

Threat Stream Data Generator: Creating the Known Unknowns for Test and Evaluation of Visual Analytics Tools

Mark A. Whiting, Wendy Cowley, Jereme Haack, Doug Love, Stephen Tratz, Caroline Varley, Kim Wiessner
Pacific Northwest National Laboratory
PO Box 999

Richland, WA 99352 USA

{Mark.A.Whiting, Wendy, Jereme.Haack, Doug.Love, Stephen.Tratz, Carrie.Varley, Kimberly.Wiessner}@pnl.gov

ABSTRACT

We present the Threat Stream Data Generator, an approach and tool for creating synthetic data sets for the test and evaluation of visual analytics tools and environments. We have focused on working with information analysts to understand the characteristics of threat data, to develop scenarios that will allow us to define data sets with known ground truth, to define a process of mapping threat elements in a scenario to expressions in data, and creating a software system to generate the data. We are also developing approaches to evaluating our data sets considering characteristics such as threat subtlety and appropriateness of data for the software to be examined.

Categories and Subject Descriptors

D.2.5 [SOFTWARE ENGINEERING]: Testing and Debugging
– *Testing tools (e.g., data generators, coverage testing)*

General Terms

Measurement, Performance, Design, Reliability, Human Factors, Standardization, Verification.

Keywords

Threat, Threat Stream, Data Generator

1. INTRODUCTION

The Threat Stream Data Generator (TSG) project at Pacific Northwest National Laboratory addresses evaluation and metrics for visual analytics technologies, under work being performed for the National Visualization and Analytics Center™ (NVAC™) [1]. A threat stream data set is a synthetic collection, consisting of homogenous or heterogeneous data, that contains one or more known, specified embedded threats and that can be used in the testing of visual analytics software. Threat stream data generation starts with consideration of a scenario and the identification and mapping of elements of the threats into data. Elements such as the users' knowledge of their domain and the hypotheses and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BELIV 2006 Venice, Italy.

Copyright 2006 ACM 1-59593-562-2/06/05 ...\$5.00.

evidence they identify as part of the threat identification process are also considered. Research in this area is critical—test data is vitally important for software evaluation. Good test data is nearly impossible to come by, and those who embark upon developing a test data set with a known embedded threat pursue a very expensive and resource-intensive endeavor. A practical TSG has the goal of substantially reducing the time for creation of test data. An additional benefit is that increased availability of realistic data

sets will help researchers in building good analytic tools. Presently, there is little for them to draw upon as inputs to their prototypes and experiments. Analysts' time is precious, and their availability is also limited by classification issues. Well thought out synthetic data may help researchers better understand the



Figure 1. Threat Stream Generator Process

analyst to a certain degree.

2. PROCESS STEPS

Creation of a reasonable synthetic data set that will satisfy the requirements of a rigorous evaluation procedure is a complex process. In general, the process follows the cycle depicted in Figure 1. The four process elements include creating a scenario, generating the synthetic data, testing the analytic tools against data with known threats, and assessing the tool performance. We have refined these steps in the process to reflect different parts of an evolving model:

Scenario Creation. A threat can have either a completely imaginary basis or a historical basis with a storyline that is fabricated. In either situation, the scenario must be sufficiently believable to allow a suspension of disbelief for an analyst who will be inspecting the data. We use existing commercial off-the-shelf (COTS) tools to support our work for this, including narrative description, Analysts Notebook-style entity relationship mapping, and timelines.

Threat Definition. This consists of the mapping from scenario elements to expressions in data. The mapping takes into account scenario time, events, cues, expression data set, data set format, elements revealed, evaluative parameters, and linkages.

Threat Stream Generation software. Our java-based system allows specification of different data forms through a “user-friendly” interface or through a threat specification language. Our software provides three primary approaches to data generation: statistical, rule-based, and semantic. Statistical data generation permits creation of data of various forms (for example, Boolean, numeric, text, fetch, and insert from previously stored files) according to various statistical distributions. The rule-based approach allows lightweight specification of data dependencies for generation. For example, if a person was born within a particular year timeframe, then his or her voter registration date should be no less than 18 years after that time and no greater than the current year. Semantic specification allows more sophisticated definition of complex relationships and the rules that the generated data should follow.

Threat Data Characterization. There are several ways to characterize the resulting test data. Overall threat subtlety is one measure of interest. Different factors involved in this include amount of evidence expressed, preponderance of cues expressed, data format diversity (homogeneous or heterogeneous format), and degree of deception. We currently use a Bayesian belief network to help determine a subtlety rating for data sets. Other characteristic measures include a threat risk determination. We have been interacting with external groups specializing in risk assessment to help us understand what kind of risk our synthetic data sets express, so that we might better understand how analysts will react to them. One approach to risk assessment is an algorithmic process to provide a numeric evaluation of comprehensive risk from sources, actions, location, targets, weapons, and opportunities. A related characteristic is consequence. We have developed an initial consequence assessment database to help us understand the consequence associated with our embedded (synthetic) threat, which may also help us understand the analyst’s reaction to a perceived threat.

Data Set - Analytic Tool Mapping and Expectations. As a final step in the data set preparation, we try to map the test data sets to the types of analytic tools to be assessed and also try to understand the extent of analysis that may be performed with these applications. For example, a homogeneous text-based data set would work well with a visual analytic tool such as IN-SPIRE.

However, only part of a heterogeneous data set, such as the one we are providing to the IEEE Visual Analytics Science and Technology symposium contest [2], will be useful to IN-SPIRE, because this tool cannot process numeric or image data. Additional tools and experts must be employed to analyze that data. We feel this is an important element of understanding the depth and breadth of analytic tool evaluation and to understand how to assess a collection of analytic tools that may be used by a collaborative team of analysts.

3. DATA SET GENERATION

The data set we have named “Spinosaurus” is a good example of a complex data set that contains challenges typical of a creating a large, multivariate data set. Spinosaurus was created to support testing a large information analytic system, and multiple versions of the set were generated to match changing formats and knowledge of operational data. NTCD-200 is one table of Spinosaurus that contains data about hypothetical border crossing events. The following is taken from the scenario description that we created, given to analysts wanting to test their software:

For this mini-scenario, a border bouncer would be a person who unsuccessfully tries to enter the United States from a foreign nation on several occasions over a relatively short period of time. This person is detained or blocked from entry into the United States due to association with terrorism activities. This person may be acting alone or with others -- other group members and/or relatives. More than one border bouncer may try to enter the U.S. from different locations over a particular period of time... Some guidelines for the task include: use the definition of border bouncer we present above. It may not match how you know other people or organizations addressing the problem of searching for border bouncers, but it is applicable for this task. We are interested in the bouncers themselves and associated people of interest. Canonicalization issues of people, place, or organization names may occur. There is no intentional deception in this dataset, although there may be false leads. Please identify these as such if you find any.

On the surface, this may appear to be a straightforward task of finding repetitious names in the dataset. However, when canonicalization issues and individuals working with other individuals as a team are introduced, the problem becomes more difficult.

Two initial considerations we make in creating a dataset include the type of analytical tool we are targeting to test and the subtlety of the threat to be embedded in the data. For this dataset, we are clearly targeting tools that can handle large amounts of multivariate data. As part of the subtlety considerations, we initially made the dataset two orders of magnitude larger than the current distribution size, but were asked to reduce the size for initial tests. We also did not include any explicit deceptions, although there is a false lead that appears suspicious, but does not match the given definition of a border bouncer.

As part of our data generation process, the threat data can be generated by hand or can be created using our generation

software, or both. When we determine exactly the characteristics of the threat data to be included in the dataset, the generation method is selected. For example, for Spinosaurus, both methods were used. Specific people fitting the definition of border bouncer were created by hand and inserted into the dataset by hand. The false lead was generated by the software by selected manipulation of parameters.

Of course, the majority of the data in the dataset represent the surrounding data environment. Some refer to this data as “noise”, but this noise must be carefully understood and generated so as to maintain the required suspension of disbelief for an analyst. NTCDD has 28 fields and 200 records, although we have generated a 2 million record version. The fields include names, addresses, vehicle descriptions, actions taken, and miscellaneous information such as classification level and accuracy rating. Here are some examples of NTCDD data fields and the considerations required to generate them:

- Incident Number: Simply a random, non-repeating integer from 1 to 9999 to index a record. The software can also generate random numeric data using a variety of distributions including normal, chi-square, Poisson, and Student. This data can also be used as indices into tables of other types of data.
- Incident state and zip code: Where the activity occurred. This represents a simple dependency that needs to match reality. We maintain two tables of reference data - zip code by city, and cities in states. We select a city and state according to a desired distribution (or no city, when that data is to be left blank), and then match the city to an eligible zip code.
- Incident longitude and latitude: This represents a sparse data collection, where less than half of the entries would include this data. We select the number of data elements that will contain geo-spatial data, distribute the selection across the data, and then perform a city-state lookup for longitude and latitude for that entry.
- Incident date and date of entry: The rule for these fields is that the incident date must precede the date of entry and the date of entry must be prior to a selected date. The dates generated for these fields must be relevant to the analytical problem, so the incident date is selected between 2003 and 2004. The date of entry could be between a day to a couple years following the incident date.
- Resolution description: While our software does not yet have the capability to generate free text, text replacement gives the user the ability to write up templates which can then be completed using values from other fields. This field uses multiple templates with full or partial names taken from the Suspect field. We generate items such as “Marcian Macrosson was stopped by police while attempting to take suspicious items into target location” and “Pauray was released after a detailed inspection of vehicle turned up nothing suspicious.”

Dependencies among data elements is a large challenge, and we continue to improve our approach to specifying semantic associations among data. Another example of this in Spinosaurus

reflects the canonicalization issue, or more generally, names with aliases. We are able to specify a relationship such as: generate a person, assign a primary name, and allow data records to be generated using the primary name and up to two aliases. This process involves generating temporary fields that do not get included with the final dataset.

The threat data and the environment data are merged to create a final dataset. Various tools are used to review the data set against characteristics of interest. One characteristic is temporal distribution of data elements. We have created a timeline viewer that maps one temporal dimension of the data (such as Incident Date), so that we can view the range, distribution, and clustering tendencies of the data. It also allows visual editing of the data so that data set developers can move one or more records around on the timeline.

4. DIRECTIONS

We have created six TSG data sets to date, shown in Figure 2. Our first set, Trilobite, is a homogeneous collection of about 1400 text data articles. It presents a terrorism threat in a historically accurate setting. Triceratops is a multivariate data set with an invented scenario of political intrigue. This data set is the basis

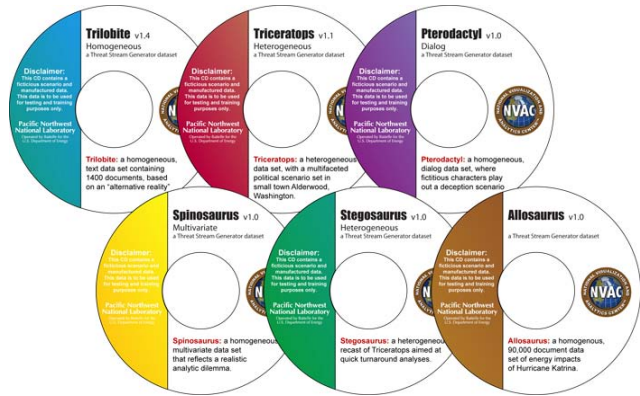


Figure 2. Threat Stream Data Sets

for the 2006 VAST contest. Pterodactyl is a dialog data set with fictitious characters, featuring deception-laden activities. Allosaurus is a 90,000 document data set featuring energy issues resulting from Hurricane Katrina.

The goals for our work for NVAC include the creation of a collection of synthetic data sets with embedded threats that may be used for test and evaluation of visual analytic tools, a viable threat stream generator software system, a mathematic model describing the process from scenario creation through analysis, and a process of evaluating the resulting data sets so that we know what benefits we bring to the evaluation process.

5. REFERENCES

- [1] National Visualization and Analytics Center website. <http://nvac.pnl.gov>
- [2] Visual Analytics Science and Technology Symposium contest website. <http://www.cs.umd.edu/hcil/VASTcontest06/>