surveyed other network visualization techniques beyond our paper's scope.

In order to visualize multivariate data, and to display the maximum amount of data relative to the available screen space, a pixel-based visualization was developed by Keim et al. [23]. In the pixel-based visualization, each data attribute is assigned to a pixel, and a predefined color map is used to shade the pixel to represent the range of the data attribute. Thus, the amount of information in the visualization is theoretically limited only by the resolution of the screen. Borgo et al. [9] present how the usability of the pixel-based visualization varies across different tasks and block resolutions while Ko et al. [25] demonstrate the effectiveness of pixel-based visualization in the task of analyzing corporate competitive advantages. Unlike the pixel displays, the matrix displays assign fewer nodes on both axes of a matrix, and the relational attributes of two nodes are visualized in a link location where the two nodes meet. Matrix displays have been widely used for network visualization due to their effectiveness in providing an overview of the connections in dense networks [14, 16, 21]. Our system utilizes both pixel and network displays, not only to visualize multivariate data (e.g., airports–airlines), but also to describe the network (e.g., airports–airports). We use the term "link" for matrix displays that corresponds to "a pixel" in the pixel displays. Heatmaps present attributes through different shadings of rectangular tiles in a data matrix [45]. We use the heatmap shading approach in the calendar representation [44] that is incorporated in our system.

To help users visually explore multivariate data, many systems have been developed in research and commercial areas [47] (e.g., Spotfire [6], QlikView [4], and Tableau [5]). Common among these systems is that they make extensive use of interactive techniques for brushing, linking, zooming, and filtering to refine the user's queries. Of the systems, Tableau [5], which has become popular due to its flexible operation, allows analysts to easily access and effectively analyze their data [47]. Although multivariate and time-series data analysis is possible in the tool, comparison among multivariate, spatial-temporal, and network-based attributes with geographical components is not well supported by Tableau. In our system, all attributes and characteristics in the data are incorporated and visualized using multiple linked views for simultaneous comparisons. For visualizing multivariate data, Duffy et al. [13] use a glyph encoding some 20 variables while Scheepens et al. [35] focus on a method for reducing visual clutter and occlusion among glyphs.

Lee and Ziang [27] provide an overview of using information-theoretical measures for anomaly detection, including entropy, conditional entropy, information gain, and information cost. A number of case studies are also provided in the domain of network security. Chandola et al. provided a comprehensive survey on methods for anomaly detection [11]. Arackaparambil et al. [7] use information theory to monitor network streams for anomalies in network traffic, and to explore the challenges of providing a scalable implementation using a distributed approach to computing entropy and conditional entropy. Kopylova et al. [26] investigate the use of mutual information in network traffic anomaly detection using Rényi entropy rather than the traditional Shannon entropy measure.

3 Multivariate Network Visualization

To effectively reveal as many aspects of the data characteristics as possible, we explore the data in a series of linked visualizations. Fig. 1illustrates how our system provides comprehensive multivariate network information in multiple linked views. For illustration, we use a flight delay network dataset [1] as an example of multivariate geospatial network data, but any multivariate network data can be populated into our system. Multivariate network information is provided in the geographical view (F) where any operational variable can be used for coloring the node (e.g., anomaly score). The user can explore the data in either a matrix view or a parallel coordinate view (G). Note that (G) has two tab views at the

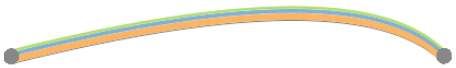


Figure 2: *Thread* example, showing a single link with multiple *threads*. Width of each *thread* within a link is adjusted based on the contribution of each variable. Contribution of each variable in this example is as follows: Variable 1 (Orange) = 0.5, Variable 2 (Blue) = 0.3, Variable 3 (Green) = 0.2.

bottom, and a parallel coordinate view example in (G) is shown in Fig. 7. Similarly, time-varying variables (e.g., delays) are presented in different linked visualizations for efficient exploration in the calendar view (A), clock view (B), and line graph view (C). In the bar chart view (L), the user can interactively compare 1) five delays in all petals, 2) five delays in a petal pie wedge, and 3) five delays for an origin-destination pair *thread*. The height of the bars is normalized and the numeric delay information (longest) is presented in (L). The hourly delay of paths view (J), is designed to allow users to explore attributes in a series of nodes on the paths that a user specifies. With the flight delay data, the user can compare the delays between a direct flight and stop-over flights. As an example, DFW–ORD–IND is shown in (J), where we see that a delay will possibly be maximized if a traveler leaves after 1pm from DFW and between 7pm–9pm from ORD. Users can select airports for analysis in (I) and choose the time in (K). In the system, the line graph view presents temporally aggregated data (e.g., weekly, monthly, yearly). The parallel coordinate view (discussed later in Section 5.2) can be be used to explore the attributes and their value distributions, as well as designing and selecting Query Conditional ATtributes (QCATs, discussed in Section 4) for anomaly detection. Based on characteristics of the data, perceptually appropriate color maps are chosen from both sequential and qualitative color maps from ColorBrewer [20].

3.1 Spatial Multivariate Network Visualization

Unfortunately, a barrier exists in analyzing multivariate network data because visual clutter and complexity often occur in visualizing multiple variables for a node with multiple links between nodes in the map. To reduce such clutter and complexity in the analysis, we design *threads* (see Fig. 2) and *petals* (see Fig. 3) for exploring multivariate link network data. *Threads* connect an origin to each destination and visualize multiple link variables. Because visual clutter around the origin is often generated by link visualization and our *threads*, we also design *petals* to present aggregated and simplified many-to-many network link data. *Threads* and *petals* are designed based on the following requirements for the visualization:

- R1 A visualization should present multiple variables describing the relationships between an origin and multiple destination nodes on the map. Here, users should be able to see an overview of the multivariate relationships and discern at least the largest variable in the visualization for both one-to-one and one-to-many relationships.
- R2 The visualization should provide simplified one-to-many multivariate spatial networks with minimum visual clutter. Use of node rearrangement techniques (e.g., force-based model algorithm [31]) is not allowed to maintain geospatial semantic meanings.
- R3 Users should be able to discern in the visualization for R2 which one-to-many network has the largest aggregate value and which variable has the largest contribution for the largest aggregate value of the one-to-many network.
- R4 Multiple variables describing the statistics for a node should be visually presented.

For goal R1, we design *threads*, and for goals R2–R4 we design *petals*. In the following sections, we explain their visual representations in detail.

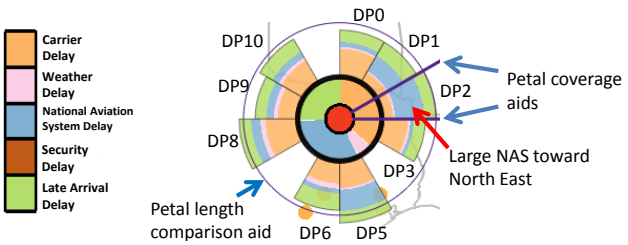


Figure 3: To show the *petal* coverage including destinations, *petal* coverage guide lines are provided (Other coverage aids examples are in Fig. 5 and Fig. 9). For comparison of *petal* lengths, equal-radius circles are drawn on all *petals* as shown. The radius of the circles is the length of the *petal* where a user's mouse is hovering.

3.1.1 Thread Visual Representation

We design the *thread* visualization for representing multiple link variables with a focus on the relationship in an origin-destination pair (R1). Each network link consists of multiple *threads*, and each *thread*'s width is scaled based on a link variable's value. Therefore, each link has the same number of *threads* as the number of link variables, but with varying *thread* widths. While GreenGrid [46] utilizes the force-directed layout [31] and presents a (combined) variable on its links, the *threads* are placed on physical locations and present multiple variables. Users can choose the node variables to be encoded in the *thread* link width. Fig. 2 illustrates an example presenting how link variables can be mapped to *threads*. In this work, we use the departure delay times for each cause of delay as the link variables. This visual representation helps users easily identify which link has the largest delay and which delay type contributes most to the delay. In addition, when a link is specified as an anomalous link, it is located on the top in the stack of *threads* and other links become transparent so that the anomalous link can be highlighted as shown in Fig. 1 (F). To show the direction, an origin node is larger than other destination nodes and has a black outline on the node. Note that Bezier curves are utilized for the link visualizations, and *threads* can be sorted (e.g., departure delays or anomaly scores in our implementation). We incorporate general Bezier curves [32] but we configure the control points of the curves so that long-haul flights tend to be straighter than short-haul curves. In addition, we invert the direction of the normal vectors of the curves alternatively to prevent the case that all control points are moved to one direction in each quadrant. To help user perception, our system provides zooming (with a mouse wheel) and allows users to select the *thread* base width.

3.1.2 Petal Visualization

We introduce *petals*, a new directionally-aggregated radial visual representation as shown in Fig. 3 (Dallas, TX). In this representation, we can provide aggregated directional multivariate network link visualization with minimal visual clutter because we avoid link crossings [8]. Moreover, the spatial and multivariate characteristics are preserved and emphasized. Each directional *petal* (DP) encodes various information between one origin and multiple destinations in a given aggregate direction. Many transportation and logistics problems do have variable variation that is directionally dependent due to transportation paths, weather, routing, etc. By radiating from the origin location to multiple directions (one- to-many), a *petal* presents the geospatial relationships (R2). The *petal* length encodes a selected variable value (R2). Additional variable information is then encoded as radial sections within each *petal* (R3). For example, with the flight delay network data, the average departure delay for the flights heading for airports in a certain radial direction is mapped to the length of the *petal*. Then, the five types of delays are encoded by length (i.e., a segment on a radius) inside the *petal* presenting the contributions of each delay type. Thus, we interpret

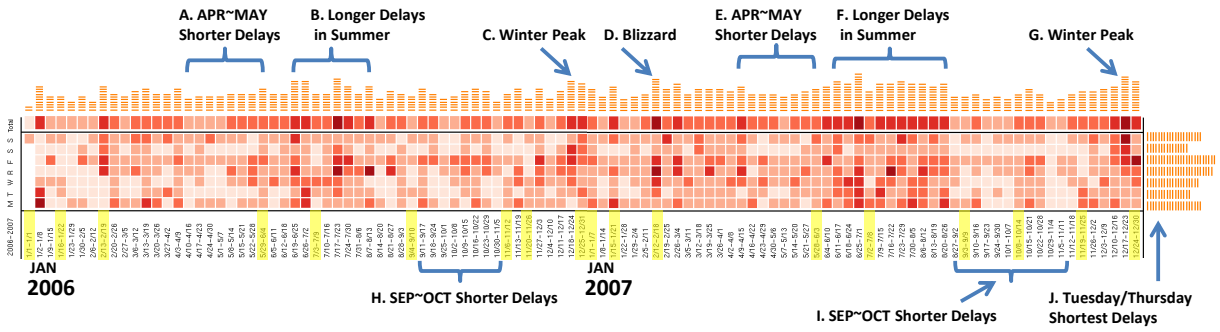


Figure 4: Calendar view showing delay patterns for 2006–2007. In general, there were long delays in the summer and winter seasons, while APR–MAY and SEP–OCT did not have as many delays. Some delays increased around the holidays (highlighted in yellow), but not all holidays had much impact on the delays.

that DP2 in Fig. 3, has a large NAS (National Aviation System, pointed by an red arrow) delay from Dallas. This indicates a large air traffic delay for the destinations, especially toward the airports in New York. Within a *petal*, we insert a pie chart visualization to show comprehensive overviews and comparisons among multiple variables in a node (R4). In the system, users can turn the *petal* display on and off. By default, we assign 12 *petals* for each origin, but users can change the number of *petals*, merge two adjacent *petals* or split one *petal* into many *petals* if necessary. To help users easily recognize the destinations included in a *petal*, our system provides *petal* coverage guide lines as shown in Fig. 3. In addition, when the mouse hovers on a *petal*, the destinations included in the *petal* turn red for better recognition. Lastly to ease comparison of *petal* lengths, equal-radius circles are drawn on all *petals*. The radius of the circle is the length of the *petal* where a user's mouse is hovering (e.g., the radius of the current circle in Fig. 3 is the length of DP2). A tooltip presenting numeric information is provided when a mouse hovers at the center of the *petals*. This can be used for comparing a variable at one location to a variable at another location. Note that the data for the destinations within a *petal*'s coverage are aggregated and visualized together in the corresponding *petal*.

3.2 Network Matrix Displays

Matrix displays have been adapted in various network visualizations because they are effective in providing an overview and relationships of nodes in a dense network. We utilize the matrix displays in our system to provide more complete multivariate network information, as shown in Fig. 1 *(G)*. Our system allows up to three matrices, where the *y*-axis of all matrices are the origins while the *x*-axis represent destinations. For example, for a flight delay network data for 20 airports, we place the departure delay matrix in *(G-1)*, the dominant delay type matrix (e.g., weather, security) in *(G-2)*, and the anomaly Z-score (or standard score, $z = \frac{x-\mu}{\sigma}$ where μ is mean and σ is standard deviation) matrix *(G-3)* from our information-theoretic model as discussed in Section 4. Note that a Z-score filter is applied so that red links have Z-scores larger than 2 (97.7%) and purple links have Z-scores between 1 and 2 (84.1%). In our implementation, users can optionally make *G-3* present additional delay information (e.g., delay by airplane ages and by airlines as shown in Fig. 6 (c) and (d)). When a mouse hovers on a link, a tooltip pops up to display detailed information including delays of different types, the number of flights, and the anomaly scores, as shown in Fig. 1 (G-1). This interaction method is useful when a user wants to find out whether a delay type presented as a dominant type in (G-2) is indeed dominant among all delay types.

3.3 Time Series Displays

In order to present temporal trends, our system provides various time-series views: a calendar view *(A)*, clock view *(B)*, and line graph view *(C)* in Fig. 1. With the calendar representation [44] that applies a calendar metaphor to effectively reveal seasonality and cyclic trends, our system presents the delays by using different shading levels. For instance, the longer delays are presented with darker red. In addition, to help users identify any holiday effect, the week including a holiday has a yellow background. In order to supplement the functionality of the monthly trend line graph, our calendar representation provides additional weekly information on the right side of the calendar *(A-1)* and day of weekly patterns at the bottom of the calendar *(A-2)* in Fig. 1. The clock representation *(B)* is an efficient tool to detect houly trends [17], and we encode variables using areas to enhance visual perception according to Stevens' power law [39]. The line graph view *(C)* presents the types of aggregated delays as well as statistics such as the number of total flights, delayed flights, and average delay time.

4 Anomaly Detection and Highlighting

The visualizations in our system are able to draw upon an information-theoretic model for anomaly detection in a context-sensitive manner, utilizing the anomaly data for a consistent highlighting strategy shown throughout the visualization pipeline. For example, while Fig. 1 *(G-3)* explicitly encodes the anomaly score as the primary visual attribute, Fig. 1 *(F)* focuses on highly anomalous routes with thin outlines. In this case, attribute $a_{origin} = DFW$ (Dallas) is set as the condition in the model. What defines an 'anomalous' record depends upon the user's design and definition of individual anomaly detectors, *QCATs*, discussed in detail in this section. From a visual analytical perspective, these QCATs provide an overview of records where important attributes deviate from usual for specific conditions.

4.1 Overview of Anomaly Detection Method

Chandola et al. provided a comprehensive survey on methods for anomaly detection [11], categorizing them based on the nature of inputs, instance types, algorithmic mechanisms, and forms of outputs. For multivariate network data, we are interested in methods that can:

- Handle multi-dimensional records – because the main flight data concerned is a structured data stream consisting of 29 attribute dimensions (e.g., ≥ 10);
- Address the need for detecting contextual anomalies – which can provide a high-degree of flexibility and accommodating dynamic data and task variations in different detection scenarios;
- Facilitate an unsupervised algorithmic mechanism – alleviating the lack of training data in many situations;
- Generate anomaly scores as outputs that can be effectively conveyed by most visualization techniques.

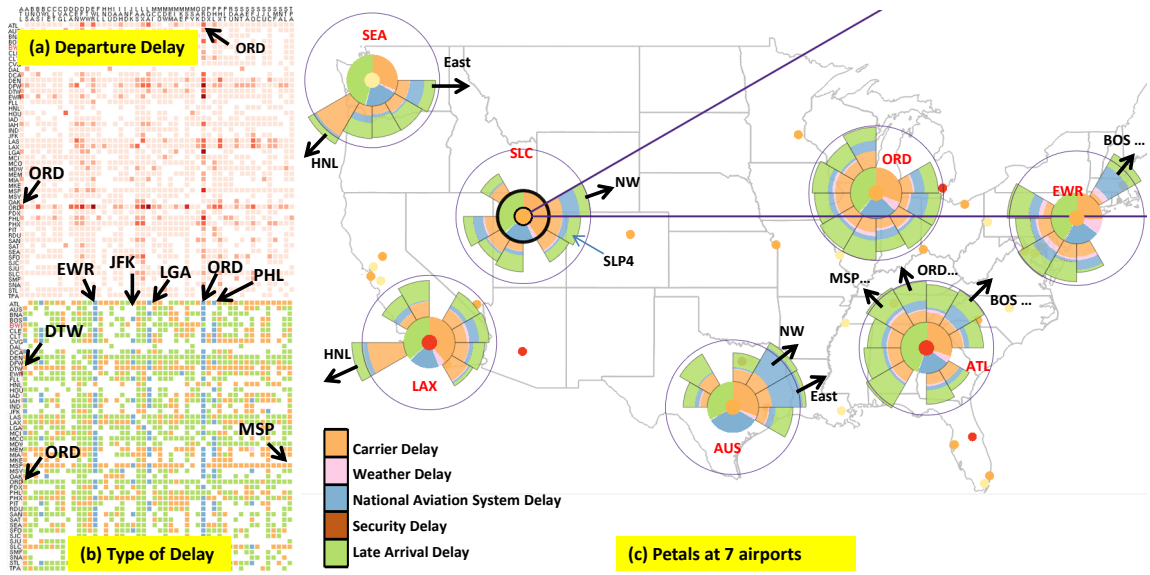


Figure 5: (a) ORD is the most congested airport for both in-bound (vertical) and out-bound (horizontal). It is notable that carrier delay is the prevalent (out-bound) delay for DTW and MSP while NAS delay is the prominent delay for the incoming flights (vertical) in at EWR, JFK and LGA (b). (c) Flights heading to Hawaii from west cost airports in winter had long delays. Flights heading for ORD, ATL and airports from mid-east and east usually suffer from NAS delays.

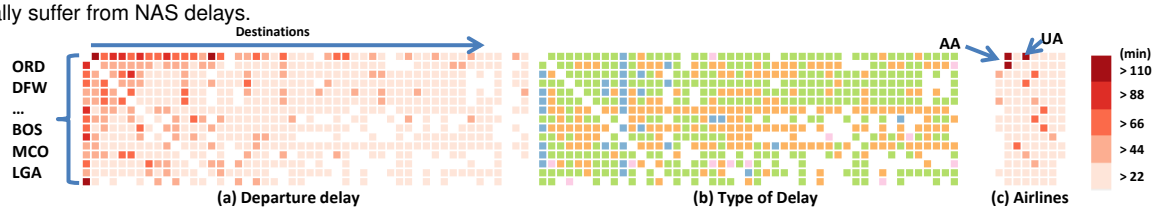


Figure 6: Airports are sorted by delays. ORD shows the longest delays in many out-bound flights in (a). The dominant type of delay was carrier delay and LAD in (b). UA and AA had the longest delays in ORD when ORD was top by delays in (c).

In general, the family of statistical and information-theoretic methods can address the above-mentioned requirements better than the families of classification-, nearest neighbor- and clustering-based methods. As information theory is fundamentally built on probabilistic and statistical measures, information-theoretic methods may also be considered as a subset of the family of statistical methods. In this work, we use an information-theoretic method because of advantages as highlighted in [11]. "(1) They can operate in an unsupervised setting. (2) They do not make any assumptions about the underlying statistical distribution for the data."

Let $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ be a set of n variables. Each data record, $R = \{v_1, v_2, \ldots, v_n\}$ be a n-tuple, where v_i represents a valid value of attribute $\mathbf{a}_i$. In a practical scenario, an attribute, $\mathbf{a}_i$, may have a very large or infinite number of valid values. Binning is normally used to facilitate more accurate estimation of the probability of each valid value. In the following discussion, the probability distribution of an attribute, $p(\mathbf{a}_i)$, is assumed to be estimated in conjunction with an appropriate binning scheme.

The attribute set, $\mathbf{A}$, is divided into three mutually-exclusive subsets, $\mathbf{A}_{cnd}$, $\mathbf{A}_{von}$, and $\mathbf{A}_{ins}$. As anomalies are context-sensitive, $\mathbf{A}_{cnd}$ defines the context of a type of anomaly as a particular condition, such that all attributes in $\mathbf{A}_{cnd}$ are associated with specific values. For example, we may have $\mathbf{a}_4 = 1(Monday)$, $\mathbf{a}_{17} = JFK$, $\mathbf{a}_{18} = LHR$. The attributes in $\mathbf{A}_{cnd}$ are referred to as *conditional attributes*. In some situations, a conditional attribute may also take a range of values, e.g., $\mathbf{a}_4 = 1,2,3,4$ or 5 (Monday–Friday).

The attributes in $\mathbf{A}_{von}$ play the primary role in determining an anomaly score for each record that has met the condition defined by $\mathbf{A}_{cnd}$. These attributes are referred to as *Variants of Normality* (VON). The remaining attributes, which are grouped into $\mathbf{A}_{ins}$, are considered to have "insignificant" influence on the type of anomaly concerned and are therefore excluded in the computation. Such a decision is usually made based on some known factors or logical reasoning by the user.

A combined configuration of $\mathbf{A}_{cnd}$ and $\mathbf{A}_{von}$ in relation to the overall attribute set $\mathbf{A}$, subsequently, determines how anomaly scores are estimated for each record. Given a record R, we first retrieve all records that have the same conditional attribute values as R. Let this collection of records be $R_1, R_2, \ldots, R_W$, where W is usually a very large number. We now consider only the variants of normality defined by $\mathbf{A}_{von} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_s\}$. In conjunction with a binning scheme, each attribute, $\mathbf{x}_j$, may take valid values that are mapped to a set of t_j bins $B_j = \{b_{j,1}, b_{j,2}, \ldots, b_{j,t_j}\}$. For the s attributes in $\mathbf{A}_{von}$, there are a total of: $t_1 \times t_2 \times \ldots \times t_s$ different combinations of bins across different attributes. These combinations collectively define an alphabet $\mathscr{Z}$, and each unique combination is a letter $z \in \mathscr{Z}$.

The selection of an appropriate binning scheme for each attribute $\mathbf{x}_j$ is essential for ensuring that the total number of letters $|\mathscr{Z}|$ is smaller than the total number of records W. Ideally, we have $|\mathscr{Z}| << W$. We can, then, estimate the probability of each letter $z \in \mathscr{Z}$ based on the collection of records $R_1, R_2, \ldots, R_W$, resulting in a probability distribution function $p(z)$. For the given record R, we obtain its probability $p(R)$ by mapping it to its corresponding letter in $\mathscr{Z}$. The level of self-information is $I(R) = -\log_2(p(R))$, which is also called *surprisal*. We use this surprisal value as the anomaly score for the given record R. The level of uncertainty of this score can be defined as $H(\mathscr{Z})/\log_2(|\mathscr{Z}|)$, where $H(\mathscr{Z})$ is the entropy of the alphabet $\mathscr{Z}$.

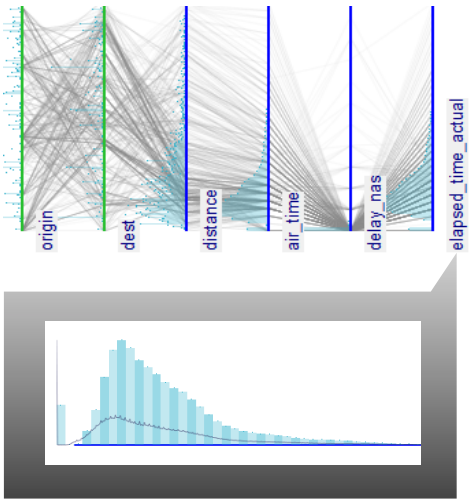
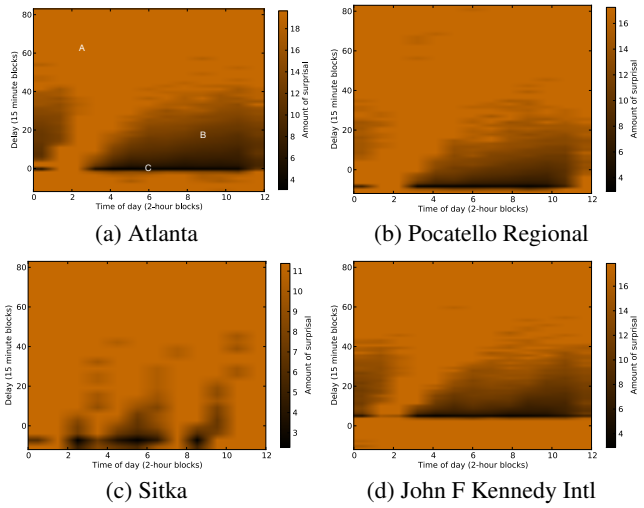


Figure 7: Using Parallel Coordinates to Design QCATs: *(top)* Exploring the attributes as a parallel coordinate plot; *(bottom)* Specifying an individual attribute's bin specification



(a) Atlanta (b) Pocatello Regional

(c) Sitka (d) John F Kennedy Intl

Figure 8: Heatmaps representing the surprisal spaces of flights leaving four different airports, with *(x)* Time of day *(bin size: 2 hrs)*, and *(y)* Departure delay *(bin size: 15 mins)*

It is necessary to emphasize that the anomaly score obtained for R reflects only the type of anomalies encoded by the specific configuration of $\mathbf{A}_{cnd}$ and $\mathbf{A}_{von}$. Hence, each configuration is only for queries of a specific type of anomaly in a particular context. We call each configuration a QCAT (**Q**uery **C**onditional **AT**tributes). It is not difficult to see that a visual analytics system can be equipped with one or more QCATs. For a given record, scores obtained using different QCATs can be aggregated, though it is necessary to understand the semantic implication of combining different QCATs and the difference between different aggregation methods (e.g., mean or max). Section 5.2 discusses the workflow for working with QCATs in a visual analytical system.

The information-theoretic method for anomaly detection is not an algorithm in a traditional sense. Using this approach, anomalies are defined mathematically based on the probability of events captured by the historical data. So in relation to this definition of anomaly, the probabilistic ranking of events using the method is always correct. On the other hand, machine learning methods mostly use a different definition, where an event is anomalous if it is subjectively annotated as an anomaly. So the goal of a learned algorithm is to mimic human perception of an anomaly. One cannot compare the accuracy of these two methods directly. For qualitative comparison, refer to the survey by Chandola et al. [11], where a few other approaches are considered. The mathematics is not new in this algorithm [12,41] but to the best of our knowledge, this kind of probabilistic measures have not been used in visualization, or for the flight data.

4.2 Implementation & Scalability

We have conducted a series of tests on the scalability of QCATs. Two implementations, client- and server-based, have been developed using PostgreSQL [3]. The former performs the grouping and aggregation on the client (i.e, in native code), and the latter uses a stored procedure hosted by the database server. Both server- and client-based implementations show that QCATs are linearly scalable in relation to the number of records used in the computation; the server-based implementation is about 2.5 times faster than the client-based implementation. Additionally, the client implementation is more sensitive to the network bandwidth and latency to the database server.

In our scalability tests, we have found that the performance of the server-based solution can be seriously affected by the number of VONs in $\mathbf{A}_{von}$, while the client-based implementation shows steady linear scalability in relation to the increasing number of VONs. The largest factor is the amount of shared buffers provided to PostgreSQL. The scalability of entropy computation is linear but does rely on recomputing past data due to updated probability masses. However, Arackaparambil et al. [7] show that a distributed method for conditional entropy computation is feasible, while Guba et al. [18] demonstrate entropy estimation in streaming insert-only datasets. In the following sections, we describe how our system presents multivariate network data and visualizes the detected anomalies.

5 Geospatial Multivariate Network Data Exploration

As an example, we will use US domestic flight delay data from the Bureau of Transportation Statistics (BTS) [1] where each data row provides information for an individual flight including origin, destination, day of week, day of month, scheduled (departure/arrival) time, and real (departure/arrival) time and type of delay. There are five types of delays. Carrier delay is a problem within the airlines' control including mechanical problems of aircrafts, while NAS delay is caused by the control of the National Aviation System (NAS) including heavy traffic volume. Late Arrival Delay (LAD) is caused by the late arrival of the same aircraft at a previous airport. Security delay includes re-boarding time due to security breach and waiting time at the screening equipment. Weather delay means delay caused by extreme weather conditions at point of departure or arrival. Note that NAS delay and Security delay might be caused by the government organizations, while Carrier delay and LAD are caused by the airlines. We use the top 50 airports according to the number of passenger boardings that encompasses FAA's OEP-35 (Operational Evolution Partnership 35) airports accounting for more than 70% of the entire number of passengers [2].

5.1 Flight Delay Network Exploration

In this section, we explore the flight delay network data from 2006-2007 and summarize delay patterns in terms of temporal (e.g., summer, winter, holidays, weekly, hourly, and day of week) and spatial effects including special conditions such as severe weather (e.g., blizzards). First, we use the calendar view to investigate data patterns. In Fig. 4, we can see long delays as prominent seasonal patterns in the summer (B, F) and winter (C, G), while shorter delays were recorded during April–May and September–October. Another visible pattern is that there were fewer delays on Tuesday and Saturday in (J). We find that the patterns are related to holidays that are

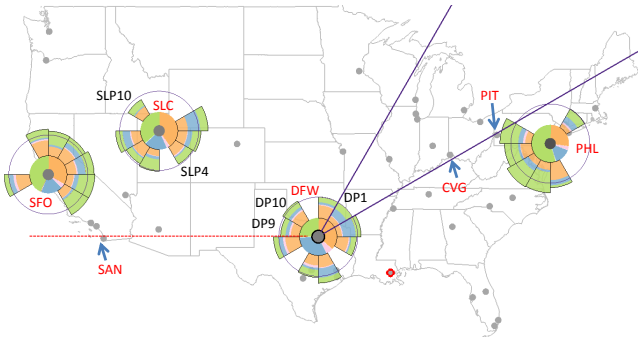


Figure 9: An example for a *petal* experiment. With the visual aid, users could better tell that CVG is included in DP1 while PIT is included in DP2.

concentrated in summer and winter (e.g., Independence Day in July, Christmas in December, personal vacations) but long delay patterns are not indicated for Martin Luther King Day in January and Labor Day in September. Moreover, long delay patterns tend to increase in 2007, especially in the summer (B and F). Also, there is a sudden spike (D) shown with the darkest red that might be another point for investigation.

Next, we can explore the aggregated delays for two years in the matrix view as shown in Fig. 5 (a, b), where we see some interesting patterns. The most prominent pattern is the series of horizontal and vertical dark red links (long delays) generated at the Chicago O'Hare Airport (ORD) in (a), which indicates that both in-bound (horizontal) and out-bound (vertical) flights were severely congested. We also observe that such delays in ORD were caused mainly by late arrivals of aircraft (horizontal green line) shown in (b). In addition, we notice that there are five distinguishable vertical blue lines in the matrix (b) and four of them (EWR, JFK, LGA, and ORD) were regulated by the High Density Rule (HDR) enacted in 1969 by the FAA due to severe congestion. This may indicate that the rule might not be strong enough to prevent such long delays. The delays in DTW (Detroit) and MSP (Minneapolis), which are two of the biggest hubs of Delta Airlines, are not very long compared to those in other top congested airports. However, it is interesting that the major type of delay is carrier delay (orange) caused by the airline itself.

Since one of the highest delays is observed in winter as shown in Fig. 4, we use our *petal* visualization with winter seasonal data for finding patterns and types of delays in the network as shown in Fig. 5 (c). We can select as many *petals* as designed for the exploration as long as minimal visual clutter is maintained. One interesting finding is that the flights heading for HNL (Hawaii) from the west coast airports (SEA and LAX) have relatively long delays (e.g., 120 minutes on average) and the prevalent cause for the delay is carrier delay. Moreover, those airports also have relatively long NAS delays for flights heading for north-east destinations (ORD, and airports around New York).

The next interesting aspect is the delay distribution by time as shown in Fig. 1 (B) in the proportional mode with area encoding for each delay type. Here, we see a trend showing that delays increased from 6 am and had a peak around 6 pm. It is noted that this is the same pattern shown in the late aircraft delay while other types retained their proportion. This suggests that delays propagate during the day, a problem that Mazzeo termed "cascading delays" [29]. Such trends may imply that delays might be effectively reduced because these delays can be controlled either by the airlines (carrier delay/late arrival delay) with enough of an interval or layover time between two consecutive flight schedules, or by a government agency (e.g., Federal Aviation Administration) with advanced systems for air traffic control. *Threads* can be a good means for understanding delay patterns, as well as the concentration of delays and

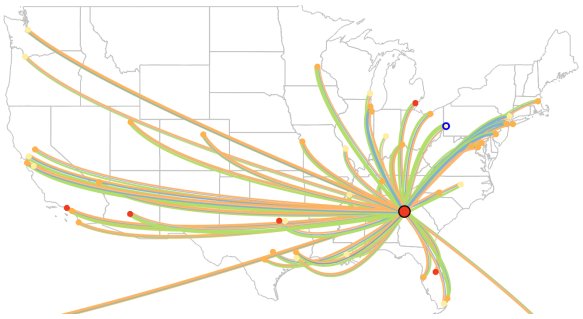


Figure 10: An example of a *thread* experiment. 40% of the participants answered incorrectly that green was the prevalent delay due to the severe color concentration around the origin.

complications at hub airports. Figure 1 (F) presents an example of the complicated network status at DFW, a hub of two major airlines in *Threads* (April 2012). Here we see that the airport has many connections with airports across the U.S., and the major type of delay is late arrival delay (green) and carrier delay (yellow). In addition, the flights heading for New York City suffer from NAS (National Aviation System) delay.

Of primary interest are the patterns in the length and types of delays that can be can be better explored by sorting airports. We see that the ranks change with little variation based on seasons, but most delays are caused by major airports including ORD (Chicago), ATL (Atlanta), LGA (New York City), EWR (New York City), DTW (Detroit), LAX (Los Angeles), LAS (Las Vegas), and DFW (Dallas) as shown in Fig. 6 (a). From the type matrix Fig. 6 (b), we notice that in many highly-ranked airports, the main type of delay is the late arrival delay in busy travel seasons while the NAS delay is dominant at other times. This implies that the NAS might not be properly adapting to the current increasing traffic in terms of delays. On the other hand, we notice that the two distinguishable airlines causing delays are AA (American Airline) and UA (United Airline) in the two most delayed airports as shown in Fig. 6 (c). The dominant delay type matrix in Fig. 6 (b) indicates that the airlines are responsible for solving the delay problem because the dominant types of delays were carrier delay and late aircraft arrivals.

5.2 QCAT Workflow

As discussed in Section 4, our system features an information-theoretic anomaly detection system that is comprised of a set of user-defined QCATs. The design of a QCAT can be based on a specific hypothesis, or as a more general monitoring system for one or more attributes. Ideally, in a deployed system, the roles of QCAT *designer* and overall *analyst* would be disparate, with the analyst analyzing the data for anomalies and reporting back to the designer to refine the QCATs based on new trends.

To assist the user in defining the QCATs in the system, we provide a design tool based on parallel coordinates (see Fig. 7 *(top)*) while the user is able to explore the attribute space by adding/removing attribute dimensions, observing their value distributions (e.g., probability mass functions), as well as viewing the record relationship between attributes afforded by a standard parallel coordinates representation. The role of an attribute can be toggled between conditional (green) and VON (black) using the right mouse button. The user is also able to explore an individual attribute in more detail by clicking the left mouse button that expands the attribute to the full view to show its distribution in more detail (Fig. 7 *(bottom)*). The detail view also shows the attribute's bin width specification, which can be modified per QCAT. The user's choice of bin width has an effect on the anomaly results and reflects the user's knowledge of the attribute's semantic meaning. The system maps data types to suitable bin width granularities automatically. For example, timestamp datatypes are divided into bins of n

Petal Index	Difference (%)	Accuracy (%)	Time (s)
DP1	4 (Small)	76.7	8.2
DP10	21 (Large)	100	3.7

Table 1: Participants found the longest delay inside a *petal* more accurately in less time as the difference became larger (HP1).

Petal Index	Difference (%)	Accuracy (%)	Time (s)
DP9	3 (Small)	46.7 (83.3)	6.6 (5.4)
SLP10	12 (Large)	96.7 (100)	3.9 (2.6)

Table 2: As the difference became larger, the participants better detected the shortest delay (HP2). Visual aids improved both the accuracy and efficiency (HP5).

minutes; categorical data such as strings are unbinned. Since integer types may represent categorical, interval or ratio measurements, we assume a default bin width of 1 and let the user decide upon a more suitable width.

Once the user has defined a QCAT, it can be saved to the QCAT library and selected as the active QCAT. Anomaly-supporting visualizations in our system such as the network matrix view update to reflect the anomaly scores by completing the relevant conditionals in the QCAT (i.e., origin and destination pairs) and executing the QCAT on the data to obtain statistics (i.e., mean, max, variance) on the surprisal values for records matching the conditionals. Our system by default displays the maximum surprisal value as the anomaly value mapped to a visual attribute (i.e, outline on *threads*) in the visualization. The anomaly values in the visualizations guide the user to identify abnormal flights based on their own criteria specified in the design of each QCAT. Anomalous results can then be explored further using the available visual analytical tools to understand why the anomaly value was high and report these findings to the QCAT's designer.

For a QCAT consisting of two VONs, we can illustrate the anomaly distribution using a heatmap. Fig. 8 shows the anomaly space for flights leaving four different airports for the years 2006 and 2007. The x-axis shows the time of day (morning) divided into two-hour blocks, and the y-axis shows the amount of delay in 20-minute blocks (notice that flights can leave early). Areas of low surprisal value are black and become amber with higher surprisal values. It is clear that for this airport, flights around 4AM are uncommon, and the amount of delay seems to increase steadily throughout the day until late afternoon before leveling out. For the Atlanta airport, three example records, A, B and C are shown of high ($\approx$19.68), slightly above average ($\approx$14.36), and low ($\approx$3.478) surprisal values, respectively. Investigating these flights using *threads* shows that late aircraft were largely to blame for both A and B; however, in the case of A the high surprisal value indicates that such a large delay is unusual at this time of the morning. At C, we find ourselves in the 'usual' low-anomaly area for this airport, where delays are close to zero for most of the day.

A professional analyst from an industry-leading company that deals with flight delay data evaluated our system and our approaches used in this work. The analyst mentioned that, at this company, they do not have such visual tools that can enable visual analysis of multiple variables at different locations and different times. Therefore, our system is excellent for dealing with challenging data in the flight delay domain, and it is cutting-edge work for the industry. In particular, the information theory based anomaly detection approach is very intriguing, and it has not been applied to analyses in the industry as of today. Lastly, the analyst suggested visualizations of correlations and propagation of delays (or cascading delay) as key properties of an interconnected network to enhance our system because such visualizations allow for a form of root-cause analysis to help analysts see what is driving delays in the network and what is happening to the delay debt.

Petals	Diff. (%)	Accuracy (%)	Time (s)
SLC+SFO+PHL	2 (Small)	36.7 (73.3)	10.9 (6.8)
SLC+SFO+PHL	12 (Large)	96.7 (96.7)	4.7 (5.1)

Table 3: Users had difficulty finding the longest delay among distant *petals* with a small (2%) difference (HP3). The visual aid helped the users better answer with a small difference (HP5).

6 User Study

In order to evaluate the *petal* and *thread* designs, we performed a user study with 30 participants recruited from various majors at our university. In the study, the participants were given computer-based tasks for verifying hypotheses. Various difference levels in the flight delay network data were used in the tasks. Note that the *difference level* in this section means the difference between the longest (shortest) and second longest (shortest) delays. Note that the numbers in parentheses in Table 2, Table 3, and Table 4 are the results with visual aids. We use a paired t-test to check if our experimental result obtained is significant (p-value < 0.05) within a 95% confidence interval.

6.1 Petal User Study Results

We first set up the following hypotheses for the *petals* visualization as follows:

HP1 As the difference becomes larger, users will show high accuracy and speed in detecting the longest delay inside a *petal*,

HP2 As the difference becomes larger, users will show high accuracy and speed in finding the shortest (or longest) delay among *petals* for one operational place (e.g., airport),

HP3 Users will show lower accuracy in finding the shortest (or longest) *petal* among the *petals* at multiple operational places,

HP4 Users will show low accuracy and speed in finding whether an airport is included in *a petal* as the distance between the *petal* and the airport becomes longer and as an airport is close to the boundary of the *petal*, and

HP5 Visual aids will improve accuracy and speed.

TASK1 for verifying HP1 asked the participants to choose the longest delay inside a *petal* in 2 locations: DP1 (delay difference: 4%) and DP10 (21%), as shown in Fig. 9. The participants showed higher accuracy and speed as the difference increased (Table 1, p-value < 0.05). In TASK2, for verifying HP2, the participants were asked to select the shortest *petal* in 2 locations: DFW (3%) and SLC (12%). For a small difference (3%), 46.7% of the participants answered correctly. As the difference became larger and the visual aid (circle) was provided (HP5), both accuracy and speed were improved (Table 2, p-value < 0.05). TASK3 was the same as TASK2 but multiple *petals* at Salt Lake City (SLC), San Francisco (SFO), and Philadelphia (PHL) were presented concurrently. Here, the participants showed lower accuracy (from 46.7% to 36.7%) and slower speed (from 6.6s to 10.9s) compared to the results in TASK2. The visual aid (HP5) improved both accuracy and speed in the lower difference (Table 3, p-value < 0.05). In order to evaluate if users accurately recognized the coverage of each *petal* (HP4), TASK4 asked the participants to select airports that were included in DP1 and DP9, as shown in Fig. 9. As summarized in Table 4, the participants showed low accuracy (23.3% and 60%). The main reason for such low accuracy was that it was hard for them to find whether CVG (Cincinnati) and PIT (Pittsburgh) were included in DP1. In the same context, only 60% of the participants correctly found that SAN was not included in DP9. However, with the visual coverage line (HP5), both the accuracy and speed improved.

Petal Index	# of Airports	Accuracy (%)	Time (avg.)
DP1	8	23.3 (83.3)	1.96 (1.18)
DP9	8	60.0 (93.3)	2.3 (1.3)

Table 4: The participants had difficulty finding whether CVG and PIT were included in DP1 (HP4). The visual aid helped the users better recognize if an airport was included in a *petal* or not (HP5).

Difference (%)	Accuracy (%)	Time (s)
3.1 (Small)	66.7	5.2
28 (Large)	100	2.9

Table 5: The participants made errors and spent more time finding the longest delay when the difference in *threads* was small (HT1).

6.2 Thread User Study Results

Next, we set up hypotheses for the *threads*. As the difference between the longest and the second longest delays becomes larger, the users will produce better results in HT1) detecting the longest delay inside the *threads*, and in HT2) choosing the most prevalent delay among all the *threads*. TASK5 for verifying HT1 asked the participants to select the thickest *thread* for small (3.1%) and large (28%) difference levels. As summarized in Table 5, when the difference was small (3.1%), it was hard for the participants to tell the longest delay (66.7% accuracy). On the other hand, when the difference became larger, they answered very accurately and spent less time (p-value < 0.05). TASK6 for verifying HT2 asked the participants to tell the longest delay when all the *threads* were considered. Here we see a similar result as in TASK5: the larger the difference, the higher the accuracy and the slower the speed (Table 6). In TASK6, we had an interesting result showing that special concentration of a color may interfere with accurate visual perception. For example, we can see LAD (green) is concentrated on short-haul routes as shown in Fig. 10. In this case, 40% of the participants thought that LAD was the longest delay for flights leaving from Atlanta, but in fact the carrier delay was 23% larger than LAD. This error rate is unexpected compared to the result in TASK5 where the participants showed higher accuracy and speed with a similar difference (28.6%). Conversely, we think it is possible that users could assume that the color on long-haul routes has the largest value if the color is concentrated in long-haul *threads*. To prevent this, our system provides numeric information in the legend view that users can refer to, as shown in Fig. 1 (E).

7 LIMITATIONS AND DISCUSSION

Petals have a similar appearance to the rose or sunburst diagrams that have been adapted in various contexts [15, 30, 37]. The contribution of *petals* lies in extending the usability of the family of the rose diagram by allowing geographically-directional, multivariate, and aggregated network analysis simultaneously. Discerning widths of *thread* can be hard when each variable has similar values or when a unit *thread* within a route is not thick enough for visual perception. In addition, when a color is concentrated on long-haul or short-haul routes, it could be hard to select the largest value among all *threads*. In these cases, a line with a superimposed histogram can be utilized. To help users with these issues with threads, our system provides interactive bar charts and the numeric variable information in the legend view when a user specifies an area of *threads* (aggregated) and in a tooltip when the user's mouse hovers over an airport (origin to destination). The tooltip in the matrix view can be used for verifying that the presented dominant delay (Fig. 1 (G-2)) is indeed dominant compared to others. The scalability of threads can be limited by two factors: the number of variables and the links. In our system, the number of links can be adjusted by the on/off function in threads and the provided network matrices can complement the link analysis. Our user study implies that a thread with 30% larger value than the others can be distinguished from the others. However, a fundamental issue when a large number of variables is used becomes how many colors a human can distinguish and which colors should be used. Harrower et al. suggest 12 distinguishable colors [20] but a lower number of colors would be effective for the *threads* due to the difficulty in comparing widths.

Difference (%)	Accuracy (%)	Time (s)
11 (Small)	90	5.3
28.6 (Large)	100	2.9
23 (Large)	60	5.1

Table 6: 40% of the participants answered incorrectly with a large difference (23%) in finding the prevalent delay among all *threads*. This may indicate that color concentration on long-haul or short-haul *threads* interferes with visual perception.

8 CONCLUSION AND FUTURE WORK

We have explored complex multivariate network links with multiple tightly-integrated interactive visualizations. We have introduced two new visual representations, *petals* and *threads*, for spatial multivariate link visualization. Our sortable matrix displays have the ability to represent multiple origin and destination pairs, while the linked line graph, calendar, and clock views give opportunities to find temporal characteristics. An information-theoretic anomaly detection model was introduced based on conditional attributes, with the visualizations in the system utilizing the surprisal values for visual highlighting of anomalies in multiple visualization components in a unified manner.

It has several benefits compared to previous systems. Our system allows users to investigate the data status of a large number of operational locations by simultaneously observing various data characteristics at both aggregate (entire network) and detailed levels (e.g., origin-destination pairs) using our multiple linked view. Our new visual representations, *petals* and *threads*, help users find features of multiple spatial network variables with minimum visual clutter; the network matrices aid in analyzing the entire network in terms of multiple origin-destination pairs as well as origin-attribute pairs. Seasonal and cyclical trends can be efficiently detected in the calendar, line graph, and clock visualizations from our system. Lastly, our system provides an information-theoretic model for detecting anomalies based on conditions. For the evaluation of our system, we presented an example using flight delay network data from the top 50 airports to illustrate the use and potential of our designs and the user study results.

Our system can be easily applied to analysis with any other multivariate spatiotemporal, network-based data such as transportation and logistics, trading, and communication industries [8]. As a future work, we plan to incorporate the ability to help users find correlations using *petals* and *threads*. The capability for visualizing cascading effects and clusters of operational places that have the same characteristics will also be investigated. We also plan to use actual routes to enable comparison with length of flights. In addition, we would like to explore our anomaly detection more by investigating methods of combining the anomaly values for groups of QCATs.

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2009-ST-061-CI0001-06. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. Jang's work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1011170).

REFERENCES

[1] Bureau of Transportation Statistics (Accessed 20 Mar 14. http://www.rita.dot.gov/.

[2] Operational Evolution Partnership 35. http://aspmhelp.faa.gov/index.php/OEP_35.

[3] PostgreSQL (Accessed 11 Jun 14. http://www.postgresql.org/.

[4] Qlikview. http://www.qlikview.com/.

[5] Tableau. http://www.tableausoftware.com.

[6] C. Ahlberg. Spotfire: An information exploration environment. *ACM Special Interest Group on Management of Data Record*, 25(4):25–29, 1996.

[7] C. Arackaparambil, S. Bratus, J. Brody, and A. Shubina. Distributed monitoring of conditional entropy for anomaly detection in streams. In *Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, pages 1–8, 2010.

[8] R. A. Becker, S. G. Eick, and A. R. Wilks. Visualizing network data. *IEEE Transaction on Visualization and Computer Graphics*, 1(1):16–21, Mar. 1995.

[9] R. Borgo, K. Proctor, M. Chen, H. Janicke, T. Murray, and I. Thornton. Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):963–972, 2010.

[10] D. Brooks. What data can't do. *The New York Times*, Feb. 2013.

[11] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys.*, 41(3), July 2009.

[12] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[13] B. Duffy, J. Thiyagalingam, S. Walton, D. J. Smith, A. Trefethen, J. C. Kirkman-Brown, E. A. Gaffney, and M. Chen. Glyph-based video visualization for semen analysis. *IEEE Transactions on Visualization and Computer Graphics*, 99:1, 2013.

[14] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete. ZAME: Interactive large-scale graph visualization. In *PacificVis*, pages 215–222. IEEE, 2008.

[15] N. Elmqvist, J. Stasko, and P. Tsigas. Datameadow: A visual canvas for analysis of large-scale multivariate data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 187–194, 2007.

[16] J.-D. Fekete. Visualizing networks using adjacency matrices: Progresses and challenges. In *CAD/Graphics*, pages 636–638. IEEE, 2009.

[17] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2013.

[18] S. Guha, A. Mcgregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742, 2006.

[19] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048, 2009.

[20] M. A. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting color schemes for maps. *Cartographic Journal*, 40(1):27–37, 2003.

[21] N. Henry, J.-D. Fekete, and M. J. McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.

[22] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. In *IEEE Transactions on Visualization and Computer Graphics*, volume 6 (1), pages 24–43. 2000.

[23] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[24] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann, 2010.

[25] S. Ko, R. Maciejewski, Y. Jang, and D. S. Ebert. Marketanalyzer: An interactive visual analytics system for analyzing competitive advantage using point of sale data. *Computer Graphics Forum*, 31(3):1245–1254, 2012.

[26] Y. Kopylova, D. Buell, C.-T. Huang, and J. Janies. Mutual information applied to anomaly detection. *Communications and Networks, Journal of*, 10(1):89–97, 2008.

[27] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143, 2001.

[28] Z. Liu, S. B. Navathe, and J. T. Stasko. Network-based visual analysis of tabular data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 41–50, 2011.

[29] M. Mazzeo. Competition and service quality in the u.s. airline industry. *Review of Industrial Organization*, 22(4):275–296, June 2003.

[30] F. Nightingale. *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison and Sons, 1958.

[31] A. Noack. An energy model for visual graph clustering. In *Graph Drawing*, volume 2912 of *Lecture Notes in Computer Science*, pages 425–436. Springer, 2003.

[32] S. M. Peter Shirley, Michael Ashikhmin. *Fundamentals of Computer Graphics*. A K Peters, 2009.

[33] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.

[34] A. Z. Santovena. Big data : evolution, components, challenges and opportunities. Master's thesis, Massachusetts Institute of Technology, Sloan School of Management, 2013.

[35] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *IEEE Symposium on Pacific Visualization*, pages 17–24, 2014.

[36] D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363, 2011.

[37] Z. Shen and K.-L. Ma. Mobivis: A visualization system for exploring mobile data. In *IEEE Symposium on Pacific Visualization*, pages 175–182, 2008.

[38] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):733–740, 2006.

[39] S. S. Stevens. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. Wiley, 1975.

[40] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.

[41] M. J. Usher. *Information Theory for Information Technologists*. Computer Science. Macmillan, 1984.

[42] F. van Ham, H.-J. Schulz, and J. M. DiMicco. Honeycomb: Visual analysis of large scale social networks. In *Proceedings of International Conference on Human-Computer Interaction*, volume 5727 of *Lecture Notes in Computer Science*, pages 429–442. Springer, 2009.

[43] Wattenberg, Martin. Visual exploration of multivariate graphs. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, volume 1 of *Visualization 1*, pages 811–819, 2006.

[44] J. V. Wijk and E. V. Selow. Cluster and calendar based visualization of time series data. In *1999 IEEE Symposium on Information Visualization (INFOVIS '99)*, pages 4–9, Oct. 1999.

[45] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[46] P. C. Wong, K. Schneider, P. Mackey, H. Foote, G. Chin, Jr., R. Guttromson, and J. Thomas. A novel visualization technique for electric power grid analytics. *IEEE Transactions on Visualization and Computer Graphics*, 15(3):410–423, May/June 2009.

[47] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstädt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. A. Keim. Visual analytics for the big data era - A comparative review of state-of-the-art commercial systems. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 173–182. IEEE Computer Society, 2012.

UKVAC II Joint Research Project

Summary Report on WP1: Information-Theoretic Framework

Min Chen and Simon Walton, University of Oxford (17 December 2013)
min.chen@oerc.ox.ac.uk, simon.walton@oerc.ox.ac.uk

Overview

This joint research project was first proposed in 2010 involving five universities: Bangor, Imperial, Middlesex, Swansea and UCL. It was coordinated by Middlesex, and the Swansea team transferred to Oxford in 2011 following Min Chen's appointment at OeRC. The joint project was intended as an innovative research exercise to tackle two 'grand challenge' problems by thinking out-of-the-box. The two challenges are the Nobel Laureates problem and the flight data problem. Each of the five partners focused on different aspects of the problems. When the funding was finally approved, the reduced budget could only support some 50% of the proposed project, and the overall objectives of the project were condensed. The Oxford partner (WP1) was tasked to focus on the flight data problem, and:

(a) To formulate an information-theoretic framework for visual analytics processes in a context exemplified by the FD problem;
(b) To develop visual representations that depict information contained in a dataset as well as the uncertainty quantities and other probabilistically-derived qualities such as importance and value.

Progress

The theatrical investigation into the FD problem started in 2012. A theoretic framework for anomaly detection was formulated in earlier part of 2013 and was first presented by Professor Min Chen to the consortium and stakeholders in the May 2013 meeting. The main concept introduced was initially called FaCAS (Focus and Context Attribute Subset).

Dr. Simon Walton was assigned to the project from the beginning, but could only work on the project on a full-time basis from May 2013 due to another contract. Dr. Walton led the research effort for prototyping, testing and demonstrating the proposed framework.

We first evaluated a number of available options, and identified that the software system developed by Purdue for the FD problem had most functionality necessary for testing and demonstrating the theoretic framework. The Purdue software was in a relatively mature status in February 2013 when they were preparing a submission to IEEE VAST 2013. As Purdue has been a stakeholder of this project, as the US funding was channelled through the VACCINE program led by Professor David Ebert as its director. The discussion about Oxford-Purdue collaboration started in February. In order to avoid complicating Purdue's submission plan, we decided to start the collaborative effort in May after Purdue had submitted their work to VAST 2013 and Simon had started his full-time role.

Dr. Simon Walton implemented a test suite in the Python programming language during the summer demonstrating the correctness of the information-theoretic framework against a known data source. This was later implemented in C++ for efficiency, and in the September meeting his report showed that the theoretic framework was correct, reasonably fast, and usable in anomaly detection. We also received a full set of source code from Purdue, and Simon started to work closely with Professor Ebert's PhD student Sungahn Ko and his former research officer Dr. Yun Jang (an assistant professor at Sejong University, South Korea). After the September meeting, the concept of FaCAS was renamed as QCAT (Query with Conditional ATtributes).

Following the advice of the stakeholders, Dr. Walton conducted a series of scalability tests on the methods in September and October, including a proof-of-concept MapReduce implementation. All those tests indicated that the technique is scalable in practice. In October 2013, the initial integration of the theoretic framework with the Purdue software also indicated a likely success. We started to work on a joint publication with Purdue by building on their previous submission to IEEE VAST. A completed draft was sent to the stakeholders for approval in late October.

During October and November, there were intensive collaborative activities between Oxford and Purdue. The work on visual representations focused on presenting consistent visual cues for depicting the results of anomaly detection across several different views. The testing of anomaly detection was focused

on specific periods, for which the US colleagues were aware of the presence of anomalous events. Several different QCATs were formulated and tested. This helped researchers involved to gain a good understanding of the concept as well as its uses in practice.

In December, a joint paper between Purdue and Oxford was submitted to EuroVis 2014, and Dr. Walton presented a report to the closing meeting of UKVAC project in December 2013.

Achievement

We have completed the WP1 of the project with a number of concrete achievements.

(a) We have formulated an information-theoretic framework for anomaly detection. A relatively comprehensive literature study showed that the idea of using information theory for anomaly detection has been examined by many since the 1980s. No existing scheme was found to be identical as our proposal. It suggests a number of possibilities, including (i) our scheme is novel; (ii) the same scheme was proposed earlier but was not disclosed; (iii) the same scheme was proposed earlier but was discarded due to the lack of big data or a practical implementation (e.g., using visual analytics).

(b) We have conducted a study on the visual representations. We have identified the requirements for an effective visual design, including consistent visual cues across multiple views, and supporting both visualization and exploration. The current visual design was formulated based on an existing system that introduces many constraints such as the use of colours. We believe that the visual designs can be improved further if there is a flexibility to redesign the visual representations in different views.

(c) With the help of Purdue, we have delivered a prototype demonstration system for illustrating the potential use of information theoretic quantities in the visual analytics process for handling the FD data. This is partly one of the original objectives that were removed due to the budget reduction.

(d) We have formulated a high-level visual analytics cycle by dividing visual analytics tasks into three groups, *monitoring*, *analysis* and *model refinement*. This proposed cycle allows us to create, deploy, use, evaluate and improve automated detection techniques (e.g., information-theoretic anomaly detection) in different operational modes, by users at different skill-levels and in different numbers, and through different computational processes with different demands for database access and pre-processing. This high-level thinking was partly supported by the empirical studies carried at UCL (WP4).

Future Work

The development of the QCAT technique is only a start. There are several new directions that we wish to follow. In addition to studying its mathematical properties, we wish to establish an effective and efficient workflow for creating, testing and monitoring QCATs. There is also a suggestion of the potential benefit of creating a hybrid scheme with machine learning, such as the Bayesian model used by Imperial in WP3, and developing a visual search facilities for identified anomalies (possibly with WP5). Should there be a need to invest in a new software framework for visual analytics, we will be interested in working with many partners in the field of visual analytics (e.g., WP2).

Appendices

- EuroVis14submission.pdf — The joint paper submission with Purdue
- Oxford-FD1-Jan2013.pdf — Presentation by Min Chen
- Oxford-FD2-May2013.pdf — Presentation by Min Chen
- Oxford-FD3-Sep2013.pdf — Presentation by Simon Walton
- Oxford-FD4a-Dec2013.pdf — Presentation by Min Chen
- Oxford-FD4b-Dec2013.pdf — Presentation by Simon Walton
- QCAT-TechnicalReport.pdf — Technical report on QCAT
- SummaryReport.pdf — Summary WP1 report
- The source codes related to QCAT are available to the stakeholders. To access the source codes, please contact Dr. Simon Walton.

UKVAC II Program – WP1

QCATs (Query with Conditional Attributes): An information-theoretic approach to visual analytics

Simon Walton and Min Chen, Oxford University

ABSTRACT

We present an information-theoretic anomaly detection method called QCAT (Query with Conditional Attributes), detail its implementation, performance characteristics, and an example application for aiding query design.

Table of Contents

Technical Overview of QCAT Implementation

Our implementation of the QCAT methodology can be found within a self-contained C++ API, available upon request.

Dependencies

1. **libpqxx**: The official C++ client for PostgreSQL. [1]
2. **Boost 1.53**: Peer-reviewed C++ utility libraries. [2]

Compiling

A makefile is included in the API directory to generate the object files that can be pulled into external software.

Documentation

The API is fully documented using Doxygen. The Doxygen config file can be found in the /API folder, and both HTML and LaTeX versions of the compiled documentation are also contained within /API.

Data Source

The API has been heavily tested with the flight delay data source [3] converted from its original CSV format into a PostgreSQL database using utility scripts written in Python. These utility scripts can be found in the /DataImporter directory of the project structure.

The API accesses the data source through the `QCATDataSource` class. For efficiency reasons, the API is engineered to deal directly with libpqxx types (cursors, etc). It is of course possible to use another type of data source, but such an undertaking would require some time and tradeoff. Therefore for the scope of this project, we have concentrated only on libpqxx.

General Goal of our Method

Information Theory

We consider each of the attributes in the incoming data source:

$$X_1, X_1, \cdots X_n$$

as a letter in our information theory alphabet. In terms of probability theory, an alphabet can be considered as a *variable* and each letter in the alphabet as a *value*. For example, if X_1 above is an unsigned 8-bit integer, we would have:

$$X_1 = \{x_1, x_2, \cdots, x_n\} = \{0,1, \cdots, 255\}$$

---

[1] http://pqxx.org/development/libpqxx/ (BSD License)

[2] http://www.boost.org/ (Boost Software License)

[3] http://stat-computing.org/dataexpo/2009/

Analysing each attribute's record in the dataset would reveal its probability mass function:

$$p(x_i)$$

which gives us the probability of each value x_i occurring in our data source. The probability by itself also gives an idea of the information content: the higher the probability of a value occurring, the lower its information content since it is more expected. In Shannon's information theory, this is expressed as:

$$-log_2 p(x_i)$$

using base 2 (binary representation). Thus, the above tells us the information content of the letter x_i based on the probability mass function, expressed in the number of bits of information. From this, we can define the entropy of the alphabet, i.e., the group of all n attribute values as:

$$H(X) = -\sum_{i}^{n} p(x_i) log_2 p(x_i)$$

Equation 1 Entropy calculation

Estimating a Probability Mass Function

It is immediately apparent that for real-world data, the probability mass function of those attributes with a large value range, e.g., many numerical attributes, would be difficult to estimate. For example, a standard `smallint` datatype in PostgreSQL has a possible range of $2^{16} = 65{,}536$. In order to estimate the probability of each value in this variable, e.g., assuming on average 100 occurrences per value, we would need a dataset with 6,553,600 data points. For an alphabet is defined by a combination of two attributes of `smallint`, there would be $2^{32} = 4{,}294{,}967{,}296$ different letters. To estimate its probability mass function, one would need 429,496,729,600 data points. The flight data set has 29 attributes, but only fewer than 120,000,000 data points. Hence the big data is not really big enough, which offer poor indicators of information content as the number of possible values would be far larger than the number of actual values found in the data.

Hence with a multi-dimensional problem with an insufficient number of data points, we have to estimate the probability mass function of an alphabet with a smaller number of letters. This is achieved by grouping closely-related values into bins. If we have an attribute in the flight dataset representing distance, then realistically we may consider only 2000 values (for US-only flight data), which reduces the space somewhat; however, the probability mass function for this attribute is still too `high-frequency' for the purposes of analyzing distance as we more often than not wish to treat distances in `blocks' of a number of miles each (for example, in blocks of 10 mile, two values of 57 and 52 are treated as the same distance).

In our implementation, reduction of space is achieved using binning. The probability mass function therefore becomes a histogram of values of fixed bin width. The size of each bin may vary depending on the attribute. A graphical user interface, developed in Qt, is available in project directory under /QCATDesigner, which allows for the interactive design of a QCAT and specification of its bins.

Server-side / Client-side Implementations

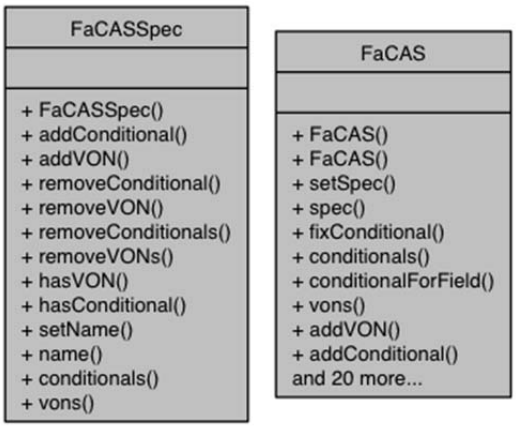
The 'grunt work' of the API, the actual QCAT computation itself, can be performed in one of two places at runtime:

- The **server** side – the QCAT conditionals are evaluated and the resulting summary is aggregated from this entirely within the PostgreSQL database server using stored procedures;
- The **client** side – the database server is used only to evaluate the conditional component of the QCAT, and the client-side C++ code performs the computation on the resulting rows.

These implementations are discussed in more detail in the coming sections, but first we detail the general strategy for evaluating a QCAT from its specification and computing a result.

Specifying a QCAT

In both cases, the relevant code can be found in the `QCATSpec` and `QCAT` classes in the API.



Two important classes: QCATSpec and QCAT.

The `QCATSpec` class defines a `QCAT` specification; that is, a set of conditionals and a set of VONs. These are specified simply as the names of fields in the target data source, which the specification object is unaware of. The `QCAT` takes a `QCATSpec` object and a `QCATDataSource` object and instantiates a `QCAT` with conditionals bound to specific values, ready to be executed. This class is then used to execute the `QCAT` specification on the data source.

The class can be instructed to perform either server or client-side execution using the `setExecutionMethod` member.

General Method

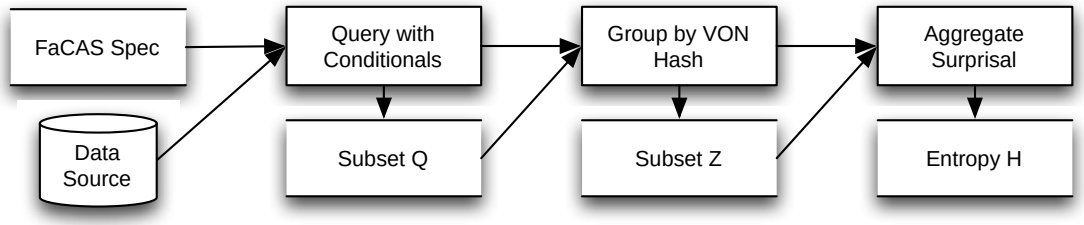


Figure 1 Implementing a QCAT

Figure 1 Implementing a QCAT gives the flow for an algorithm to compute a QCAT, including the final aggregation step to sum the surprise values and compute a final entropy value. The inputs to the algorithm are the QCAT specification (a setoff conditionals bound to values, and a set of VONs), and the data source upon which the QCAT is run.

The flow generates three datasets:

1. **Q**: The set of all records from the incoming data source that match the bound conditionals in the QCAT;
2. **Z**: The set of all records in **Q** are grouped by the set of VONs specified in the QCAT specification, with single count aggregation operation to provide the number of records matching each VON group;
3. **H**: The set of all records from **Z** are finally summed according to the entropy calculation in Equation 1 Entropy calculation. The probability mass function for each row can be inferred easily by the group counts.

Server-side Implementation

The server-side implementation is a PostgreSQL stored procedure called via the API. The `CREATE PROCEDURE` statement to generate the procedure can be found in the /SQL directory in the project structure.

The procedure takes three parameters:

1. The name of the table to use;
2. An SQL string representing the condition statement of the QCAT, e.g. `origin = 'LAX' and dest = 'JFK';`
3. An SQL string representing the VON hash

The VON hash is an identifying string that represents a unique combination of VON values. The string that is provided to the function therefore must take each VON attribute from the table and produce a hash from these values such that the procedure is able to perform a `GROUP BY` using this hash.

The procedure executes two selections on the data: the first selects records matching the conditional and grouped by the VON hash, and the second outer selection performs the necessary aggregation to output a summary of the run, including:

- `totalRowCount`: the number of rows in the selection;
- `ZCount`: the size of the alphabet ;
- `HZ`: the entropy of this QCAT execution;
- `Sum_surprise`: The sum of all surprisal values.

Client-side Implementation

The client-side implementation uses the results of selecting from the data source using the QCAT's conditionals to construct the necessary VON grouping manually. The grouping itself is performed using an `unordered_map` (from STL/C++11) with the key as the VON hash and the value as the count of the number of records matching this hash. The final map represents the dictionary of the QCAT execution, and from this, we can easily compute the entropy by iterating the dictionary and performing the sum found in the entropy equation detailed earlier in the document.

Performance Characteristics

Both the server and client-side implementations have different performance characteristics. Generally, we find that the server implementation executes more quickly than the client-side version, even when the server is running on localhost. This is due to years of performance optimizations within PostgreSQL. In this section, we discuss some of the considerations made when optimizing the running of a QCAT using PostgreSQL, including the optimizations made during development.

Custom Log_2 function

In terms of optimizing the server-side stored procedure implementation, the biggest factor was in the use of a custom log_2 implementation, written in C and imported into PostgreSQL as a user-defined function call. The original implementation using Postgres' general `log` function was extremely slow in comparison and was proving to be a problem when dealing with millions of rows.

The custom log_2 function and its makefile can be found in the QCAT/PostgresC directory within the project structure. Note that the makefile is for OSX, but can be adapted to other systems by following the guide on PostgreSQL's website[4]. Once the object file is created, it can be added to PostgreSQL as a custom function using:

```
CREATE FUNCTION log2(float) RETURNS float
     AS '/the/location/log2.so', '_log2'
     LANGUAGE C STRICT;
```

[4] http://www.postgresql.org/docs/9.3/static/xfunc-c.html

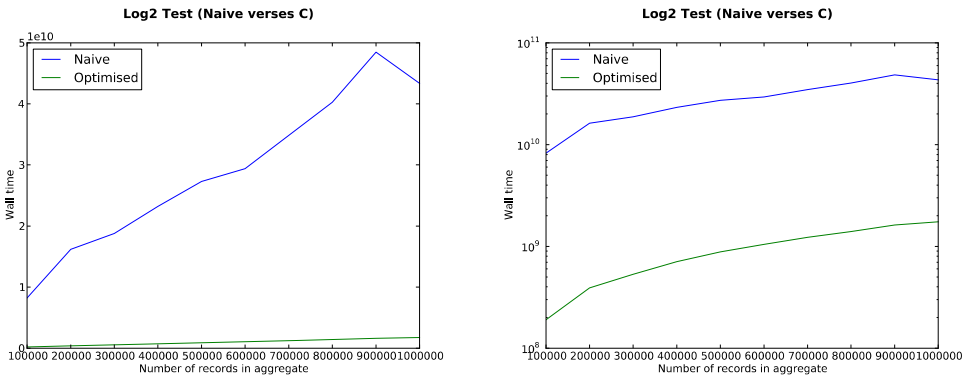


Figure 2 Wall times for naive Log2 verses Optimsed C version (left) linear Y (right) logarithmic Y

Figure 2 Wall times for naive Log2 verses Optimsed C version (left) linear Y (right) logarithmic Yshows the timings achieved for the naïve implementation (using PostgreSQL's built-in log function) and the custom C implementation, on a varying number of rows. The difference is large enough that a logarithmic chart is required to show both with clarity.

Indexing

Enabling indexing on specific columns within the database provides benefits to both the server and client-side operations as both methods use the databases' provided functionality for selecting rows that match the QCAT specification's conditionals. Indexing a given column instructs PostgreSQL to create internal data structures to evaluate conditionals more quickly, but at the expense of the additional space required to store such structures, and slower insert operations.

PostgresSQL is able to create indexes for columns, column groups, and also arbitrary expressions. The QCAT API automatically enables column indexing on columns used in a QCAT conditional to speed up future operations with that QCAT's conditionals. Creating indexes representing groups of attributes (i.e. origin and destination) provides an additional speedup when these attributes are used together in a conditional.

Enabling indexing merely on the column values however would provide only modest real benefit to the QCAT subsystem as most attributes feature a bin expression that is to be evaluated (i.e. value /10 for a bin width of 10). Therefore, the QCAT API additionally automatically creates necessary indexes for the bin expressions.

Note that as mentioned, these indexes are created at the expense of additional disk space and incur a penalty for insertion. We have found in testing that for 1,000,000 rows, a B+-tree index on a `smallint` column consumes an additional ~55MB on disk. As such, the use of indexes and bin expression indexes would have to be weighed against these factors as such a system grows in size.

Work_Mem Configuration

Inside Postgres' server configuration file is a value `work_mem`, which Postgres uses internally for internal sorting and hashing operations. Once the specified amount of memory has been exceeded, Postgres switches to memory-mapped files (effectively paging to disk), slowing the remainder of the operation. The

latter is of some interest to us, as hashing operations are used extensively when performing a `GROUP BY` operation upon the data, and the QCAT workflow relies extensively upon grouping. During our testing, we found the value of `work_mem` to be a very large factor in QCAT runtime when scaling to larger numbers of attributes. A detailed discussion can be found in the next section.

Shared_Buffers Configuration

Ideally, PostgreSQL's `shared_buffers` should be around ¼ of the available RAM. The default settings are very conservative to get around some OS installations having a very low maximum shared memory limit (such as Ubuntu).

Note that both `work_mem` and `shared_buffers` require a restart of the database server in order to take effect.

Scalability Testing

Our initial tests with the API were promising in that the execution time of an QCAT with one million rows in its final result subset was less than two seconds on a Core i5 Macbook Pro (abbreviated to MBP) with 8GB RAM. We employed the use of a dedicated server in order to perform more realistic scalability testing, for the following reasons:

1. A standard Macbook Pro does not represent the specification of the average database server deployed in a real environment;
2. Running both the API and the database on localhost is disingenuous, as it does not reflect the bandwidth and lag of a real-world database deployment.

For the above reasons, we employed the use of a virtual machine running PostgreSQL on our internal network to continue scalability testing in a more controlled and realistic manner. The machine used was a 2.8GHz Intel Xeon X5660 with 16GB RAM and 12MB cache.

Test 1: Number of Records

The first test was to evaluate the efficiency of aggregating large numbers of rows in the final aggregation operation of the QCAT algorithm. To achieve this, we used a QCAT with a simple condition that always evaluates to TRUE so that all rows are selected, and disabled binning to gain over the total number of rows aggregated.

Figures Figure 3 Testing aggregation speed using remote DB and Figure 4 Testing aggregation speed using localhost DB show the results of our scalability testing with on firstly a remote database, and secondly, for comparison, a localhost database. In both cases, we show the results for the server-side QCAT evaluation method (blue), and the client-side QCAT evaluation method (green). Note that the client machine in Figure Figure 3 Testing aggregation speed using remote DB is the same machine as the server & client machine in Figure Figure 4 Testing aggregation speed using localhost DB. We include the localhost comparison in this to illustrate the fairly constant effect

It is clear that the relationship between the number of rows and runtime is of linear complexity; although turbulence develops in both cases for the client-side method once the method starts dealing with several hundred thousand rows.

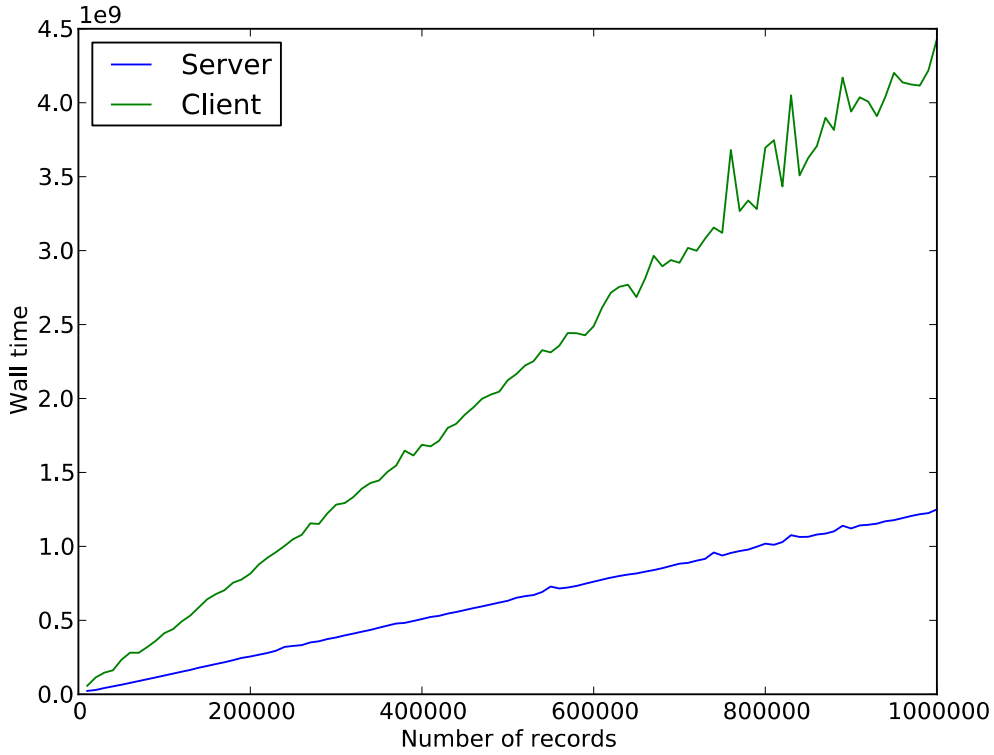


Figure 3 Testing aggregation speed using remote DB. Client: i5 MBP; Server: Xeon

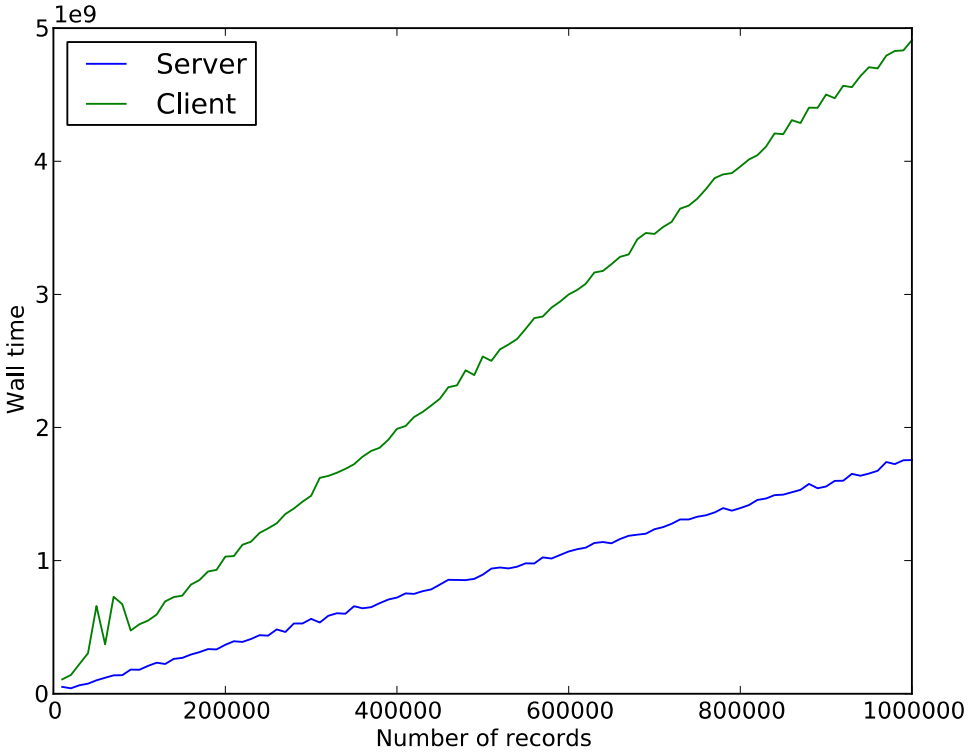


Figure 4 Testing aggregation speed using localhost DB. Client and server: i5 MBP.

We also scaled the test to 5,000,000 rows and found the same trend continued without interruption on our test machine, as shown in Figure 5.

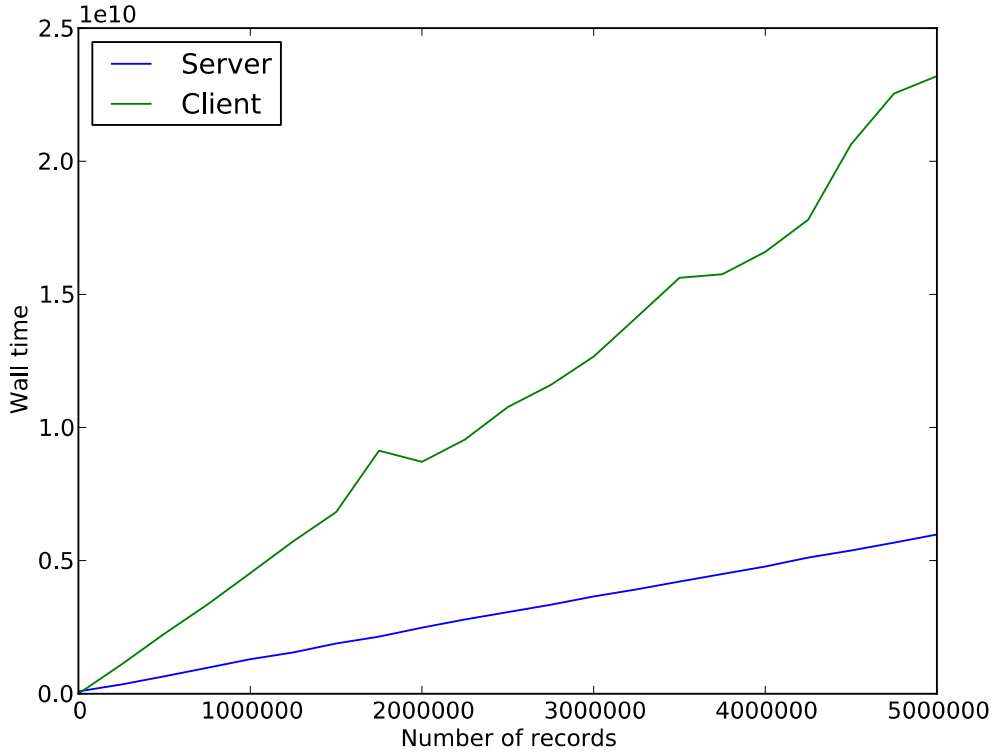


Figure 5 Scaling the number of records to 5,000,000

Using an online tool to visualize the estimated complexity of algorithms [5], we can show the complexity of the client and server evaluation strategies in terms of runtime verses varying input size. This is shown in Figure 6. In this figure, a series of complexity classes are provided along the *x*-axis, and the increasing number of rows (the *input size*) is shown on the *y*-axis. Further information on interpreting the graph can be found at the tool's website provided in the footnote.

[5] http://ovii.oerc.ox.ac.uk/cp/

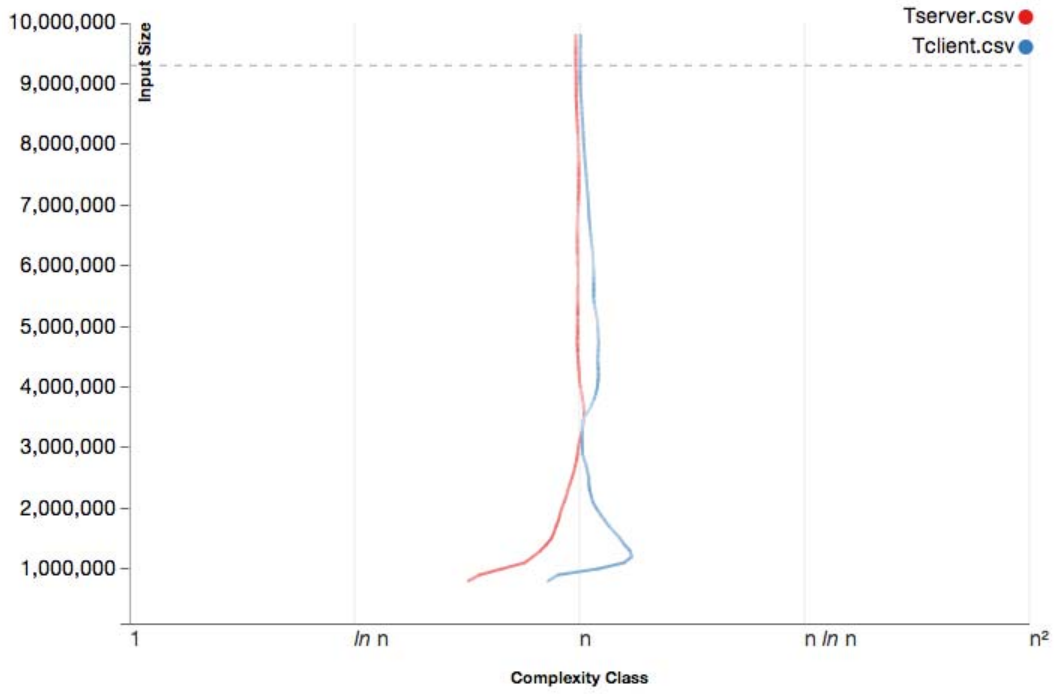


Figure 6 The complexity of varying number of aggregated rows is shown to be linear

In our tests, we did not find any internal PostgreSQL server parameters that affected this scalability in any significant manner. The speed of aggregating many rows is indeed mostly related to CPU speed, bandwidth available between storage / CPU / memory, and the speed of storage access. This is reflected in the comparison between the Xeon test machine and i5 Macbook Pro test machines, where both tests show linear complexity multiplied by the constant factor of available architecture speed and memory / storage bandwidth.

Test 2: Number of VONs

Our next test was designed to test Postgres' scalability in terms of the number of attributes per row whilst keeping the number of rows constant. In the context of a QCAT, this has implications when choosing larger numbers of VONs.

Note that the datatype of each VON also has an effect on the overall runtime as different datatypes have different expressions for creating their binned representation. Testing the different bin types is left to the next test.

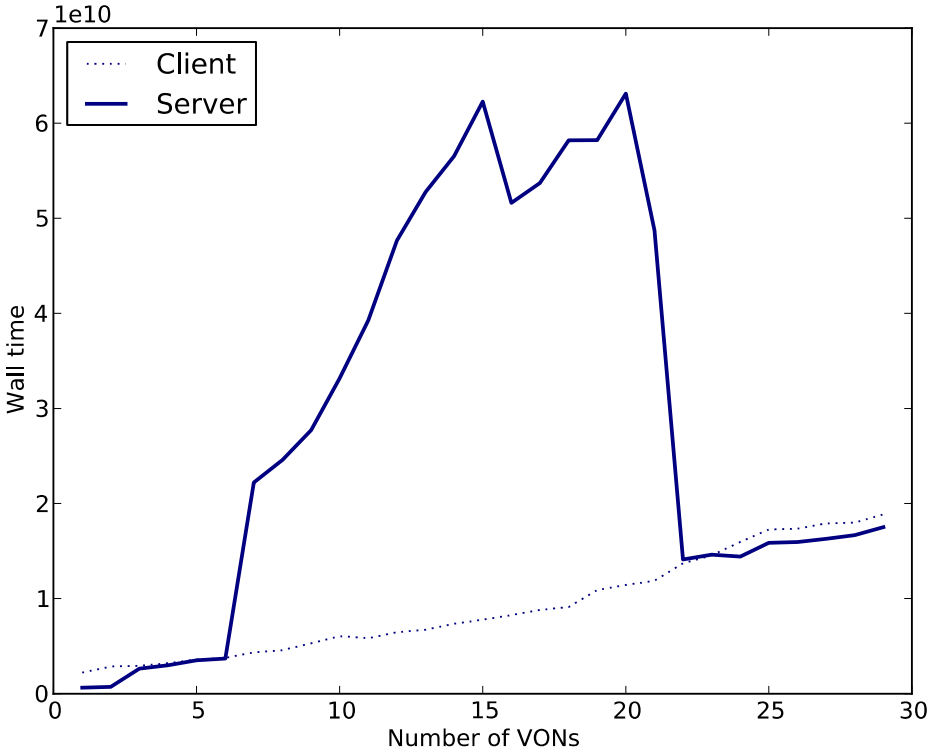


Figure 7 Testing scalability with the number of VONs on an i5 MBP

Figure 7 shows our first scalability test, where the *x*-axis represents the total number of attributes (VONs) included in the QCAT, and the *y*-axis represents the total time taken to run. The number of records is fixed at 1,000,000. The dotted line represents the client implementation, and the solid line represents the server implementation. The results for the latter are initially shocking: up until around 6 VONs, the server performance matches the client performance, after which there is a large amount of degradation of performance that continues until around 22 VONs, at which point the server performance jumps back to below the client performance as we'd expect.

What is happening is that Postgres's internal hashing operations required for grouping based on the VONs are being performed entirely in-memory up until a certain point; after which, Postgres is forced to use an out-of-core strategy. This threshold at which Postgres uses an out-of-core strategy is the `work_mem` threshold found in Postgres' internal configuration. By default, it is set to a very conservative amount of 2MB so as to avoid overloading the system memory when many connections are simultaneously executing operations that involve large hash tables (such as sorting and, in this case, grouping).

This saturation of available memory by the internal hash tables is illustrated in Figure 8, which shows the effect of increasing numbers of attributes (bins) on the operating system's virtual memory usage. The number of `pageins` jumps dramatically from around three attributes onwards, indicating a shift to out-of-core.

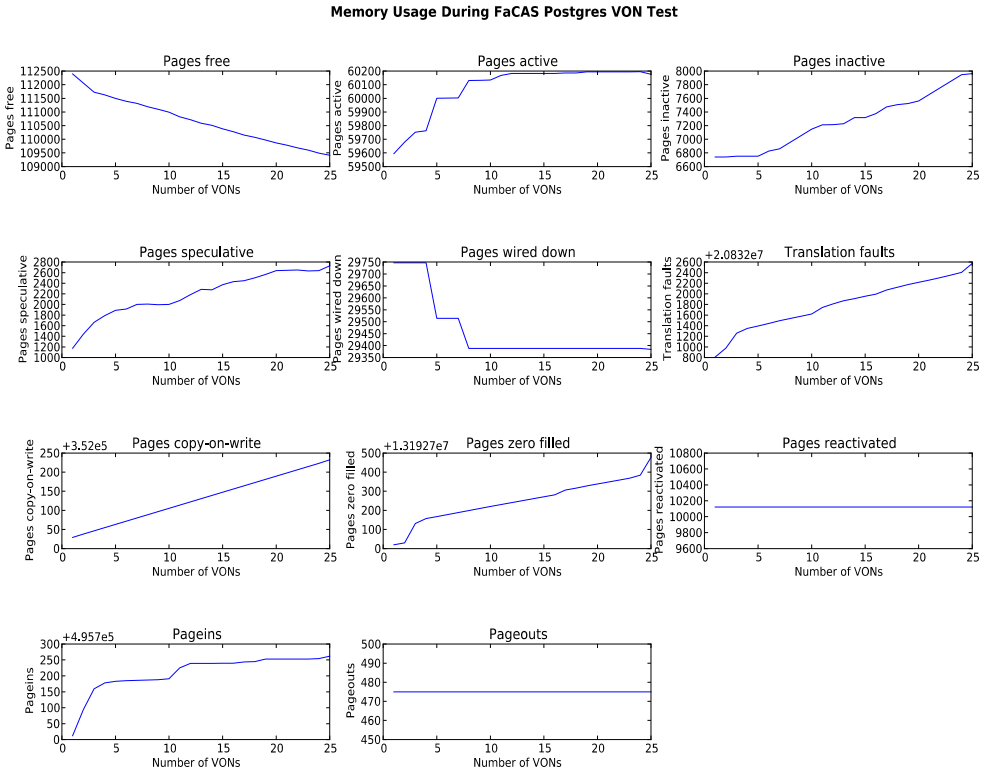


Figure 8 Virtual memory measurements during VON test

Figure 9 shows our VON test run two more times, but this time we double the amount of working memory Postgres uses for each run. It is apparent that performance improves as the working memory improves; however it is also apparent that the Postgres is deciding to switch to the out-of-core strategy for the same amount of VONs as previously. If we further increase the amount of working memory, we finally achieve the performance graph that we are looking for in Figure 10, where it is clear that a value somewhere between 64MB and 128MB has made Postgres decide to keep the hash tables in-memory.

If we further probe into the range between 32MB-64MB, we find that the initial point at which Postgres switches to the out-of-core strategy is effectively delayed by the amount of working memory available, until at somewhere between 48MB and 64MB the main stretch of degradation is mostly avoided altogether. Interestingly, it is also apparent that the apparently smooth progress after ~23 VONs can itself be further optimized by using larger amounts of working memory, as is clear from the difference between 64MB and 128MB after ~23 VONs.

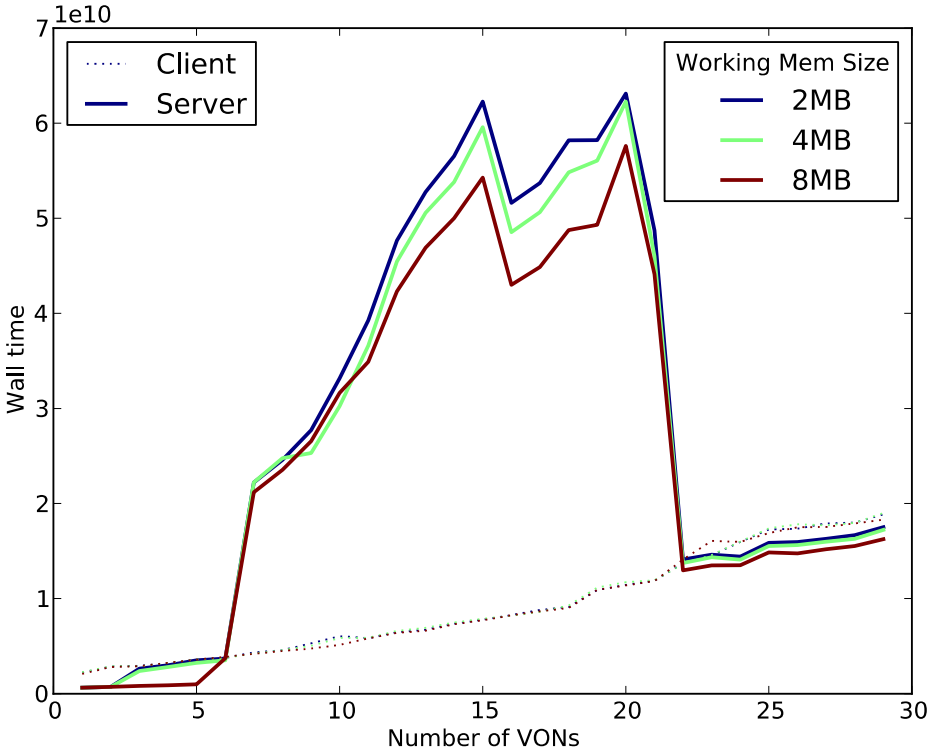


Figure 9 Increasing the size of work_mem improves VON scalability

Next, we wanted to investigate the effect of increasing the total number of records used in the VON test. The delay table was increased in size to 5,000,000 records, and the VON test re-run. Figure 12 shows the result with larger working memory sizes than with the 1,000,00 row test.

For comparison, we show in Figure 13 the results of running the server implementation with:

1. 1M rows vs 5M rows;
2. Two different values for `work_mem`.

As expected, the larger load on the internal hash tables caused by the increased number of rows forces the out-of-core implementation to take effect earlier.

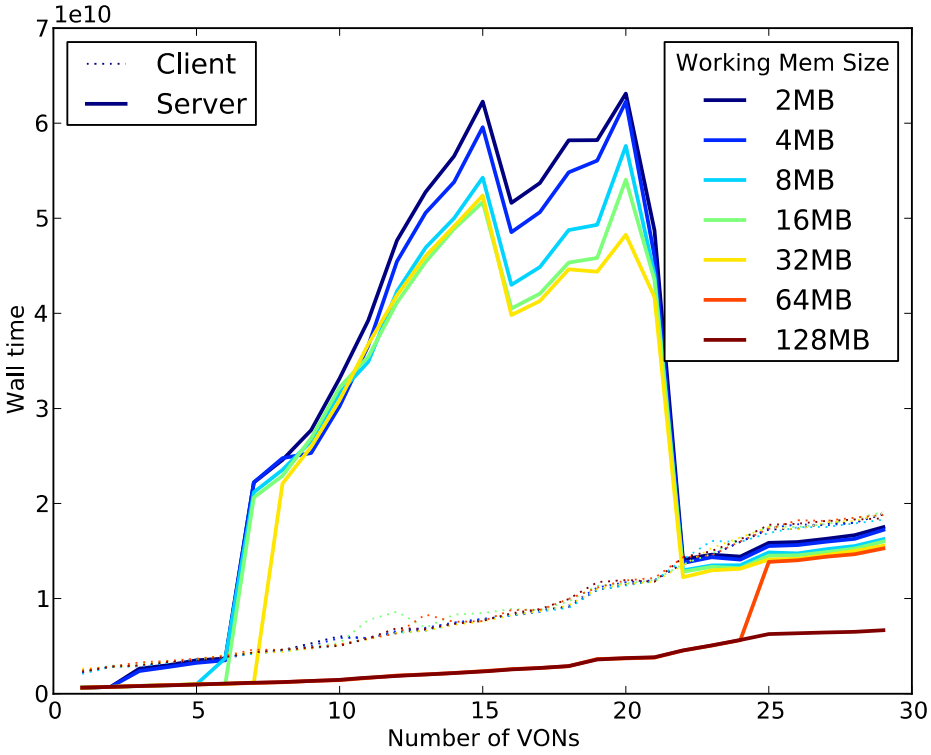


Figure 10 Increasing work_mem until server-side performance behaves as expected

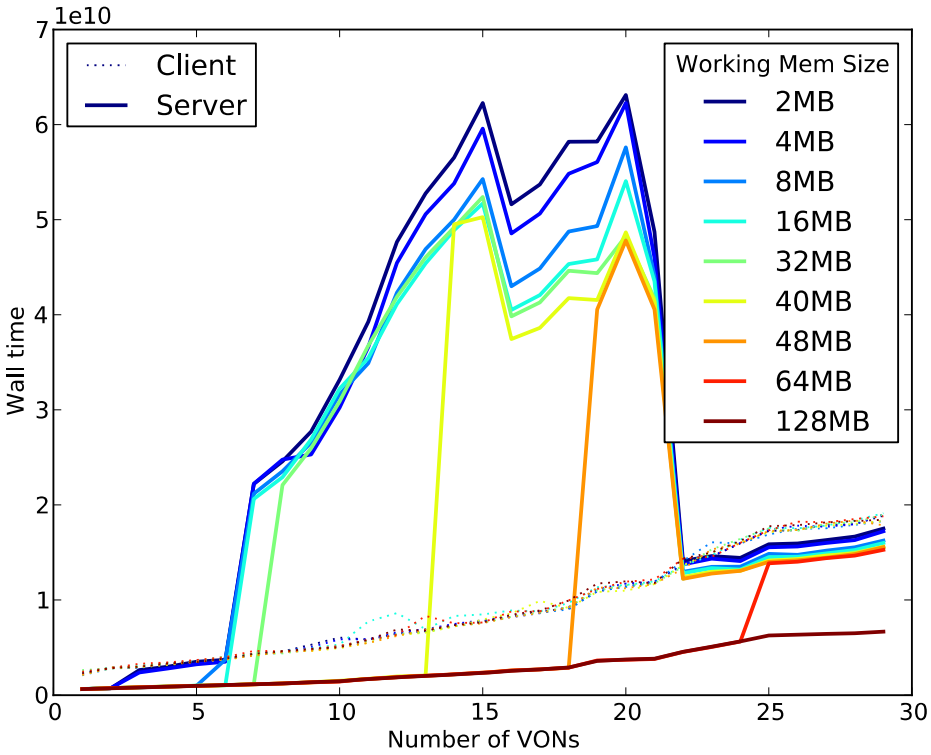


Figure 11 Further investigating the effect of work_mem on VON scalability reveals the shape of Postgres' out-of-core strategy

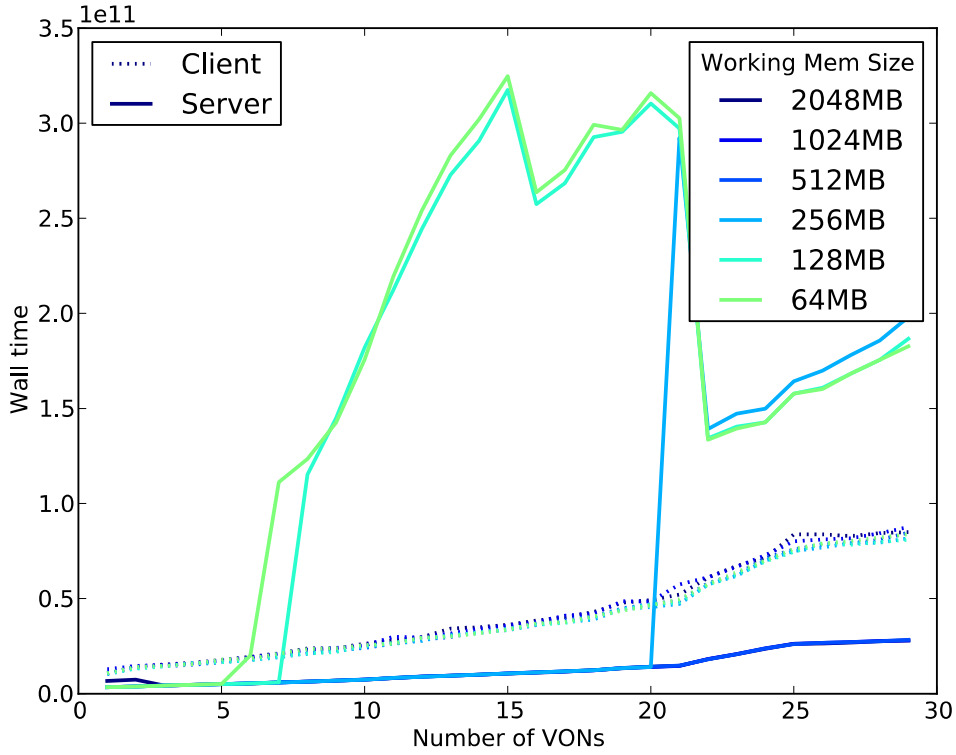


Figure 12 Further VON stress testing with 5,000,000 rows

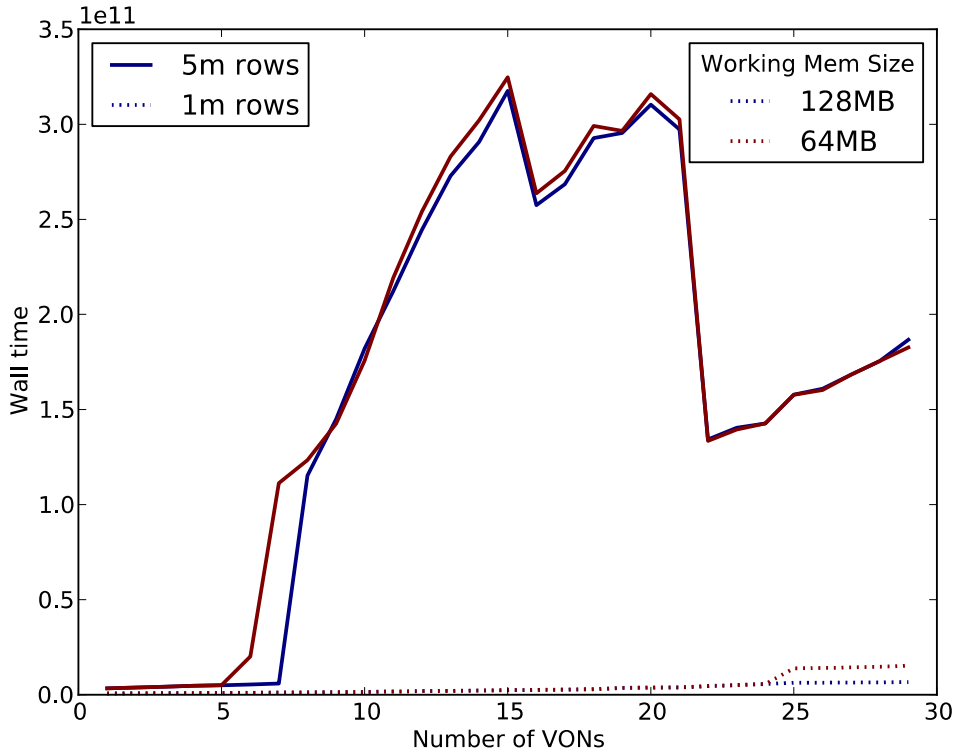


Figure 13 Comparing 1m and 5m rows with varying working mem size

Test 3: VON types

A QCAT is usually executed with mostly binned VONs; that is, the VON space has been reduced by binning the value distribution into a histogram of smaller *x*-range than the actual data. To achieve this, the VON expressions provided to the database provider must be an expression to convert the actual value obtained from the data into the binned value ready to be input to the grouping part of the algorithm.

Three classes in the API are responsible for providing expressions, one for each data type available:

1. `QCATFieldBinPassthrough` – no bin
2. `QCATFieldBinNumeric` – numeric bin
3. `QCATFieldBinTimestamp` – timespan-based bin

Each class is derived from `QCATFieldBin`. Internally, the QCAT class instantiates the correct class for each field depending on its data type, and the SQL expression to provide to PostgreSQL is generated by the appropriate classes via polymorphism at runtime. In the case of the passthrough class, used for datatypes that cannot be binned (such as Boolean and string), the field name is simply output.

In our implementation, the database server performs the binning for both client and server implementations. Each type of bin expression incurs its own overhead. The numeric bin type is the most simple with a single division by the bin width and a floor operation to convert to integer. The timestamp bin type however is more involved as the number of minutes (minutes are the chosen granularity for our implementation) since the Unix epoch are calculated from the timestamp datatype and then binned as a numeric type as above.

Figure 14 shows the evaluation speed of 1,000,000 rows with a varying number of attributes. Each chart represents a different data type. Numeric and timestamp are binned; string and Boolean are not. It is clear that the highest performance penalty comes from the timestamp bin conversion, with the lowest penalty for Boolean access. Numeric is somewhere in between but somewhat less smooth.

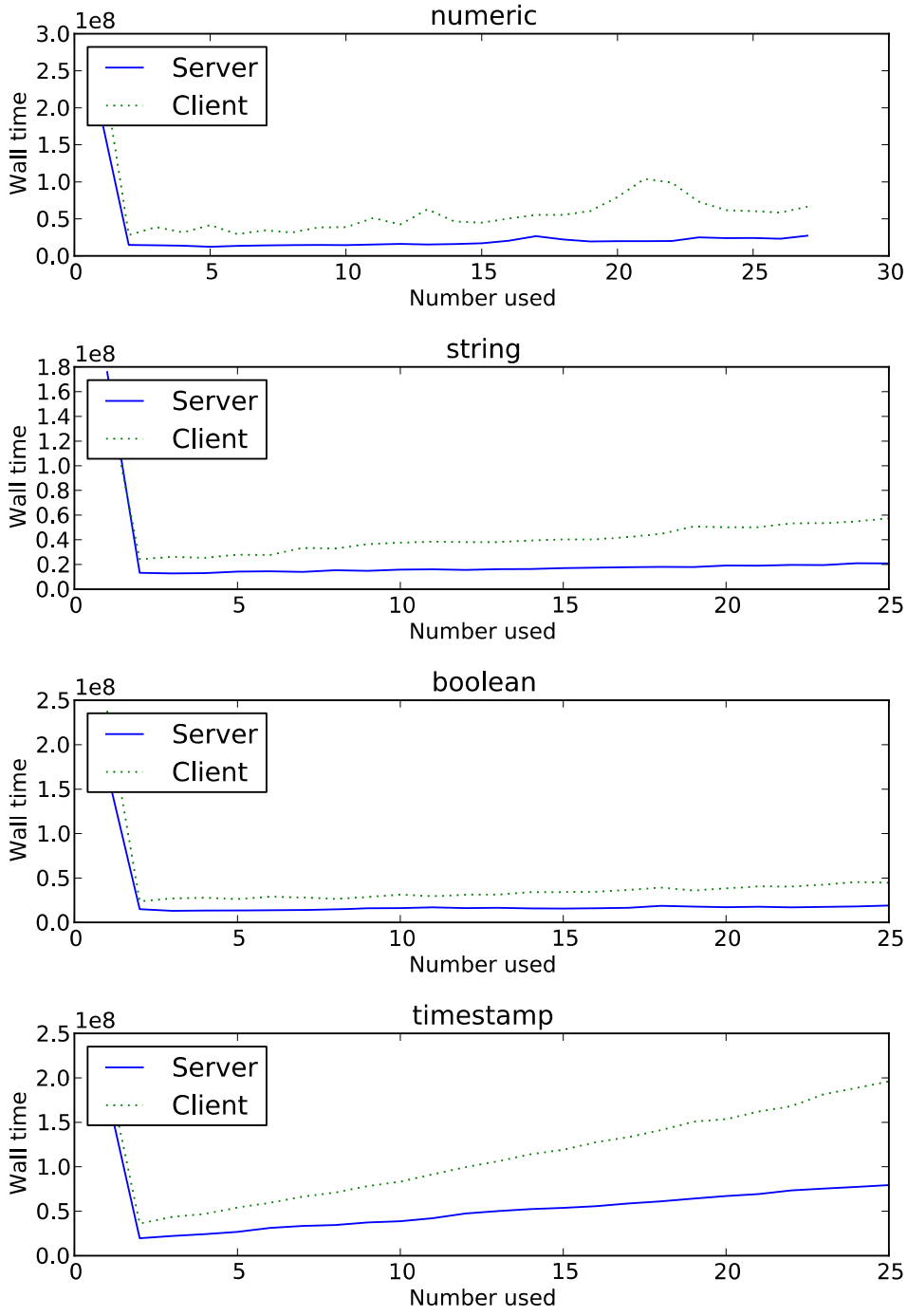


Figure 14 Testing the evaluation speed of four different data types

Utilizing a Hybrid Approach

It is clear from our experiments that the server implementation of the QCAT method is the most efficient if a large enough amount of working memory is provided to Postgres in order to perform its hashing in-memory. Past this point, a severe performance penalty occurs and the client implementation becomes the quickest method.

We would anticipate therefore that a hybrid approach of the two methods would provide the best balance in a real-world setting. For a given session execution a group of QCATs, we can measure the performance obtained from the database server to develop a coarse-scale matrix with the number of records on one axis and the number of VONs on the other. This matrix is filled in for large multiples of rows, and smaller multiples of VONs. Somewhere between the top-left and bottom-right diagonals, we should find the out-of-core divide. Using this information, the `QCAT` class can automatically switch to the client-side implementation when it anticipates that the server implementation would resort to out-of-core for an incoming execution request.

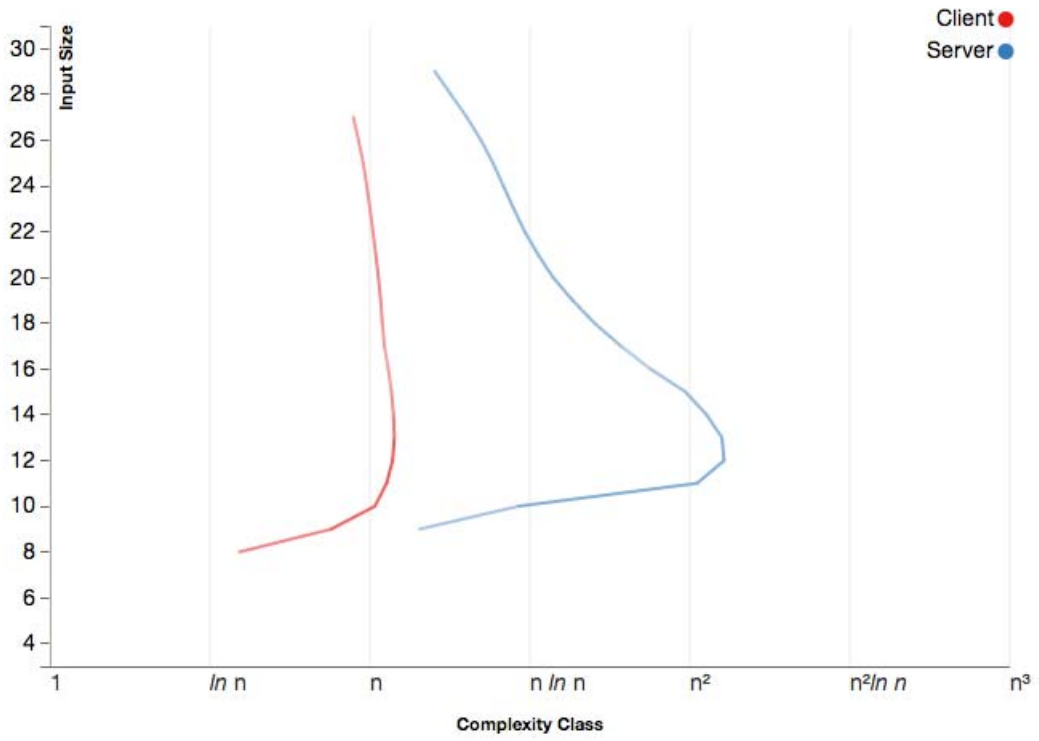


Figure 15 Increasing the number of VONs is of at most linear complexity.

Map-Reduce Implementation

As a proof-of-concept, the QCAT computation has been implemented as a map-reduce operation, and can be found within project folder /MapReduce. The implementation is provided purely as a proof-of-concept, and is inefficient compared to its native procedural implementation for the following reasons:

1. The data source is the raw CSV data, which incurs both a parsing and a bandwidth overhead;
2. Each node must complete its initial reduce stage to produce the aggregation used in the computation of the probability mass, which is performed in the second map stage.

Generally, map-reduce implementations are far less efficient than their procedural counterparts until the data becomes so large that the data cannot be feasibly stored on a single machine. Due to the scope of this project, we have not been able to perform such a scalability test, but the map-reduce implementation is theoretically correct and will produce the correct results when spanned over multiple machines to petabytes of data.

Design of MapReduce Implementation

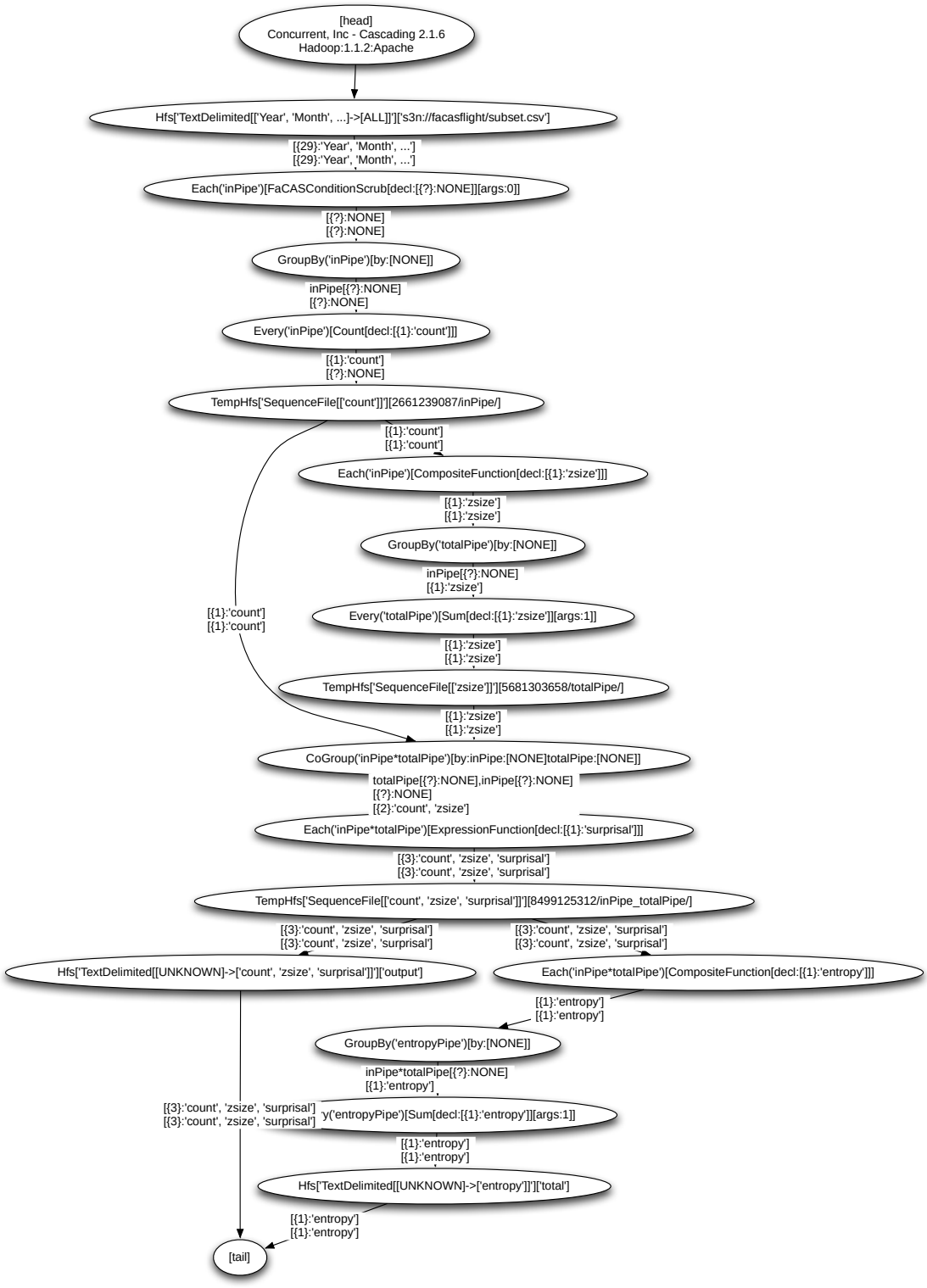


Figure 16 MapReduce implementation of QCAT Computation

Figure 16 MapReduce implementation of QCAT Computationprovides a flow diagram of the Cascading code, generated automatically from the Cascading API from the connections of taps and pipes specified.

There are three other classes, aside from `Main`, which runs the flow:

- `QCATCondition`: represents a single QCAT condition with a field name and associated value that must be satisfied
- `QCATSpec`: a simple representation of a QCAT specification, including a set of conditions and a set of VONs
- `QCATConditionScrub`: a helper class derived from Cascading's `BaseOperation` class, which performs filtering based upon conditionals.

The incoming tap is specified as the flight data source in CSV format. The first map stage in the pipeline performs an `Every` operation upon the data to output records matching the specified QCAT conditions. In our implementation, this is achieved using the `QCATConditionScrub` class. This class overrides `operate` from `BaseOperation` to extract the relevant fields from the row matching fields in the QCAT condition, and pass out the VON group if all conditionals match.

The first reduce stage in the pipeline performs a `GroupBy` based on the VON fields. At this point, we need to split the pipe into two in order to:

1. Count the number of records to calculate the probability mass function;
2. Pass the result from (1) to back into the original pipe to calculate the surprisal values.

The two pipes are combined using a `CoGroup` to attach the count into the original data stream, and a custom `ExpressionFunction` inserts a surprisal field into the stream. Finally, a new pipe is created as the result of a `SumBy` operation to sum the surprisal values into a final entropy value for the QCAT query.

Deployment

The MapReduce implementation uses the Cascading [6] library and is designed to be deployed as a single .jar file (with all dependencies bundled inside) to an Apache Hadoop [7] installation. The code can be built using Maven, and a Maven makefile is included in the directory. The Maven makefile has been constructed to automatically include all dependencies into the final .jar file such that the Hadoop installation's JVM is able to use dependencies straight from the provided bundle rather than searching its own classpaths.

To run the .jar, (assuming Hadoop is installed and in your $PATH), execute:

```
hadoop jar jar_name.jar
```

QCATDesigner Graphical User Interface

A Graphical User Interface, developed using Qt 5, is included in the project structure under the /QCATDesigner folder.

Dependencies (in addition to the QCAT API and its dependencies)

- QCustomPlot [8]

---

[6] http://www.cascading.org/ - Apache v2 Licensed

[7] http://hadoop.apache.org/ - Apache Software Licensed

[8] http://www.qcustomplot.com/ - GPL Licensed

- Qt 5 [9] (non-commercial licence)

Purpose and Usage

The QCATDesigner application is a proof-of-concept tool for interactively designing, defining and running a QCAT. Recall from our proposed workflow that there are three main roles: the designer, the analyst, and the observer. This Graphical User Interface was developed with the QCAT designer role in mind.

The interface assumes a certain level of technical knowledge from the designer, including familiarity with probability distributions and histograms. Such knowledge is not necessarily required the role of the observer, who would make use of visualizations of the QCAT outputs in order to make judgments on incoming data.

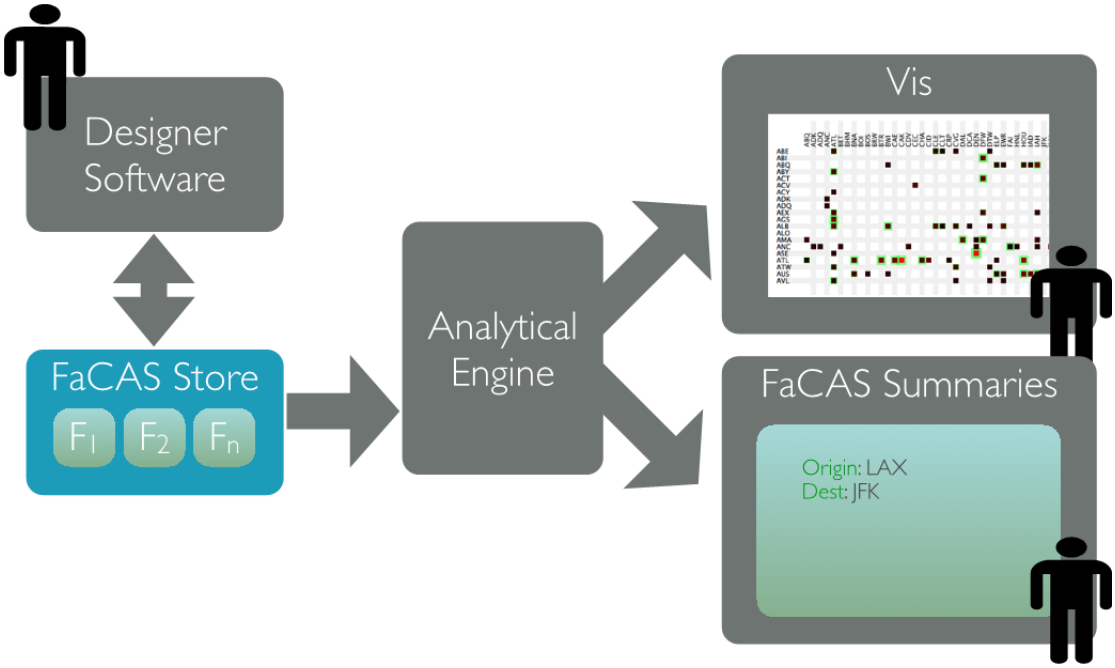


Figure 17 The three roles within our proposed workflow: designer, analyst, observer

[9] http://qt-project.org/ - GPL v3 Licensed

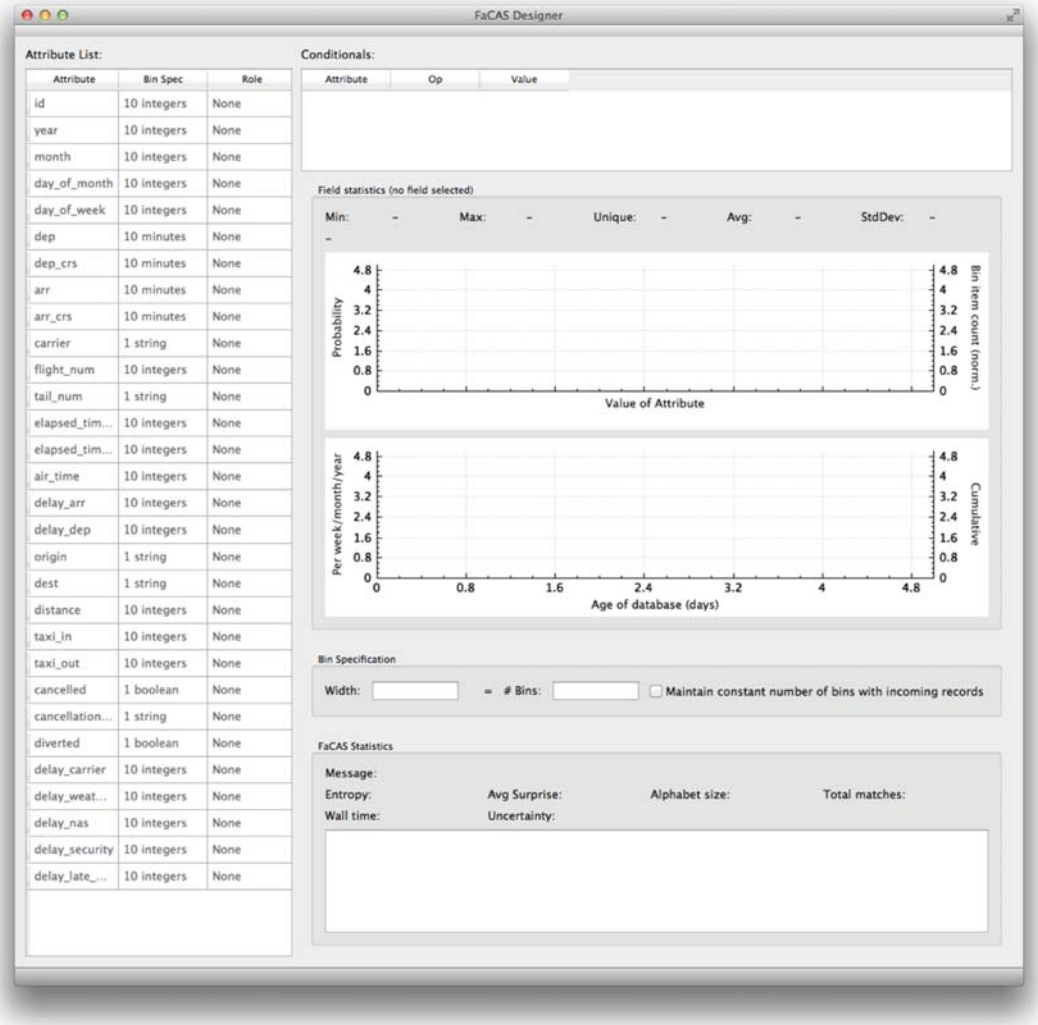


Figure 18 The QCATDesigner User Interface

User Interface Overview

Figure 18 The QCATDesigner User Interface shows the user interface. The list view to the left of the interface displays all available fields from the data source. In this case, we are showing the fields from the flight delay table. is represented by three columns:

1. The field name;
2. The current bin specification (i.e. bin width) for the field;
3. The field's QCAT role (None, Conditional, or VON).

Each field in the list represents a potential field to be included in the QCAT specification. By default, no fields are included, and hence the role for each is defaulted to *None*.

The top right of the interface provides a space to fix values to specified QCAT conditionals in order to test the QCAT output, which appears at the bottom right of the interface when the QCAT has completed running (in the QCAT Statistics group box in the bottom right of the interface).

The middle area on the right hand side provides information about the currently-selected field, and this area is used not only to judge whether a field is a suitable candidate for a particular role in the QCAT, but also on its bin specification if the field is included.

Running a QCAT

The user is able to specify and run a QCAT by selecting the role column of a desired field, and changing its role to either conditional, or VON. In these cases, the fields become highlighted in green (conditional) or blue (VON).

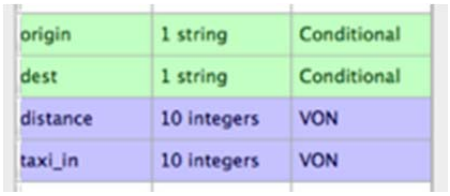


Figure 19 Conditionals and VONs in the UI

Once conditionals are specified, they appear in the Conditionals view in the top right of the interface (see Figure 20 Binding conditionals to values ready to run).

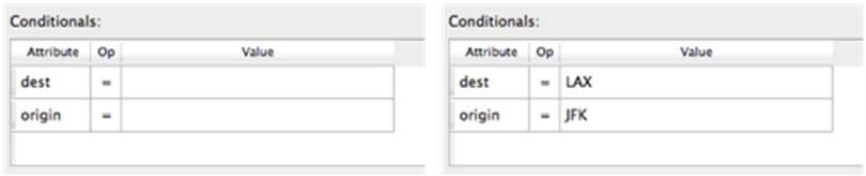


Figure 20 Binding conditionals to values ready to run

Once a QCAT has conditions bound to values, and has one or more VONs, it is automatically executed on the data source and its summary appears in the QCAT Statistics group box. Included in the summary are the following outputs:

- A message indicating any problems that occurred whilst attempting to run the QCAT (or a message of success);
- The entropy of the QCAT;
- The average level of surprisal found in the alphabet;
- The total size of the alphabet;
- The total number of records matching the conditionals before grouping;
- The wall time taken to run the QCAT;
- The uncertainty of the QCAT.

Note that if there are conditionals specified that are not bound to values, and/or there are zero VONs, then the QCAT cannot be executed.

List of Figures