Systematic Yet Flexible Discovery: Guiding Domain Experts through Exploratory Data Analysis

Adam Perer, Ben Shneiderman
Human-Computer Interaction Lab
Department of Computer Science
University of Maryland
College Park, MD 27042,USA
Tel: 1-301-405-2769
[adamp,ben]@cs.umd.edu

ABSTRACT

During exploratory data analysis, visualizations are often useful for making sense of complex data sets. However, as data sets increase in size and complexity, static information visualizations decrease in comprehensibility. Interactive techniques can yield valuable discoveries, but current data analysis tools typically support only opportunistic exploration that may be inefficient and incomplete.

We present a refined architecture that uses *systematic yet flexible* (SYF) design goals to guide domain expert users through complex exploration of data over days, weeks and months. The SYF system aims to support exploratory data analysis with some of the simplicity of an e-commerce check-out while providing added flexibility to pursue insights. The SYF system provides an overview of the analysis process, suggests unexplored states, allows users to annotate useful states, supports collaboration, and enables reuse of successful strategies. The affordances of the SYF system are demonstrated by integrating it into a social network analysis tool employed by social scientists and intelligence analysts. The SYF system is a tool-independent component and can be incorporated into other data analysis tools.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Design, Human Factors

Keywords: systematic yet flexible, guides, wizards, information visualization, social networks, exploratory data analysis

INTRODUCTION

The increasing availability of digitized information encourages users to conduct more frequent and complex exploratory data analyses. The basic string search or SQL query are no longer adequate for advanced users who seek to understand patterns, discern relationships, identify outliers, and discover gaps.

IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.

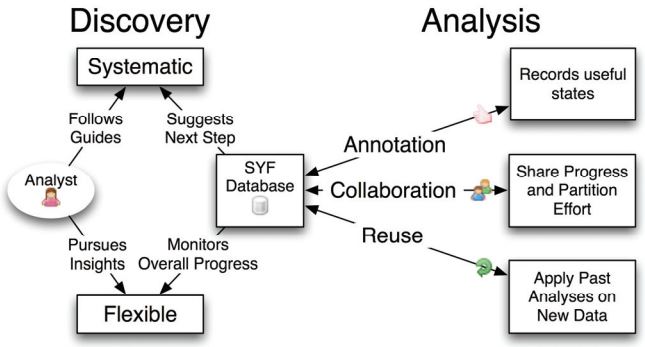

Figure 1. The SYF infrastructure facilitates discovery by providing systematic guides while also allowing users to flexibly pursue insights. SYF also facilitates analysis by allowing users to easily annotate during exploration, share exploration results with colleagues and partition effort, and reapply past exploration paths on new data sets.

Data mining strategies, cluster analysis, and search engine results are helpful tools for such exploration, which typically takes days, weeks, or months. Domain experts may be trying to sift through gigabytes of genomic data to understand the causes of inherited disease, to filter legal cases in search of all relevant precedents, or to discover behavioral patterns in social networks with billions of people. For these challenging tasks, users must conduct repeated searches, combine results, and consult with colleagues. As they grow familiar with the data, they move from divergent conjectures to more careful hypothesis testing so as to collect evidence supporting their emerging insights.

Current tools can produce useful nuggets of information, but domain experts are increasingly aware of the need to shift from opportunistic discoveries to more systematic approaches. A *systematic* approach guarantees that all measures, dimensions and features of a data set are studied. Such an approach guides new users, ensures analysts of completeness, and facilitates cooperation during analyses that may take weeks or months. However, a wholly strict

guide would undermine the *flexible* needs of an analyst, as they will inevitably wish to pursue insights based on past successes, new information, fresh hypotheses, or unproductive directions.

Legal searchers, who need to find every relevant case to avoid surprises, have developed paper-based and sometimes electronic tools to guide their work. Their goals are to ensure complete coverage, allow measurement of their progress, and enable team members to combine their partial results. Another expectation of systematic approaches is that they allow different users working independently to come up with largely similar results. Other professional examples demonstrating a similar need for systematic analysis include physicians completing diagnostic examinations, field biologists surveying forest grounds, and forensic scientists investigating murder scenes. Such professions have developed orderly strategies to assist investigators with challenging, non-trivial, multi-faceted exploration.

Systematic-only approaches may suit some users' needs, but complex problems rarely yield to clean algorithmic strategies. If real problems were that simple, their solutions could be automated. Thus, systematic yet flexible strategies are emerging as a key topic in areas such as survey completion, job applications, and business process modeling. Such strategies are all the more central in the e-science community, where scientific workflow management and record keeping are issues of vital importance. E-science researchers must also address long duration projects, collaboration complexities, and guarantees of completeness [18, 36].

Most computer users have some form of experience with systematic interfaces, as they are pervasive in many common activities. The checkout process at Amazon.com [2], shown in Figure 2, provides an overview of the four steps users are required to complete before making a purchase. The process is simple and systematic, but inflexible in that it requires users to complete their purchase following a strict order of operations, as part of a one-time process which does not allow them to return to or revise entries weeks later.

A more sophisticated interface is Intuit's TurboTax [19], which guides users safely through the complex U.S. Internal Revenue Service tax filing procedures. TurboTax steps users through the process of entering required information. The top of the interface, shown in Figure 3, features secondary navigation tabs that allow users to complete steps in any order, in case they should wish to make changes or review previously entered information. The top of the interface presents an overview of users' expected tax refunds or debts owed, and updates after each question is answered. TurboTax then verifies that all appropriate forms are filled out before allowing users to print or file their taxes. While flexible to user preferences, the TurboTax system still does not explicitly track user progress for presentation in the header overview.

Inspired by these approaches, our goal is to enhance the tools available for data analysis with *systematic yet flexible* (SYF) support. Data analysis is not as simple as a purchase on a website or filling out tax forms, so we present seven design goals to handle these more challenging tasks. We integrate these design goals into our tool-independent SYF infrastructure. This infrastructure supports discovery through *systematic* and *flexible* exploration, as well as annotation, collaboration, and process reuse (Figure 1). This integration supports orderly exploration over weeks, record-keeping to support discovery claims, and collaboration with colleagues. We also support the iterative process of returning to review earlier work and bold initiatives that break from the formulaic approach.

Social Network Analysis

We demonstrate the benefits of SYF by integrating it into a tool for social network analysis, *SocialAction* [24]. Sociologists, intelligence analysts, communication theorists, bibliometricians, food-web ecologists, criminologists and numerous other professionals are interested in understanding social networks. Network analysts focus on relationships

Figure 2. The 4 systematic steps for checkout at Amazon.com [2]. Users must step through all four stages in order, while the progress is this process is updated at each step.

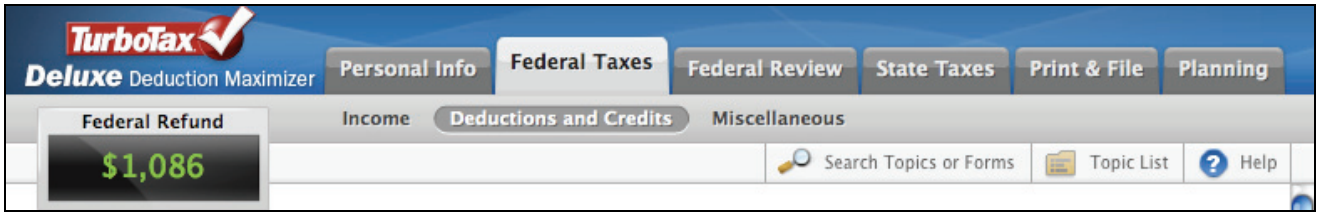

Figure 3. Intuit's TurboTax [19] guides users through the complex process of preparing tax returns in the United States. The top of the interface features tabs that allow users to complete steps in their own order, in case want to make changes or review. The interface also presents an overview of the user's expected refund or debt owed, and updates after each step.

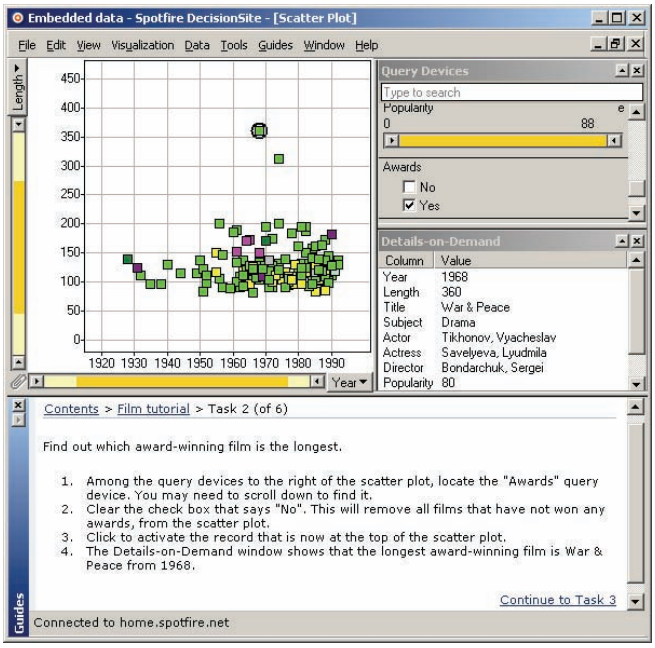

Figure 4. Spotfire, a commercial information visualization software package, allows end-users to create guides for exploring data. In this example, the guide is located on the bottom of the interface. The guide describes the current task, provides instructions on how to manipulate the data, and offers a hyperlink to the next task after users believe they have finished.

instead of just their individual elements; how the elements are put together is just as important as the elements themselves. In many previous studies, sociologists focused largely on behavioral attributes and neglected the social facets of behavior (how individuals interact and the influence they have on each other) [9]. Using newer techniques employed by the social network community, analysts can now find patterns in the structure, witness the flow of resources through a network, and learn how individuals are influenced by their surroundings.

The maturity of social network analysis tools has not advanced as fast as the popularity of social network analysis. Numerous measures have been proposed by structural analysts to statistically assess social networks [39]. With a wealth of metrics, analysts want to be certain they are not overlooking critical facets of the networks in question. A design that allows social network analysts to quickly iterate and keep track of computed metrics is critical for exploring these vast statistical measures. The ability to share results, annotate key findings and reapply past measurements on new networks allow past efforts to not be wasted. The SYF system provides such benefits critical to social network analysts.

RELATED WORK

Guides for Complex Tasks

When exploring large networks of information, maintaining a path history and providing guides can improve navigation.

The World Wide Web is a one such vast repository of information, which users navigate with hyperlinks and view pages with browsers. Most browsers feature history mechanisms, including a visual cue of changing the hyperlink's color once it has been visited. This technique is effective at alerting users to pages they have already visited, so they need not bother visiting them again [37]. Google's Notebook [12] and Grokker's Working List [13] enable easy recording of web pages that can be saved or sent to others.

However, as a task's complexity increases, more sophisticated guides can alleviate the inevitable struggles of users. "Wizards" are a common type of interface that, instead of informing users how to perform a task, break the task into a linear series of steps. This interface strategy is most successful for tasks that have standard solutions; that is, when a simple step-by-step procedure leads to success [8]. Users often wish to turn off wizards after they have learned a task, and research suggests that users have trouble transferring knowledge gained from wizards to a non-wizard environment [6]. Furthermore, secondary navigation is often preferred to allow users to complete the steps in their own order, and is featured in some commercial software (e.g. Intuit's TurboTax) [6].

Another type of guide is an adaptive interface that reduces the complexity of tasks by "understanding" the user's needs and simplifying the interface [22]. In practice, the algorithms supporting adaptive interfaces are often simple, such as Microsoft Office's Adaptive Menus, which hide the least recently used items. COACH provides pre-coded, in-context guidance, captured from demonstrations that were based on observing user behavior [28]. DocWizards allows users to more easily create *follow me documentation wizards* by learning from demonstrations using a task model [3].

For complex document assembly tasks, some systems will provide an overview of what is needed, so users can see their progress and make informed choices about what their next steps should be. For example, the U.S. National Science Foundation FastLane provides such guidance for the 20+ components that research teams must submit in grant proposals, with feedback about the last update for each component.

However, there have been few approaches specifically designed for data analysis. Spotfire, a commercial information visualization software package, allows end-users to create guides (Figure 4) [34]. After the process of analysis has been understood, end-users can compose Guides to help automate repetitive procedures and ensure consistency among analysts [35]. Spotfire Guides are presented as a series of hyperlinks that assist users in preparing data, opening standard visualizations, sorting data and even removing outliers. However, the guides do not monitor the actions of users and thus do not provide a measurement of progress. Another approach is by Groth and Streefkerk who describe a prototype system without guides that records the history of user explorations in a visualization tool, as well as the capability for users to annotate their exploration [14].

Network Analysis Tools

Since we are applying SYF to a social network analysis tool, we provide a brief review of available network analysis tools. There are dozens of software tools designed to help analysts understand social networks, such as [4, 5, 7]. These tools often feature an impressive number of analysis techniques that users can perform on networks. However, they are also often a medley of statistical methods and overwhelming visual output that leaves many analysts uncertain about how to explore their terrain in an orderly manner. Social network analysis is an inherently deductive task, and a user's exploratory process can be distracted by having to navigate between separate analysis and visualization packages.

Recently, there have been several projects focusing on improving interactive exploration within networks, although not necessarily focused on social networks. Among them, *GUESS* is a novel graph exploration system that combines an interpreted language with a graphical front end [1]. *TreePlus* allows users to explore graphs using more comprehensible enhanced tree layouts [21]. *NetLens* allows users to explore an actor-event network using iterative queries and histograms [20]. Ghoniem et al. presented the promise of using matrix-based visualizations instead of node-link diagrams [11]. *JUNG* is a JAVA toolkit that provides users with a framework to build their own social network analysis tools [23]. *NVSS* addresses the challenge of node layout by using attributes of nodes, where user-defined semantic substrates act as regions for nodes that share similar attributes [31]. *MatrixExplorer* is a recent system designed for exploring social networks using a matrix visualization as the primary view [17].

THE SYF INFRASTRUCTURE

Systematic yet flexible design goals

We propose a set of seven design goals shown in Table 1. The first four goals provide *systematic yet flexible* discovery support by ensuring analysts of completeness and guiding novices. The last three goals support analysis by enabling annotations, collaboration and reuse. Each of these goals support analysts who work over many days, weeks, or months. Furthermore, these design goals emphasize maintaining concentration to achieve task completion [32]. By showing users their prior, current and future steps, users are assisted when returning after inevitable distractions.

In order to facilitate the integration of SYF principles into data analysis tools, we provide an open-source infrastructure to tool developers. First, the tool developers register the systematic steps of exploration via SYF's application programming interface (API). Then, they register GUI events from their tool using the API and specify which steps the events belong to. SYF keeps track of user progress by maintaining a history of GUI events invoked. After developers augment their application with the SYF user interface, they can easily provide users with an overview, progress feedback, history navigation, annotation support, and the additional features listed in Table 1.

Systematic Yet Flexible Design Goals
Enable users to:
1. See an overview of the sequential process of actions
2. Step through actions
3. Select actions in any order.
4. See completed and remaining actions
5. Annotate their actions.
6. Share progress with other users.
7. Reapply past paths of exploration on new data.

Table 1: Seven design goals for systematic yet flexible interfaces

Supporting Discovery with Systematic Yet Flexible Guides

When users are exploring data, there are many paths and permutations to examine and users can easily get lost. The SYF system provides feedback to users about their current state, the actions they have already completed, and which actions remain. This information gives confidence to users that they have made progress through the rich landscape of data analysis.

The SYF system, which augments a data analysis tool's interface, provides an overview of each of the systematic steps for completeness (Design Goal 1). The left-hand side of Figure 5 presents *SocialAction 3.0*'s seven systematic steps for social network analysis derived from practitioner interviews.

Users who wish to explore the data via SYF's *systematic* guiding can use the navigation buttons, also found on the left-hand side of Figure 5. When users are ready to continue analysis, they can click the 'Next' button to bring them to the next unvisited state (Design Goal 2), or return to a previous state using the 'Back' button. If users wish to explore the data in a *flexible* way, each step button acts as a secondary navigation button, much like a tab. Users can click this button to navigate to the actions required to complete the step (Design Goal 3).

Each step button features a progress bar. These meters give users a sense of how far away they are from completing the current step, as well as the entire data analysis (Design Goal 4). If users wish to view their path of exploration so far, they can launch the history panel. In Figure 6, a user's history is shown as a tabular list that is sortable by step number, state type, user action or annotation rating.

Users can also hide the SYF panel if they wish to focus on their work. By dragging the divider panel that separates SYF and the data analysis tool, they can shrink or minimize

the guide. Even when the SYF interface is hidden, the user's actions be monitored so the benefits of SYF can be leveraged later.

SYF In Action: SocialAction 3.0's Node Rankings

One step in the defined systematic social network analysis path is ranking all nodes according to importance metrics. In Figure 5, an analyst has completed 40% of the current step. In order for users to finish this step, they must examine the rest of the node importance rankings. Information about completed rankings is not isolated to the SYF interface, but can also be integrated into the main UI of *SocialAction 3.0*. For instance, the combo box in which users select importance rankings are augmented with icons highlighting previously visited options (Figure 7). If users have already examined a ranking, a checkmark appears beside it. Similarly, if users have already made an annotation about this ranking, an annotation icon appears. *SocialAction* can look up this information about each ranking state by using the SYF system's API. Informing users in a consistent manner is important, as many users prefer to use secondary navigation instead of following all steps in order, depending on their hypotheses or experience [6].

Supporting Analysis with Annotation

Throughout the process of exploring data, users may come across important discoveries. The SYF system features a light-weight solution for users to annotate these insights quickly (Design Goal 5). Annotations are textual comments such as indications of insights, notes about informing partners about progress, or questions to be asked to collaborators. Often, annotations will deal with schedules, deadlines, reminders of tasks to be done, or the need to prepare for presentations. Useful annotations might be attached to objects being studied, such as indications of relative value of legal precedents or chemical structures. We augment these textual annotations with ratings and tags so they can be easily found later.

During any stage of data exploration, analysts have access to the annotation functionality shown in the panel on the left of Figures 5 and 8. This persistent panel allows users to quickly comment, rate and tag any state of analysis. Users can write their insights in the enlargeable text editor.

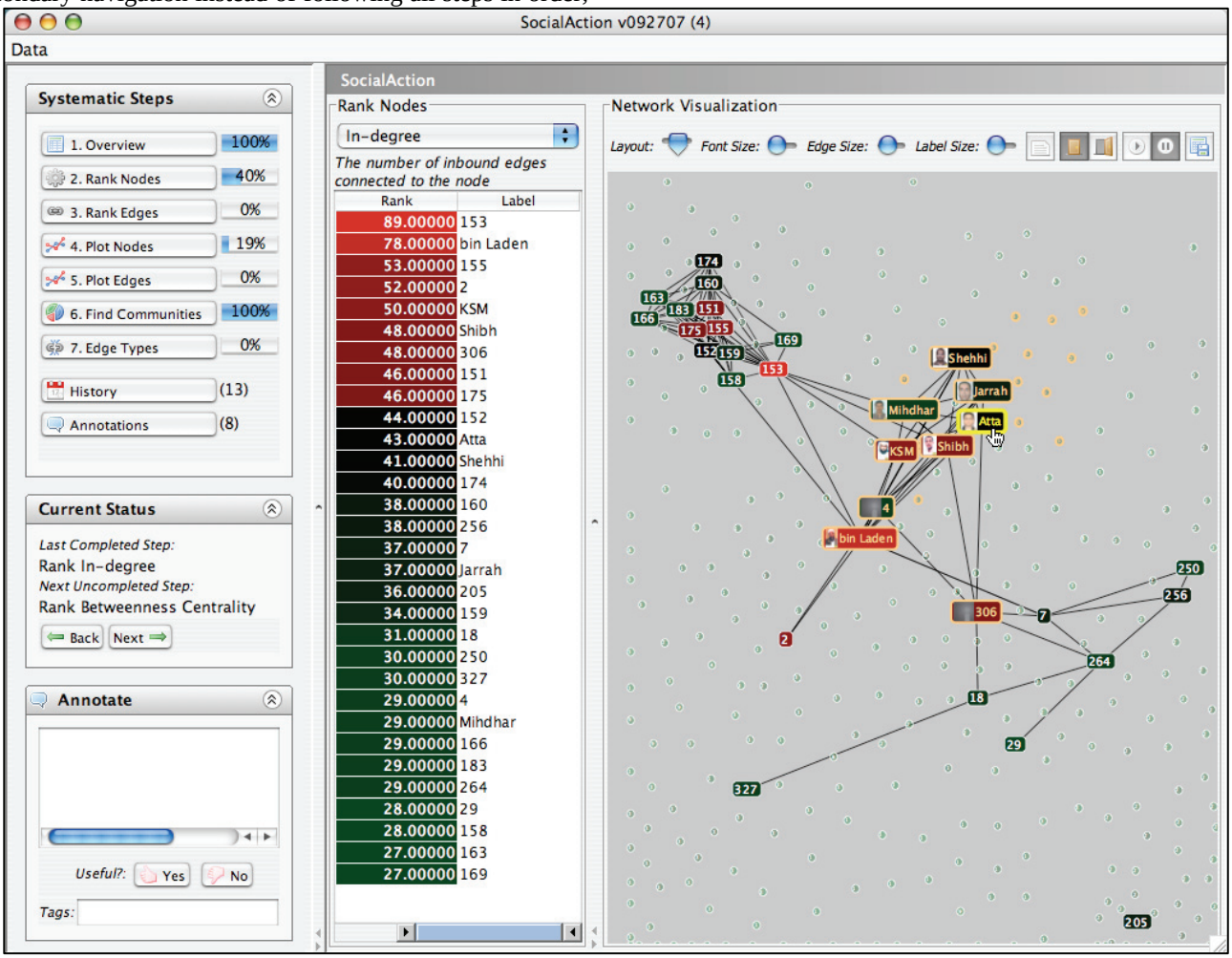

Figure 5. The SYF system integrated into *SocialAction 3.0*, a social network analysis tool. The interface to SYF is presented on the left-hand side, whereas the main UI for *SocialAction 3.0* is on the right. This figure features a "Global Jihad" terrorist network that researchers are studying using *SocialAction*. In order to protect sensitive information, node labels have been anonymized except for those individuals publicly identified in the Zacarias Moussaoui trial.

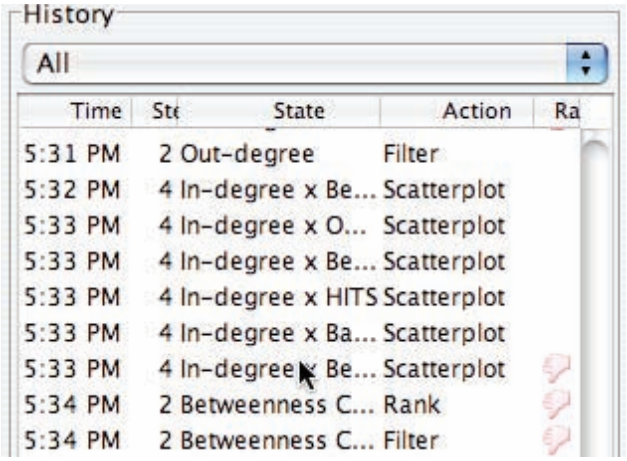

Figure 6. SYF's History panel shows users' past actions in tabular form. Users can navigate by sorting by step number, state type, user action or annotation rating. A 'Date' column also appears when analysis takes place over multiple days. Furthermore, users can filter based on "important" or "unimportant" annotations using the combo box at the top. Users can jump back to a previous state by clicking the 'Go' button.

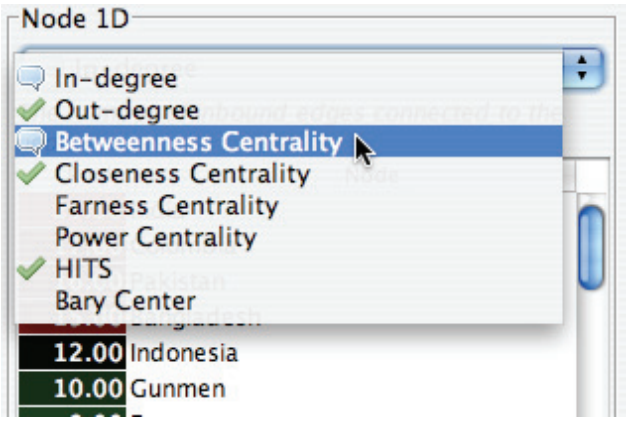

Figure 7. The Combo Box in the SocialAction 3.0 GUI provides feedback in the form of a checkmark icon to show which measures have been computed previously.

Users can also mark a state as interesting via the 'thumbs up' button, uninteresting via the 'thumbs down' button, or tag the state with meaningful words or phrases. Users can also choose to mark this state and comment with a tag in the 'Tag' text field. Whenever users return to an annotated state of analysis, the annotations reappear automatically in this space.

Users can review all past annotations by clicking the annotation button located below the systematic steps. The number next to the annotation link informs users how many annotations have been composed. In the annotation panel, shown in Figure 8, users can browse all annotations, keyword search for specific annotations, navigate using the tag cloud for tagged comments [15], or filter based on rated interestingness. Users can select individual annotations from a sortable, tabular list where they can read the comment or jump back to the state where the annotation was written.

In addition to allowing users to return to interesting states for further exploration, annotations are useful when users wish to create reports about their findings. Since useful discoveries have been recorded, users can export the images, tables and descriptions associated with interesting states into word processors or web pages.

SYF In Action: SocialAction Communities

We illustrate our annotation functionality in another step of social network analysis: community detection. One of the main goals of sociologists studying social networks is to find cohesive subgroups of nodes [10]. *SocialAction*'s algorithms automatically determine communities based on their link structure, to help users find groups of nodes that are closely connected in the network. Communities are visually represented by surrounding all members with a translucent convex hull as shown in the right side of Figure 8.

In this example, the user is browsing all annotations created with SYF. The tag cloud shows the user's tags for all annotated states, and the tabular list shows all annotations. The last annotation is selected and displayed below. The user activated this state by clicking the "Go" button and can review and continue analysis.

Supporting Analysis with Collaboration

New evidence has emerged suggesting that communication and collaborations are necessary components of successful visualization systems [38]. User studies also suggest that supporting collaboration with visual data analysis can help people explore a data set both broadly and deeply [16].

The SYF system supports collaboration by allowing users to easily share their exploratory paths and insights that were annotated during their data analysis (Design Goal 6). Since SYF monitors each interaction and allows users to specify useful states, analysts can easily export interesting states to colleagues. Furthermore, users can partition effort during analysis. After users finish a segment of analysis, they can share their completed results. Recipients will know which analyses have been performed and annotated and will be empowered to not duplicate past efforts.

Supporting Analysis with Reusable Exploration

In addition to user-to-user collaboration, SYF also supports data-to-data collaboration. Users can repeat analyses conducted on previous data to new data sets (Design Goal 7). For instance, if a user already found several useful states during exploration and marked them as useful in the annotation panel, they could reuse these "best practices" on new

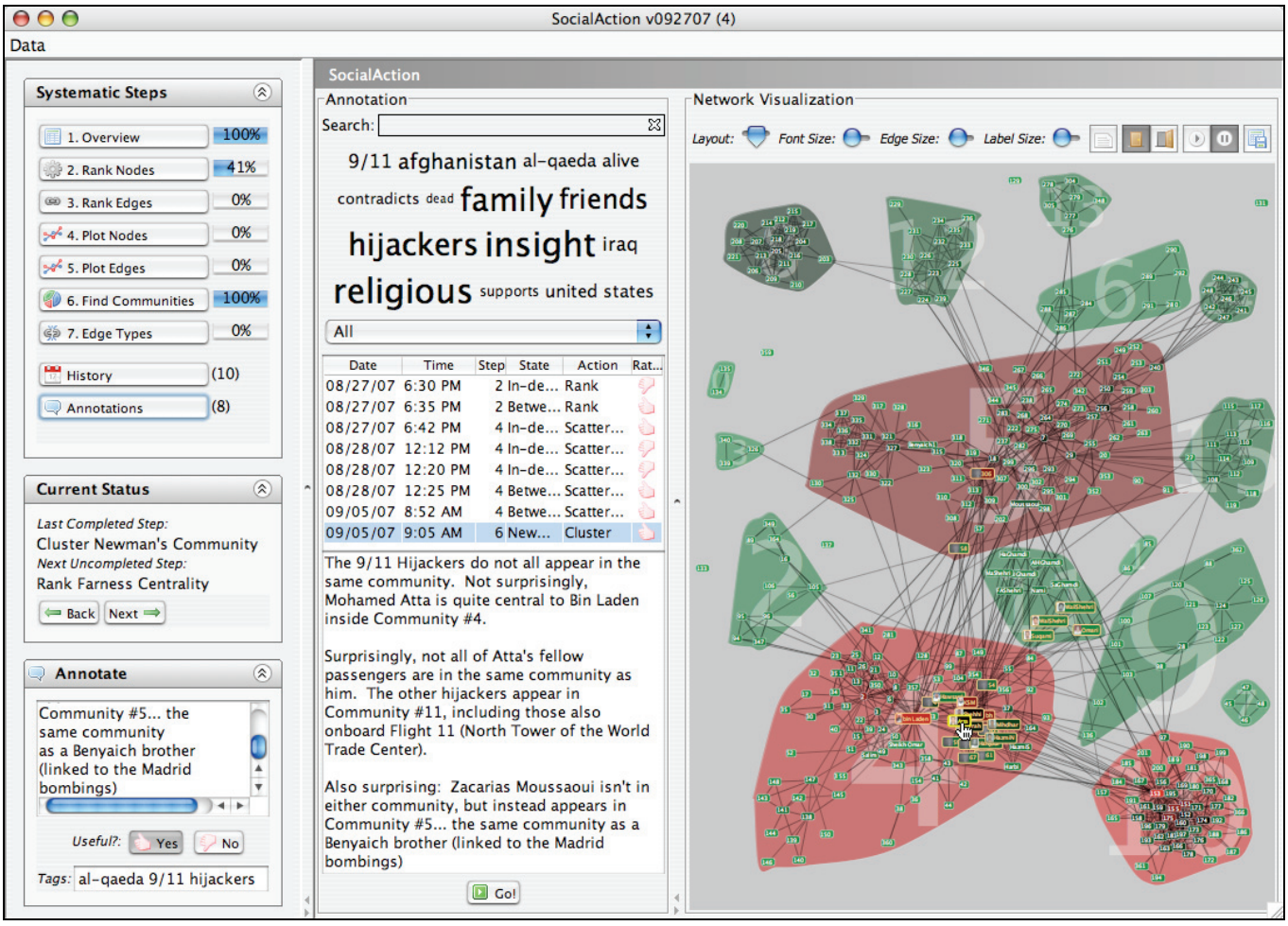

Figure 8. This figure is shows the annotation features of SYF. Users can browse their annotations by selecting the annotation button located at the bottom of the systematic steps panel. Users can keyword search, navigate using the tag cloud, or filter based on the rating to find specific comments. When a user selects an annotation from the resulting tabular list, it is displayed below. Users can jump to the state where the annotation occurred by selecting the 'Go' button.

data, as if it was a macro. Analysts can quickly see if the same patterns, gaps or outliers are present in the new data set.

SYF In Action: Comparing Networks in SocialAction

We illustrate an example of reusing past exploration in *SocialAction 3.0*. This example comes from colleagues studying social networks that span thirty years. In order to grasp the dynamics of a network, they often study a year's data independently and then make comparisons to other years. Instead of repeating calculations on every year manually, SYF allows these analysts to automatically compute and present analysis after the first exploratory path has been defined. Social scientists often collect and input data manually and sometimes the visualizations present coding mistakes in the data. In this situation, users need to fix the mistakes in the original dataset. Instead of starting over from the beginning, analysts can use SYF to reapply all past analyses and continue to make progress.

DEFINING A SYSTEMATIC PATH TO COMPLETENESS

Understanding the domain experts' tasks is necessary to defining the systematic steps for guided discovery. Although some professions such as physicians, field biologists, and forensic scientists have specific methodologies defined for accomplishing tasks, this is rarer in data analysis. Interviewing analysts, reviewing current software approaches, and tabulating techniques common in research publications are important ways to deduce these steps.

For instance, even though there are many importance rankings, clustering algorithms, and statistical techniques for assessing social networks, there is no well-defined methodology for performing these operations [39]. During the design of *SocialAction* we conducted in-depth interviews with six social network practitioners to understand their current work habits. Since most social network practitioners were not using visualizations during their exploratory analysis, these findings were augmented with several key principles from the information visualization community. The tenets of the Visual Information Seeking Mantra [30] ("Overview first, zoom and filter, details-on-demand") were kept in mind when ordering the tasks of social network analysts. Furthermore, the Graphics, Ranking, and Interaction for Discovery (GRID) principles [29] ("Study 1D, study

2D, then find features. Ranking guides insight, statistics confirm") also shaped our systematic method for analyzing social networks.

The resulting 7-step methodology for social network analysis, integrated into *SocialAction 3.0*, is:

1. Overall network metrics (e.g. number of nodes, number of edges, density, diameter)
2. Node rankings (e.g. degree, betweenness, closeness centrality)
3. Edge rankings (e.g. weight, betweenness centrality)
4. Node rankings in pairs (e.g. degree vs. betweenness, plotted on a scattergram)
5. Edge rankings in pairs
6. Cohesive subgroups (e.g. finding communities in networks)
7. Multiplexity (e.g. analyzing comparisons between different edge types, such as friends vs. enemies)

This is not the only systematic method for social network analysis, but one that will assure analysts they have explored relevant features in *SocialAction 3.0*. This methodology is evident in the SYF user interface that augments *SocialAction 3.0* (left side of Figures 5 and 8).

PRELIMINARY EVALUATION

Although computer applications, such as *SocialAction*, shift from productivity support to creativity support, research evaluation methods are still predominantly based on older strategies. Controlled experiments with dependent variables such as time for performance of benchmark tasks are still valuable, but they may be inadequate to study tools that support creative exploration [26, 27]. These new tools may require substantial learning, changes to problem-solving strategies, and exploratory use of tactics that defy controlled experimentation.

New research evaluation methods based on ethnographic observation and longitudinal study are being refined to meet the needs of these type of tools [33]. We are using Multidimensional In-depth Long-term Case studies (MILCs) with academic and professional social network practitioners. These long-term case studies shift the strategy to working with small numbers of domain experts over longer time periods. We are using these MILCS to understand the power of the *systematic yet flexible* ideas for data analysis.

Four long-term case studies were conducted using *SocialAction 2.0* which did not yet have SYF support [25]. In these case studies, participants received a training session for two hours in order to become proficient in the data exploration techniques, followed by weekly interviews. In these interviews, the authors would often demonstrate certain features of *SocialAction 2.0* they had not yet explored. Typically, the participants were exploring the social networks with their best practices from previous software tools. However, as new features were demonstrated in the interviews, the participants were often excited and eventually led to new hypothesis testing. These experiences inspired us to develop the *systematic yet flexible* infrastructure as part of *SocialAction 3.0*.

Initial feedback from social network practitioners using SYF is promising. Some partners were dubious that systematic methods will lead them to discover new insights, as they believed examining every permutation might be significantly more work for little added benefit. However, after using the software, they noticed the ease of exploring the rich features of *SocialAction 3.0* by clicking "Next". This led them to examine measurements they might have otherwise skipped and think about their data in new ways. Their current software tools made it difficult to initiate new permutations, so they often carefully and cautiously chose standard routes of exploration. This freedom to explore, with the system keeping track of their hard work, excited them to the point of considering reanalyzing data they previously analyzed with other tools.

Our MILC partners are also enthusiastic about the annotation, collaboration and reusable functionality that the SYF infrastructure provides. These features are missing from each of the analysis tools they have used in the past, so they believe such features could impact how they conduct research in the future.

As expected, not every partner agrees with the systematic steps defined for *SocialAction*. However, these partners also admit the flexible freedom that SYF affords to work in any order but to provide a route when they feel lost. They hope they will also have the freedom to customize the SYF steps once they find paths of exploration that are most effective for their needs.

IMPLEMENTATION

The SYF system is tool-independent and implemented in JAVA. It features an API that interface designers can use to integrate with any software, not just social network analysis tools. After developers set up their application with the API, SYF stores all linked application state events into a history database. Similarly, an indexed database of all annotations is maintained for fast searching and browsing. When users save their analysis, the original network file and both databases are stored into a flat file for easy distribution. The SYF then maintains internal databases after importation.

Among its features, SYF allows developers to specify their own systematic steps, provides a consistent way to log and navigate to interface state events, and an easy way to integrate collaboration.

FUTURE WORK

To date, the SYF system has only been integrated into one data analysis tool, *SocialAction*. However, since the SYF system is designed as a modular component, we are planning to integrate it into other data analysis tools as well. Several tool designers that were given a preview of the SYF system immediately saw the benefits it would offer to their users. In addition to providing guides, developers would

obtain critical features that users demand for free, such as history keeping and supporting "undo".

We also plan on advancing the current collaboration functionality we offer. Although users can take turns and share their exploration, we offer no way to merge them if they are concurrent. We are interested in looking at ways to support small groups (2-10) and larger teams (10-100) of researchers who work together.

Expert users might also wish to rearrange or design their own steps for social network analysis. Currently, step design is left up to the developer using the API. However, since most expert users are end users and not developers, it makes sense to afford them this capability. This feature would also be useful in allowing users to compose smaller steps for more specific tasks. If analysts are only interested in a small subset of measurements, having a way to measure progress based on those instead of the overall features is important. For these reasons, we are building a systematic customization feature for experts.

DISCUSSION AND CONCLUSION

Systematic yet flexible support has implications beyond data analysis tools. Wizards and tabs are pervasive in the user interfaces of many applications. SYF combines the *systematic* properties of wizards with the *flexible* properties of tabs, while providing users feedback about progress. For any interface that requires steps to be completed, and where order of completion is not restricted, we believe the SYF interface would improve the user experience.

In conclusion, we present a novel user interface infrastructure to provide support for the challenging task of data analysis. To assist discovery, SYF offers *systematic* guides that provide users the ability explore relevant analytical features. SYF also supports *flexible* diversions to pursue insights while still maintaining overall progress. To assist analysis, SYF provides annotation, collaboration and reuse capabilities. These three tasks offer analysts a way to record, share, and more easily find new insights. After all, data analysis is all about finding the useful nuggets. SYF still relies on human analysts to find these nuggets, but empowers them by maintaining their history, measuring their progress, and most importantly, keeping them informed.

ACKNOWLEDGMENTS

We thank Georg Apitz, Bita Azhdam, Nicolas Chen, Krzysztof Gajos, François Guimbretière, Tessa Lau and Catherine Plaisant and the anonymous reviewers for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 0633843.

REFERENCES

1. Adar, E. GUESS: a language and interface for graph exploration. In *Proc. SIGCHI conference on Human Factors in computing systems*. ACM Press (2006), 791-800.
2. Amazon.com *Amazon.com*. http://www.amazon.com, (2007).
3. Bergman, L., Castelli, V., Lau, T. and Oblinger, D. DocWizards: a system for authoring follow-me documentation wizards. In *Proc. ACM symposium on User interface software and technology*. ACM Press (2005), 191-200.
4. Borgatti, S., Everett, M. G. and Freeman, L. C. *UCINET 6*. Analytic Technologies, (2006).
5. Brandes, U. and Wagner, D. *visone - Analysis and Visualization of Social Networks* In *Graph Drawing Software*, M. Junger and P. Mutzel. Springer-Verlag (2003).
6. Burton, M., Wickham, D. P., Phelps, L., Kelly, S., Kelly, Crews, J. and Rich, N. Secondary navigation in software wizards. In *Proc. CHI '99 extended abstracts on Human factors in computing systems*. ACM Press (1999), 294-295.
7. de Nooy, W., Mrvar, A. and Batageli, V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge (2005).
8. Dryer, D. C. Wizards, guides, and beyond: rational and empirical methods for selecting optimal intelligent user interface agents. In *Proc. International conference on Intelligent user interfaces*. ACM Press (1997), 265-268.
9. Freeman, L. C. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press (2004).
10. Freeman, L. C. *Graphic Techniques for Exploring Social Network Data* In *Models and Methods in Social Network Analysis*, P. J. Carrington, J. Scott and S. Wasserman. Cambridge University Press, Cambridge (2004).
11. Ghoniem, M., Fekete, J.-D. and Castagliola, P. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In *Proc. IEEE Symposium on Information Visualization*(2004).
12. Google *Notebook*. http://www.google.com/notebook/, (2007).
13. Grokker *Working List*. http://www.grokker.com, (2007).
14. Groth, D. P. and Streefkerk, K. Provenance and Annotation for Visual Exploration Systems. *IEEE Transactions on Visualization and Computer Graphics*, 12, 6 (2006), 1500-1510.
15. Hassan-Monteroa, Y. and Herrero-Solanaa, V. Improving Tag-Clouds as Visual Information Retrieval Interfaces. In *Proc. International Conference on Multidisciplinary Information Sciences and Technologies*(2006).
16. Heer, J., Viegas, F. B. and Wattenberg, M. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization. In *Proc. ACM Conference on Human Factors in Computing Systems*. ACM Press (2007), 1029-1038
17. Henry, N. and Fekete, J.-D. MatrixExplorer: A Dual-Representation System to Explore Social Net-

works. *IEEE Transactions on Visualization and Computer Graphics*, 26, 5 (2006), 677-684.
18. Hodgman, C. An information-flow model of the pharmaceutical industry. *Drug Discovery Today: BIOSILOCO*, 1, 6 (2001), 1256-1258.
19. Intuit *TurboTax 2007*. http://turbotax.intuit.com/, (2007).
20. Kang, H., Plaisant, C., Lee, B. and Bederson, B. B. NetLens: Iterative Exploration of Content-Actor Network Data. In *Proc. IEEE Symposium on Visual Analytics Science and Technology*. IEEE Press (2006), 91-98.
21. Lee, B., Parr, C. S., Plaisant, C., Bederson, B. B., Veksler, V. D., Gray, W. D. and Kotfila, C. TreePlus: Interactive Exploration of Networks with Enhanced Tree Layouts. *IEEE Transactions on Visualization and Computer Graphics*, 12, 6 (2006), 1414-1426.
22. McGrenere, J., Baecker, R. M. and Booth, K. S. An evaluation of a multiple interface design solution for bloated software. In *Proc. SIGCHI conference on Human factors in computing systems*. ACM Press (2002), 164-170
23. O'Madadhain, J., Fisher, D., Smyth, P., White, S. and Boey, Y.-B. Analysis and Visualization of Network Data using JUNG. *Journal of Statistical Software*, VV, 2 (2005).
24. Perer, A. and Shneiderman, B. Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12, 5 (2006), 693-700.
25. Perer, A. and Shneiderman, B. Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis. *(Under Submission)*(2007).
26. Plaisant, C. The challenge of information visualization evaluation. In *Proc. Advanced visual interfaces*. ACM Press (2004), 109-116
27. Saraiya, P., North, C. and Duca, K. An Evaluation of Microarray Visualization Tools for Biological Insight. In *Proc. IEEE Symposium on Information Visualization*. IEEE Press (2004), 1-8.
28. Selker, T. COACH: a teaching agent that learns. *Communications of the ACM*, 37, 7 (1994), 92-99.
29. Seo, J. and Shneiderman, B. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data. *Information Visualization*, 4, 2 (2005), 99-113.
30. Shneiderman, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. In *Proc. Visual Languages*(1996), 336-343.
31. Shneiderman, B. and Aris, A. Network Visualization by Semantic Substrates. *IEEE Transactions on Visualization and Computer Graphics*, 12, 5 (2006), 733-740.
32. Shneiderman, B. and Bederson, B. B. Maintaining concentration to achieve task completion. In *Proc. Conference on Designing for User eXperience*. AIGA: American Institute of Graphic Arts (2005).
33. Shneiderman, B. and Plaisant, C. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. In *Proc. Beyond time and errors: novel evaluation methods for Information Visualization, Workshop of the Advanced Visual Interfaces Conference*. ACM Press (2006), 1-7.
34. Spotfire *DecisionSite* http://www.spotfire.com, (2007).
35. Spotfire *Introduction to Spotfire DecisionSite Analysis Builder*. http://spotfire.tibco.com/spotfire_downloads/whitepapers/analysis_builder.pdf, (2007).
36. Stevens, R., McEntire, R., Goble, C. A., Greenwood, M., Zhao, J., Wipat, A. and Li, P. myGrid and the drug discovery process. *Drug Discovery Today: BIOSILOCO*, 2, 4 (2004), 140-148.
37. Tauscher, L. and Greenberg, S. How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47, 1 (1997), 97-137.
38. Viegas, F. B. and Wattenberg, M. Communication-Minded Visualization: A Call to Action. *IBM Systems Journal*, 45, 4 (2006).
39. Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press (1994).