

Grand Challenge Award: Data Integration Visualization and Collaboration in the VAST 2008 Challenge

Donald Pellegrino*
Drexel University

Chi-Chun Pan
Penn State University

Anthony Robinson
Penn State University

Michael Stryker
Penn State University

Junyan Luo
Penn State University

Chris Weaver†
Penn State University

Prasenjit Mitra
Penn State University

Chaomei Chen
Drexel University

Ian Turton
Penn State University

Alan MacEachren
Penn State University

ABSTRACT

The VAST 2008 Challenge consisted of four heterogeneous synthetic data sets each organized into separate mini-challenges. The Grand Challenge required integrating the raw data from these four data sets as well as integrating results and findings from team members working on specific mini-challenges. Modeling the problem with a semantic network provided a means for integrating both the raw data and the subjective findings.

KEYWORDS: Visual analytics, investigative analysis, intelligence analysis, information visualization, geovisualization, data integration.

INDEX TERMS: E.1 [Data]: Data Structures—Graphs and networks; H.5.3 [Information Systems]: Information Interfaces and Presentation—Group and Organization Interfaces

1 INTRODUCTION

The 2008 VAST Challenge introduced a new format with four mini-challenges in addition to a Grand Challenge, all using synthetic data sets constructed to represent real world problems. The mini-challenges were self-contained problems that could be worked on independently. The Grand Challenge however required integrating all of the mini-challenge data and providing a higher-level assessment of the data. The heterogeneous nature of the mini-challenge data sets gave each of them unique properties that could be exploited by customized visualizations. Making sense of the full scenario required aggregating the data and findings in a way that facilitated making connections across the mini-challenges. Modeling the problem as a semantic network provided a representation that preserved the properties of the original data while supporting the addition of mini-challenge subjective findings necessary to build the aggregate high-level assessment.

2 METHODOLOGY

To address the challenge a team of researchers was brought together from the North-East Visualization and Analytics Center (NEVAC), a Regional Visualization and Analytics Center coordinated from Penn State in State College, PA and Drexel University in Philadelphia, PA. The team immediately recognized the need for a Computer Supported Collaborative Work Environment. To meet this need an Adobe Connect system was used to support synchronous collaboration and a Wiki instance was created to support asynchronous collaboration. A Wiki environment was selected to create familiarity with the Wiki data that was provided as part of the challenge and to simulate experiences that might be encountered by professional analysts working with Intellipedia. Team members took individual

responsibility for separate mini-challenges. Specific responsibility was also assigned for the Grand Challenge and the data integration that it required.

The first of the four data sets was in Wikipedia Page History Format. The data was in plain text and did not include the links to the historic versions of the pages. Additionally the final version of the page was not included. Although a Wikipedia formatted page was included in the collection the history did not apply to that particular page. This left the analyst to make sense of the list of changes without reference to the content or full context of the change itself. To address this mini-challenge a customized *Improvise* [1] visualization was created and is shown in Figure 1. This customized system used multiple coordinated views for exploratory analysis. This system was supplemented by a custom implementation of algorithms for visual analysis of controversy as described by Brandes and Lerner in [2]. A k-Means clustering was performed on the results of Brandes and Lerner's algorithm. These clusters were then analyzed and compared with the *Improvise* views to develop high-level hypotheses. The assessments were written to the project's Wiki page and shared with the rest of the team. Additional details on the team's approach to the Wiki mini-challenge are given in [3].

The second mini-challenge data set was provided in XML format and included synthetic data on Coast Guard interdictions as well as landings of migrant vessels. As with the wiki mini-challenge data this set was analyzed using both data immersion and computational techniques. A customized, map-based *Improvise* visualization was created to interactively explore geographic, temporal, and other patterns in the data and is shown in Figure 2. This proved to be an extremely useful approach to answering the questions in this mini-challenge as they required macroscopic overviews of the full data with various amounts of temporal and spatial binning. These assessments were written to the project's Wiki page. This data set was also transformed into a customized Google Earth KML file which provided another view of the macroscopic patterns.

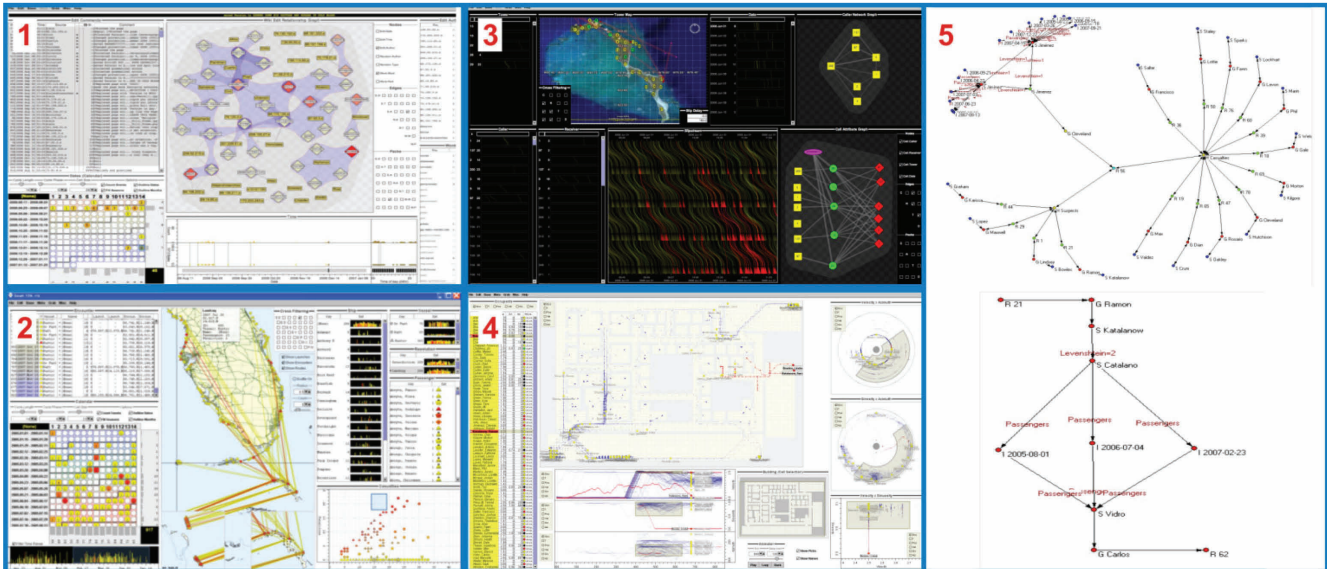
A structured file formatted using comma separated values was provided to report cell phone calls placed on an island during a ten day period. Although names of the owners of the phones were not provided the challenge instructions provided hints to the phone identifiers that might be involved in the primary story line. Again a customized *Improvise* visualization was created and is shown in Figure 3. This was supplemented by computational approaches.

The final mini-challenge data set contained locations of RFID badges over the time immediately before and immediately after an explosion occurring inside a building. The customized *Improvise* visualization created for this mini-challenge was particularly sophisticated as many derived variables, such as velocity were calculated and added to the visualization shown in Figure 4. The values and relationships for these variables were examined closely using the interactive features of the tool.

The challenge data package included additional supplementary material such as a Wikipedia page on the "Paraiso Manifesto." These additional unstructured data sources functioned as a fifth data set.

*e-mail: don@drexel.edu

†e-mail: cweaver@psu.edu



Integration of the data and findings was done by using an associative network as the fundamental data structure. This provided the greatest degree of abstraction while preserving the critical connectedness between the different types of data. A transform was created for each of the four mini-challenge data sets. The transforms created nodes in the network to represent the entities from the source data and edges to represent their connectedness. The hypotheses and assumptions captured in the Wiki were represented as derived nodes and edges in the network. These constructs helped to assign higher-level meaning to the data making the model a semantic network rather than simply a set of associations. By combining higher-level constructs such as hypotheses with the raw data, analysis results became more useful. The top panel of Figure 5 shows Cleveland Jimenez as both a suspect and a casualty. The bottom panel of Figure 5 shows the result of analyzing the network with Pajek revealing how two RFID badges may be related through passenger rosters.

3 REFLECTIONS AND FUTURE DIRECTIONS

Although the data sets in the 2008 VAST Challenge were significantly different from earlier years there were some similarities in the methodology applied here and the approach used by the South-East Regional Visualization and Analytics Center (SRVAC) team in 2007 [4]. The SRVAC team used a divide-and-conquer approach by assigning subsets of the reports to be read by individual team members. In the 2008 challenge the mini-challenge data sets provided a natural way to decompose the problem. The mini-challenge specific tools provided exploration and data immersion for the structured data sets. The effect of the divide-and-conquer approach is to distribute the subjective knowledge across the team. Stasko et al describing Jigsaw in [5] state that "Trial use of the system also suggest the need for better tools to help analysts organize their thoughts and document the models and plans they are constructing." This issue is partially addressed by explicitly capturing the analyst's notes in the Wiki environment. It is notable that the Stories module for GeoTime described in [6] provides a similar means for organizing such thoughts. The methodology described here builds on the Wiki and Stories shared discursive approaches by modeling the hypotheses formulated by team members during work on the mini-challenges into a unified semantic network. Modeling the hypotheses in the semantic network could also be viewed as an extension to the Analysis of Competing Hypothesis tabular structure described by Heuer in [7]. The semantic network model has the advantage of

contextualizing the hypotheses and evidence with the rest of the data and providing a data structure that can facilitate data analysis.

There are a number of limitations to the methodology described here. A customized visualization was developed for each mini-challenge data set. While Improvise provided a framework for quickly creating customized visualizations the tool requires training and experience to develop proficiency. This could be a limiting factor for situations where the data diversity is much wider and expert designers are unavailable. Construction of the semantic network required writing custom transforms for each of the mini-challenge data sets. Addition of the nodes and edges representing the hypotheses and other assumptions was a manual processes. Tighter integration of the semantic network, the visual analysis applications and other tools, the Wiki, and report writing may increase usability and is an opportunity for future research.

ACKNOWLEDGEMENTS

This work is supported by the National Visualization and Analytics Center, a U.S. Department of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL). PNNL is a U.S. Department of Energy Office of Science laboratory.

REFERENCES

- [1] C. Weaver, "Building Highly-Coordinated Visualizations in Improvise," in *IEEE Symposium on Information Visualization*, Austin, TX, 2004, pp. 159-166.
- [2] U. Brandes and J. Lerner, "Visual analysis of controversy in user-generated encyclopedias," *Inf Visualization*, vol. 7, pp. 34-48, 2008.
- [3] C.-C. Pan, D. Pellegrino, C. Weaver, and P. Mitra, "VAST 2008 Wiki Editors Mini Challenge - Identifying Social Networks using Wiki.viz," in *IEEE VAST '08* Columbus, OH, 2008, p. DVD.
- [4] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Jigsaw meets Blue Iguanodon - The VAST 2007 Contest," in *IEEE VAST '07* Sacramento, CA, 2007, pp. 235-236.
- [5] J. Stasko, C. Görg, Z. Liu, and K. Singhal, "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," in *IEEE VAST '07* Sacramento, CA, 2007, pp. 131-138.
- [6] R. Eccles, T. Kapler, R. Harper, and W. Wright, "Stories in GeoTime," *Inf Visualization*, vol. 7, pp. 3-17, 2008.
- [7] R. J. Heuer, Jr., "Analysis of Competing Hypotheses," in *Psychology of Intelligence Analysis*. Central Intelligence Agency, 1999, pp. 95-110.