Analysis Within and Between Graphs: Observed User Strategies in Immunobiology Visualization

Caroline Ziemkiewicz
Brown University
Providence, RI
cziemki@cs.brown.edu

Steven Gomez
Brown University
Providence, RI
steveg@cs.brown.edu

David H. Laidlaw
Brown University
Providence, RI
dhl@cs.brown.edu

ABSTRACT

We present an analysis of two user strategies in interactive data analysis, based on an observational study of four researchers in the immunology domain. Screen captures, video records, interviews, and verbal protocols are used to analyze common procedures in this type of visual data analysis, as well as how these procedures differ among these users. Our findings present a case where skilled users can approach a similar problem with diverging analysis strategies. In the group we observed, strategies fell within two broad categories: within-graph analysis, in which a user generates a few graph layouts and interacts heavily within them, and between-graph analysis, in which a user generates a series of graphs and switches between them in sequence. Differences in strategies lead to distinct interaction patterns, and are likely to be best supported by different interface designs. We characterize these observed strategies and discuss their implications for scientific visualization design and evaluation.

Author Keywords

Visualization; immunology; task analysis.

ACM Classification Keywords

H.5.2 [Information Interfaces And Presentation]: User Interfaces - Evaluation/Methodology;

INTRODUCTION

Evaluating visualization presents unique difficulties, in large part because realistic visualization problems are difficult to capture in traditional usability testing [4]. Often, visualization is used to perform open-ended analysis. They may not know precisely what they are looking for, and they may not know how to tell when they have found it. Under these conditions, people can easily find patterns that do not correspond to real phenomena in the data [7]. In real-world analytical and scientific visualization applications, these complex user tasks are not special cases; they are the core reason people use the software. These tasks are both a challenge to evaluate and one reason visualization is a critical application area. A better understanding of these tasks and how to support them through design would improve visualization practice, advance theory and user models, and impact the many fields in which visualization is a vital part of the analysis pipeline.

Immunology is one such field, and one in which we have been collaborating with researchers in the CBDM Laboratory at Harvard Medical Center. These researchers study the genetic factors that affect immune response in mice as a model for the human immune system with the goal of understanding autoimmune disorders such as diabetes and arthritis. As part of an ongoing effort to model this type of scientific visualization task, we observed four researchers from this lab performing typical data analyses. These researchers manage copious data from a range of experimental procedures, including gene expression data and T-cell counts in mice from varying genetic lines. In our observations, datasets ranged from 25,109 to 46,632 genes, but experiments can generate up to 100 samples from 30 populations for each gene, producing tens of millions of data points that cannot be simultaneously analyzed using existing tools. Currently, our users study these data using a web-based tool called GenePattern [5] to generate scatterplots and highlight groups of genes based on filtering criteria known as signatures. Discovering the gene signature that is characteristic of a particular variation in immune system behavior is the primary goal of this type of analysis.

Our results revealed that, even when users have similar goals and are using the same software, different participants used clearly diverging strategies. We characterize the strategies we observed as within-graph and between-graph analysis. In within-graph analysis, the user changes the layout rarely but interacts heavily with individual data points through selection and filtering. In between-graph analysis, the user continually generates new layouts. If selection is used, it is used to facilitate comparisons between graphs in sequence. These differences in interaction style suggest these strategies would be best supported by different visualization designs.

Close qualitative analyses of the kind presented here have made valuable contributions to the visualization field. For example, Pirolli and Card's cognitive task analysis [3] produced the Sensemaking Loop model for intelligence analysis, which has become a guiding principle in the visual analytics domain. Springmeyer et al. performed a task analysis [6] of scientific data analysis that placed visualization use in the context of the larger research pipeline and were among the first to argue that visualization systems should include a way to record users' analytic processes.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

Figure 1. The GenePattern analysis environment. Users can generate one or more scatterplots based on their data. Here, P3 is viewing a plot of changes in gene expression between two study conditions (right) and a volcano plot of the significance of those changes (left). The blue highlight identifies a specific gene signature.

The present work contributes to this body of knowledge by focusing on a specific scenario in scientific visualization that illuminates issues in designing for a variety of users. As Plaisant has argued [4], such case studies of specific use scenarios are a necessary step towards improving the metrics by which we evaluate visualization. From the data we collected, we find that it is insufficient to evaluate how well a visualization supports this particular task without acknowledging major differences in approach among users. Rather than using idealized tasks as a model for visualization evaluation and design, we suggest further study of user strategies.

METHODS

Our analysis is derived from an observational study of users in the immunology domain. We observed our users in a scenario as close as possible to the day-to-day analysis they perform when no observers are present. Participants were given an hour to analyze their own data and were encouraged to use the software and methods they were accustomed to while a researcher observed and recorded their behavior.

Observation

The participants were four postdoctoral researchers in the CBDM Laboratory, recruited during a visit to the lab by an author who requested volunteers for a human-computer interaction study. These four volunteered since they were, at the time, working on the data-analysis portion of their research. The participant group included two women and two men, all with comparable levels of experience in their field and all working on different but related experiments. Each of these users was at a similar point in the workflow of their overall project. Each had an individual experiment they had performed, and their data had recently been returned from an outside laboratory that perform processing on the gene expression results they produced. In this phase of the project, they were looking through those processed results to test hypotheses and come up with new ones for follow-up work. The four observation sessions took place during a single day and were performed on a single workstation. The primary analysis tool used by these researchers is a web-based system for gene expression analysis called GenePattern [5], specifically the Multiplot Visualizer function (Figure 1). Multiplot was used to generate scatterplots showing correlations between two gene expression variables as well as "volcano plots," used to show p-values for a given analysis. One of the participants (P2) also used Tibco Spotfire S+ [1] to generate data tables.

Participants were asked to perform a typical analysis for no more than an hour and to stop when they were satisfied with their results. Completion times ranged between 24 minutes (P2) and 46 minutes (P4). P3 and P4 both exclusively analyzed new data they had not seen before, while P1 and P2 analyzed a mix of new and older data. All of their analysis goals fit a similar mold: find a gene signature that is represented to different degrees in a control group of mice and in the test group. Prior to analysis, the participants were asked to briefly describe their research problem and the data they would be analyzing. While the participants performed their analysis, a researcher observed and recorded their actions. We captured video of the screen during the analysis as well as video and audio of the participant. The observer also took notes during this session, and would periodically prompt the participant to describe what he or she was doing. After the analysis, the observer questioned the participant in a brief unstructured interview about their analysis methods.

Analysis

The observation sessions yielded nearly five hours of video and accompanying notes. This video was coded for analysis in two passes. The first pass, performed by two coders, was a high-level analysis of interaction patterns based on the coding scheme used by Springmeyer et al. [6], expanded to include the low-level visual analytic tasks identified by Amar and Stasko [2]. A second low-level coding pass was performed by one coder to record quantitative data about the number and duration of graphs generated, and of changes to filtering and highlighting schemes. We considered a graph to be new if the variables used as axes were changed. This pass also recorded changes to the filtering and highlighting schemes used in the scatterplots and, if two graphs were visible at once, the duration of the secondary graph.

RESULTS AND DISCUSSION

For the two coders who performed the high-level coding pass, overall agreement as measured by Cohen's κ was 0.44, indicating moderate agreement. Disagreement was largely driven by uncertainty about what a participant was doing during periods where they observed a graph; for example, when looking at a scatterplot, is the participant performing a Correlate or Characterize Distribution action, or both? This is a difficult limitation when studying visualization use through recordings or interaction logs, and future work may benefit from the use of eye-tracking to address this problem. Nonetheless, the results of the high-level coding pass showed a general repeated analysis pattern that was common to all participants while using GenePattern: viewing a graph, filtering, and retrieving individual values. Essentially, participants were looking for patterns in their data, identifying a possible gene signature of interest, and examining specific genes in that signature to test their hypotheses.

Although there were common patterns in the four participants' behavior, there were substantial differences in the de-

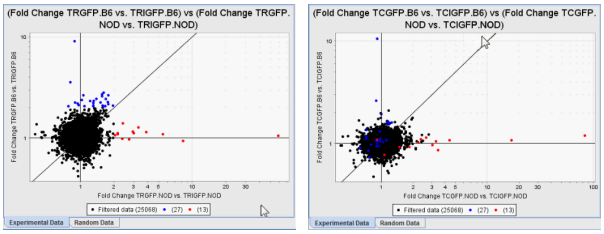

Figure 2. Two steps in P1's interaction, before and after a change in graphs. P1 used highlights such as the red and blue selections in this example to compare gene expression behavior across multiple conditions.

tails of their analysis. P2's session was somewhat anomalous: he preferred to use the scripting capabilities of the S+ system rather than perform visual analysis using GenePattern at all. This made his analysis session much shorter, since he was applying previously developed analysis scripts, and suggests that he may represent a type of user who either would not benefit from visualization or is not inclined to adopt it. Among the other three, we found that P1 and P4 exhibited similar patterns of behavior, while P3 behaved differently enough that she appeared to be using a different strategy altogether. P3 used a relatively small number of primary graphs with frequent use of secondary graphs and constant changes in highlighting schemes. P1 and P4 generated greater numbers of graphs in quick succession and sometimes used fixed highlighting to facilitate comparison between graphs. We propose that these differences in behavior reflect different approaches to visualization, which we characterize as analysis *within* a graph and analysis *between* graphs. Examining the users' strategies in detail shows how they differ both in terms of behavior and what they need from a visualization design.

P3: Analysis Within Graphs

Uniquely among the participants, P3 always had two scatterplots visible at once (Figure 1). Over the course of a 47-minute session, she generated eight primary and four secondary graphs. Each was visible for a fairly long period of time: the average duration of each primary graph was 261 seconds and that of a secondary graph was 523 seconds. She changed the highlighting scheme a total of 28 times and set seven filters. This indicates that, while she used multiple views, she maintained each overall layout of those views for a long period of time. The bulk of her interaction involved changing the appearance and filtering parameters of existing graphs, not generating new graphs. For this reason, we consider her analysis as primarily taking place within a limited number of graph layouts, rather than extending across a sequence of graphs. This within-graph strategy treats the visualization as a fixed externalization of the analysis, and focuses interaction on individual data points rather than layouts. It emphasizes understanding information in context and making visual connections between parts of the data.

The large number of highlights performed by P3 arise partly because GenePattern does not let the user make a highlighting change across multiple graphs at once. Therefore, whenever she wanted to change the highlight in both views, she had to change her highlighting scheme twice. While she did not do this for every highlight change she made, it was a common action. Essentially, she was performing manual brushing and linking between views. This, along with her use of multiple graphs, suggests that P3 was trying to make her system behave like a coordinated multiview visualization.

Also unusual in P3's analysis is that she saved her graphs as image files five times over the course of her analysis session; none of the other participants saved or otherwise recorded their graphs. This bolsters the observation that P3 tended to treat the visualization as the primary output of her analysis, rather than input to a more abstract analysis process. When asked why she used multiple views, P3 responded that it was important to see both behavior and significance at once. "If I don't have two [views], I have to go back and forth. Going back and forth, you can forget and lose time." She added that she would prefer four views to two, but that there wasn't enough room on the screen. In this, P3 shares characteristics with P2. Although P2 generated numerical tables rather than graphs, he also focused on presenting as much data at once as was possible. In some ways, P3 is the ideal user for whom visualization researchers tend to design: interested in maximizing the amount of information on screen at once, focused on efficiency, and open to many multiple views. However, not every user is like P3, as P1 and P4 demonstrate.

P1 and P4: Analysis Between Graphs

A different strategy was used by P1 and P4. For these users, most interaction focused on switching between graphs, rather than interacting within a fixed layout. Neither participant ever had more than one scatterplot visible at a time, and both changed graphs more often than P3. Over a 32-minute analysis, P1 created 12 primary graphs, each of which was visible for an average duration of 90 seconds. On the other hand, P1 created only nine highlighting schemes and three filters. P4 created 14 graphs, with an average duration of 121 seconds, and performed no highlighting or filtering actions. By at least one measure, P1 viewed the same amount of information as P3: both viewed a total of 12 graphs over the course of their analysis, although P3 viewed several of those graphs in parallel. However, focusing interaction on switching between views indicates a different conceptualization of the analysis problem. P3 preferred to have a large amount of information on screen at once, and manipulate that information within a set visual layout. P1 and P4 preferred to manipulate the visual layout itself, treating each graph as a step in the analysis.

A common action for P1 was to highlight one or two groups of genes in a color or colors, and then change the axes of the graph to see how that group behaves under different views (Figure 2). At a high level, this is the same analysis that P3 performed by setting these two graphs next to each other and applying the same highlight scheme to both. From one point of view, then, P1 is just being less efficient than P3. However, when asked about his use of visual analysis, P1 revealed another possible explanation for viewing the graphs in sequence rather than in parallel: "I like to turn the data upside down and sideways, looking for 'realness.' If you try different plots, different views, and still see something, you can be more reassured that these genes are differentially expressed." This

perspective suggests that changing graphs frequently might be a way to increase confidence in a result. While P3 worried that going back and forth would make her forget something, P1 treated that very process as a way to confirm or disconfirm his hypotheses. Similarly, when P4 encountered an error in her data during the course of her analysis, she followed up her discovery by quickly switching between several views (contributing to her high number of graphs). If a user is concerned that a pattern seen in one view might be biased or illusory, replacing that view entirely with another could be a tactic to view the data with fresh eyes. This view of the value of interaction differs from that usually emphasized in visualization research, yet recalls the graphical inference strategy proposed by Wickham et al. [7] as a method for determining the significance of visual patterns in data.

For all of these researchers, the final output of this process is expected to be a set of rigorously validated statistical results showing differences or lack thereof between pairs of experimental conditions. Since visualization is an exploratory tool seen as an intermediate step between the experiment and a statistical result, researchers are given a good deal of freedom in how they use those tools provided they lead to a validated finding. Analyses are considered productive if they are rigorous and fast, but validation is always the primary concern. Still, the participants emphasized different priorities for judging a good analysis tool. P1 was concerned with increasing confidence in his results, and P4 emphasized validity of the data as her primary concern. P3 was more concerned with efficiency and viewing lots of data quickly. This perspective was shared by P2, who preferred to use non-visual statistical software in his analysis. These priorities appear to be reflected in the reasoning behind their analysis strategies.

Broader Implications

Our findings add to the body of evidence that, in complex analytical environments, designing for only the "average user" is not realistic. The more complicated and abstract a goal is, the more likely it is that users will have different yet equally reasonable approaches to it. Further research on how individual differences such as personality and cognitive ability affect visualization use may help us to predict these differences in approach. Another challenge raised by these results is that different analysis strategies are likely to require different interface designs. The within-graph strategy we observed would clearly be better served by a coordinated multi-view visualization. P1 and P4's between-graph strategy would benefit from a system that allows for quicker changes between views, and possibly one which supports animated transitions to avoid change blindness. In applications at this level of complexity, analyzing the varying strategies within a single task allows for a more complete picture of users' needs.

Limitations

These results are from a small-scale observational study that focuses on only four users, and there is clearly a limit to how much they can be generalized. Although our participants fell into two categories, there is no guarantee that future users can be sorted into these categories as well. We hypothesize that within-graph and between-graph analysis can be extended to describe most user strategies in this domain, and perhaps in related domains as well, but controlled studies will be needed to test this hypothesis. Additionally, the open format of our observation sessions means that other factors may have been at play in the differences between participants. Participants who studied a mix of old and new data may not perform the same type of analysis as those studying only new data. The fact that P4 encountered an error in her dataset, and spend a portion of her session diagnosing the problem, is also likely to have affected her interaction behavior. While we acknowledge these limitations, they are a consequence of focusing on a few highly realistic scenarios for close study.

CONCLUSION

Our observations revealed that, in the scenario we studied, expert users can approach the same problem and the same system with diverging analysis strategies. One user focused on within-graph interaction, manipulating data within a few graph layouts; two others preferred between-graph interaction, rapidly switching between views. These findings illuminate the analysis tasks and interface needs of users in a critical domain. As immunologists progress towards building a road map of gene expression networks in the immune system, the size of these data will become increasingly overwhelming, as will the number of new users attempting to make sense of it. A complete understanding of how these users and their analysis strategies vary will make it possible to design analytical tools to support this vital area of research.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Christophe Benoist and colleagues for their participation and kind assistance. This material is based upon work supported by the National Science Foundation under Grant No. CIF-B-195. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Tibco Spotfire S+. Accessed 9/21/2011.
2. Amar, R., Eagan, J., and Stasko, J. Low-level components of analytic activity in information visualization. In *Proceedings of IEEE InfoVis* (2005), 15–23.
3. Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (2005).
4. Plaisant, C. The challenge of information visualization evaluation. In *Proceedings of AVI* (2004), 109–116.
5. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. GenePattern 2.0. *Nature Genetics 38*, 5 (2006), 500–501.
6. Springmeyer, R. R., Blattner, M. M., and Max, N. L. A characterization of the scientific data analysis process. In *Proceedings of IEEE Visualization* (1992).
7. Wickham, H., Cook, D., Hofmann, H., and Buja, A. Graphical inference for infovis. *IEEE TVCG 16*, 6 (2010), 973–979.