The Case for Visual Analysis Provenance Cases

T.J. Jankun-Kelly
Department of Computer Science and Engineering
Mississippi State University
tjk@acm.org

ABSTRACT

While visual analytics holds potential for assisting users to make timely sense of vast quantities of complex data, research into providing such tools is stymied by the lack of real-world records of the process used by analysts. Visualization researchers are able to develop tools and gather usage and insight data from their collaborators, but analytical provenance "in the large" is missing. In this position paper, I briefly outline the benefits and need for a collection of the visual analysis process, and end with a call to action.

Author Keywords

visual exploration, visual analytics, provenance

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*Theory and Methods*

THE GROWING FOCUS ON PROCESS INFORMATION

Visual analytics combines data processing, visualization, and analysis in order to make decisions and solve tasks. Like all analysis problems, it is an iterative process with possible loops, restarts, and conclusions [12, 13]. The visualization and the analysis is ideally a tight loop, where exploration of the visualization parameter space sparks new insights and analysis of the previous results suggests new avenues of investigation or conclusions. As such, methods to model and capture the user's visualization exploration facilitate understanding the possible insights garnered via that exploration.

Over the past decade, there has been a growing interest in methods to capture, represent, or analyze the visual exploration process. Broadly, these can be categorized into two approaches: Tree-like representations of the process (e.g., VisTrails [1] and Tableau [4]) and graph-based representations (e.g., the GDE [10] and P-Set [6] models). All of these capture **the fundamental operation of visual exploration:** The application of a set of parameter values to the visualization transform to generate a visualization result [6]. However, lacking from these models is a notion of the insight generated by these interactions—the fruit of the visualization. Recent work has begun to address this issue [7]. A melding of these two approaches is the goal of the analytical provenance research thrust. Some initial success in this area I overview next, to demonstrate the benefits of this research.

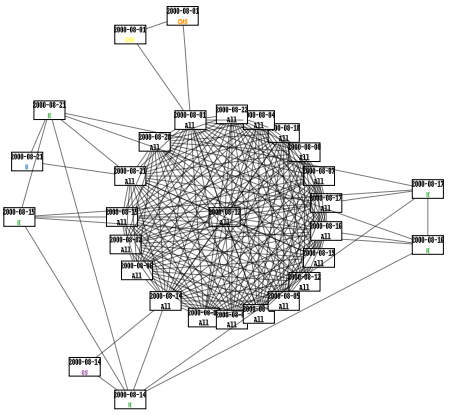

Figure 1. By gathering metrics from usage data of a network visualization tool, the visualization was improved [5]. This visualization depicts a more efficient traversal of the visualization parameter space.

BENEFITS OF PROCESS ANALYSIS

Metrics for Visual Exploration

Given a formal representation of a visual exploration, it is natural to attempt to quantify its effectiveness. To this end, quantitative metrics are used as a starting point to evaluate the session. Both the GDE [9] and the P-Set model [5] provide such metrics. These use different graph measures to describe the visualization: How redundant it was, how broad it was, and so on. Figure 1 demonstrates one of graphs generated during this analysis: It shows cliques within parameter sets, were a link indicated a shared parameter value. This graph is significantly tighter than the graph of the original visualization of the same data, showing a less scattered exploration. Using the feedback from the various metrics, the redesigned visualization increased the overall efficiency of exploration (a measure of the lack of redundant result generation) from 70% to 85% in common cases [5].

Provenance for Improving Exploration

While the metrics above were used to improve the design of a visualization (and thus the insights gained from it), provenance information has also be used to improve the process directly. VisTrails [1] captures the changes of parameters

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

and visualization pipelines during exploration. To accelerate the creation of visualizations, the VisTrails team collected a database of over 2,800 pipelines from their provenance in order to suggest possible pipelines as a user constructs them [8]. Their predictions eliminated 30–75% (50% on average) construction operations. Such optimizations do not lead directly to insight from the visualization, but they do accelerate the discovery of such insights.

Extracting Insight from Exploration

Even without explicit recording of gathered insights, the strategies, methods, and findings of a visual analysis can be extracted from provenance. Dou et al. [3] used provenance (video logs of a think aloud protocol combined with recordings of the fundamental operation of visualization) from a financial visualization system [2] to determine what level of the original insights could be derived from the process information alone. Four additional analysts examined only the provenance and recovered a 60–79% of the original insights generated by the explorations. It is a first step in automating the provenance capture of insight, which will lead to insight-based metrics and exploration suggestions.

CALL TO ACTION: PROVENANCE NOW!

There are obvious benefits for collecting visual exploration process data for understanding insight. Unfortunately, researchers are limited by the lack of real-world data available upon which to base their insight process research. None of the project aforementioned have released their provenance data to the research community: Not the single system data from the P-Set work, the 30 student's explorations with VisTrails, or the 200 minutes of transcribed video and provenance from the WireVis analysis. While privacy, confidentiality, and informed consent concerns likely prevent the release of this data, a concerted effort must be made to gather a repository of the exploration and insight provenance in order to further our research. The NSF has begin a similar effort in gathering and sanitizing sensitive computer security data for that research domain [11]; their lessons can be used to spearhead such an effort for our field. Three steps are needed:

- **Provide Provenance Software** The lower the barrier to capture process information and annotate it, the more data will be generated. VisTrails is open source and the P-Set software is available on request, but integrating such software into toolkits such as protovis should be pursued.
- **Get Big Players On Board** ManyEyes, Tableau, and Oculus have multitudes of users generating data an insight. While a difficult problem, an effort should be made in order to capture "scrubbed" data from these players to assist our research. This was one of the main concerns of the NSF workshop [11]—getting corporate buy-in.
- **Get Some Data Now** While we wait for the a solution to the scrubbing issue, a stop-gap solution can be employed: Gather data from the VAST Challenge. Requiring (or strongly requesting) a digital record of the exploration process and the insights garnered (other than the final written report) would provide a wealth of open data for further analysis. The VAST data itself has proven a boon for the development of tools; the gathering of contest's insight process would as well.

A concerted effort in establishing visual analytics is now paying dividends; our next investment should be in understanding how to better improve the insights derived from such analysis.

REFERENCES

1. Bavoil, L., Callahan, S. P., Crossno, P. J., Freire, J., Scheidegger, C. E., Silva, C. T., and Vo, H. T. VisTrails: Enabling interactive multiple-view visualizations. In *Proc. of the Vis '05* (2005), 135–142.
2. Chang, R., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D., and Sudjianto, A. WireVis: Visualization of categorical, time-varying data from financial transactions. In *Proceedings of IEEE VAST 2007* (2007), 155–162.
3. Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R., and Chang, R. Recovering reasoning processes from user interactions. *IEEE Comput. Graph. Appl. 29* (May 2009), 52–61.
4. Heer, J., Mackinlay, J., Stolte, C., and Agrawala, M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Trans. on Vis. and Comp. Graph. 14*, 6 (2008), 1189–1196.
5. Jankun-Kelly, T. J. Using visualization process graphs to improve visualization exploration. In *Proc. of IPAW '08*, Springer (June 2008), 79–91.
6. Jankun-Kelly, T. J., Ma, K.-L., and Gertz, M. A model and framework for visualization exploration. *IEEE Trans. on Vis. and Comp. Graph. 13*, 2 (March/April 2007), 357–369.
7. Kadivar, N., Chen, V., Dunsmuir, D., Lee, E., Qian, C., Dill, J., and Shaw, C. Capturing and supporting the analysis process. In *Proc. IEEE VAST '09* (2009), 131–138.
8. Koop, D., Scheidegger, C. E., Callahan, S. P., Freire, J., and Silva, C. T. VisComplete: Automating suggestions for visualization pipelines. *IEEE Trans. on Vis. and Comp. Graph. 14*, 6 (Nov./Dec. 2008), 1691–1698.
9. Lee, J. P. *A Systems and Process Model for Data Exploration*. PhD thesis, U. of Mass. Lowell, 1998.
10. Lee, J. P., and Grinstein, G. G. An architecture for retaining and analyzing visual explorations of databases. In *Proc. of the Vis '95* (1995), 101–108.
11. NSF. Dear colleague letter: Trustworthy computing. `http://www.nsf.gov/cise/news/2011_trustworthy.jsp`.
12. Pirolli, P., and Card, S. K. Information foraging. *Psychological Review 4* (1999), 643–674.
13. Thomas, J. J., and Cook, K. A., Eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, 2005.