Visually Exploring Large Social Networks

Nathalie Henry

INRIA Futurs/LRI & University of Sydney
Bat. 490 University Paris-Sud, 91405 Orsay, France
J12, University of Sydney, NSW 2006, Australia
nathalie.henry@lri.fr
http://insitu.lri.fr/~nhenry

This PhD is a cotutelle co-advised by:

- Dr. Jean-Daniel Fekete [1], INRIA Futurs, France
- Pr. Peter Eades[2], National ICT Australia & University of Sydney, Australia.

Research Area and Topic

Information Visualization and Human Computer Interaction. This PhD focuses on visualization and interaction to navigate, explore and present large social networks.

1 Introduction

Vast new datasets are available for social scientists to analyze with the increasing use of internet technologies. Email clients, instant messenger and chat; photo sharing and peer-to-peer file exchange; open-source programming platforms and online editable encyclopedias such as wikipedia—all give social scientists ready-to-analyze data about how people communicate and collaborate.

This data avalanche raises new challenges these datasets are *far larger* than those they traditionally analyzed. (For example, the English version alone of Wikipedia contains 1.7 million articles). Also, they frequently contain *richer* information such as the history of each item: who contributed, when, for how much, and they *evolve through time* (the network structure changes as articles are added, transformed or removed).

The stakes of social network analysis are rising : intelligence agencies struggle to discover terrorist networks or epidemiologists to detect and contain outbreaks of diseases such as avian influenza and SARS.

Analysts require effective tools for handling these large, rich and dynamic social networks, to perform reliable yet flexible analysis at many levels, from overviews of the whole to a detailed analysis of important sections. The goal of this PhD is to provide them with visual interactive tools to support both their exploration process and the communication of their findings.

[1] jean-daniel.fekete@inria.fr
[2] peter.eades@nicta.com.au

2 Related work

Social networks are composed of actors (people or groups) linked by relationships (for example kinship, communication or collaboration). As social networks are graphs, their analysis is closely related to the exploration of graphs in general. There are many programs designed to support network analysis. The International Network for Social Network Analysis repository lists more than 50 different programs, and 10 new ones were introduced at last year's Infovis conference (30% of the articles). I classify these systems in two categories.

Menu-based systems and programming packages provide a wide range of functionalities to analyze and visualize social networks. Popular systems such as UCINet[1] and R[2] offer many features for statistical analysis, and common graph software and packages such as Pajek[3] and JUNG[3] also provide a broad range of algorithms to create visual representations of a network. However, mastering all the functionalities of these systems requires a considerable effort for novice users, as they require knowledge of how the algorithms work and how to combine or sequence them. Guess [4] is designed to support a more exploratory process. It provides a simple script language for manipulating the visualizations. However, even it remains inaccessible for many novice users, as it is unclear that social scientists will invest time learning it.

Visual exploration systems have emerged recently. They provide interactions to navigate and manipulate networks, which makes them accessible to novice users. Following Ben Shneiderman's mantra [5]: "Overview first, zoom and filter, then details-on-demand", they provide users with dynamic queries [6] (operations with a direct feedback on the representation).

Systems such as SocialAction[7] start the analysis with a node-link diagram of the full graph. For large and dense graphs, however, node overlap and edge crossing quickly makes these representations unreadable. Users must filter or aggregate nodes to get a readable visualization. SocialAction's strongest feature is its ranking of possible operations the user can perform at each step, providing guidance for the exploration process.

Since providing a readable representation of the whole network is challenging, several systems completely gave up on providing an overview. For example, Vizster[8] and TreePlus[9] concentrate on displaying and navigating in only a small part of a network centered on a specific actor. This "ego-centered" strategy lets users have a readable representation on the screen at all times. Other systems take other radical approaches. PivotGraph [10] starts the exploration from a high-level aggregated network. The user visualizes nodes' categories and their relationships, and then interacts with the visualization to explore lower levels. Finally, NetLens[11] completely gave up the graph representation. It uses simple visualizations such as histograms to explore the graph by its attributes, filtering them back and forth to answer questions.

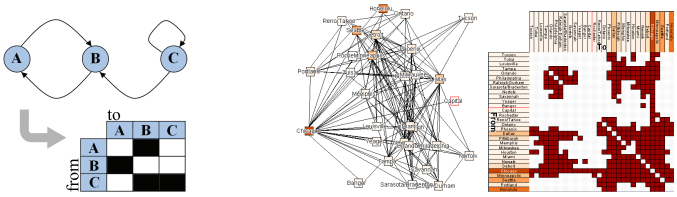
[3] JUNG `http://jung.sourceforge.net`

3 Approach

This PhD follows statistician John Tukey's concept of Exploratory Data Analysis [12]: the primary purpose of visualizing and exploring is to raise questions and gather insights about a large quantity of data. Unlike most statistical work, which evaluates *a priori* questions according to a model, exploration by information visualization has the potential to start analysis without assumptions, or open new perspectives on a previously-analyzed dataset. For these purposes, overviews of the whole network are crucial.

While traditional node-link diagrams are user-friendly, readability suffers for large and dense networks. These factors often make it impossible to use them to visualize the entire network. We have sought alternatives to these representations. I believe adjacency matrix representations (Figure 1a) have a vast potential to investigate large and dense graphs. Ghoniem et al. published a study[13] comparing readability of both representations for several basic tasks of exploration. Results show that matrices outperform node-link diagrams for most of these tasks, especially when the network becomes dense. Figure 1b shows an example of the better readability of matrices for dense networks.

Social networks vary from very sparse (genealogy trees) to very dense (tables of goods exchange) including a locally dense category (small world networks). My approach is to take advantage of both representations, improving them, combining and merging them to handle many different cases.



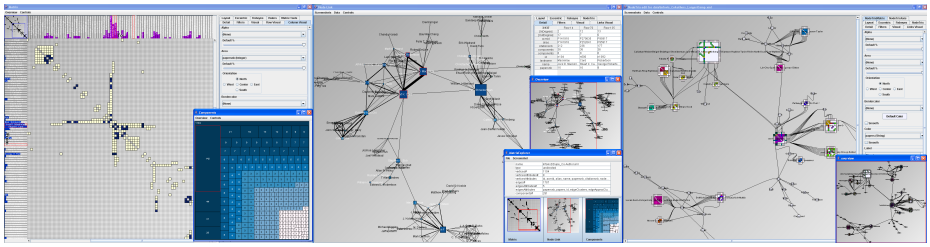
(a) a matrix (b) a dense network

Fig. 1: Matrix and node-link representations

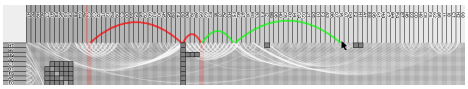
4 Contributions

The major contribution of this PhD is a visual and interactive system to help social scientists analyze large networks. My expected contributions are:

- Use participatory design techniques to determine social science analysts' needs and requirements for an interactive exploration system;
- Assessment of matrix-based representation readability as well as their improvement in general (ordering of their rows and columns) and on specific tasks important for social networks analysis (path-related tasks);
- Create novel visualizations designed for locally dense networks (small-world networks) as well as associated interaction techniques;
- Combine existing and novel visual representations into a system oriented toward interactive exploration.



(a) MatrixExplorer



(b) Zoom on MatLink

(c) Zoom on NodeTrix

Fig. 2: MatrixExplorer, MatLink and NodeTrix

MatrixExplorer: an Exploration System[14, 15]. From a serie of interviews followed by a participatory design workshop, I collected a set of requirements for visually exploring social networks. One of the major outcome was the need for multiple representations of a same network and tools to help analysts find a consensus on their findings. From this study, I designed MatrixExplorer, a system combining matrices and node-link diagrams (Figure 2a). When users apply dynamic queries on one representation, they can observe the results on the other as well. Matrices are generally used to manipulate the network (filtering, ordering, clustering) and node-link diagrams to visualize the resulting one (smaller and sparser) and finally communicate findings. I observed that ordering rows and columns of matrices was crucial to better understand them. Thus, I developed an ordering algorithm based on heurisitics for the traveling salesman problem.

MatLink: Improving Matrix Representations[16]. Ordering a matrix helps identifying communities and central actors, both important tasks for analysis. However, matrices still suffer of a weakness for path-related tasks (how many actors connect A to B?). I designed an interactive solution to solve that major disadvantage of matrices: MatLink (Figure 2b). The principle is to overlay a linear node-link diagram on the matrix headers as well as display interactively the shortest path between selected actors. Currently, I am working on integrating MatLink into ZAME[17], a multiscale matrix explorer. MatLink can provide visual cues on what is not directly visible on the screen, and thus aid navigation.

NodeTrix: a Hybrid Representation[18]. A large category of social networks are globally sparse but locally dense. In this case, the structure is readable with a node-link diagram, but dense sub-parts are not. To solve that problem, I created a hybrid representation : NodeTrix , a node-link diagram visualizing dense sub-parts as matrices (Figure 2c). To smoothly manipulate NodeTrix, I designed a set of interaction techniques based on direct manipulation of the nodes using drag-and-drop. A video is available at `http://insitu.lri.fr/~nhenry/nodetrix/`.

5 Evaluation

Evaluating representations readability can be quantitatively done on a small set of tasks using controlled experiments or on a broader context by running case studies with benchmarks. In this case, results are more qualitative but allows a quick comparison with other systems. My first attempt at assessing matrix readability was a controlled experiment[19]. It partly failed because of the difficulty to operationnalize the exploratory process and to objectively compare subjects' interpretations and findings. To solve that problem, I worked with researchers from HCIL on defining a task taxonomy for graphs [20]. The second experiment I performed was much more focused, using five tasks and a technique I developped to generate representative datasets[4]. It ended with significant results showing that MatLink improved matrices[16]. However, it is hard to generalized these results to the global process of exploration. To validate NodeTrix in a more realistic context, I chose to perform a case study using benchmarks.

Evaluating a visual exploration system is much more complicated as the process to control is long and difficult to operationnalize [21], which exclude controlled experiment. I chose to validate MatrixExplorer *a priori*, by implying users before and during its design. I am currently running a case study, describing how MatrixExplorer is used to explore a large quantity of data and what visualizations are created. Future evaluation would include a longitudinal study. These studies requires effort from both the system creator and the users in term of time and implication. However, they provide rich feedback and materials to analyze the exploration process and improve greatly the tool.

6 Directions for Future Research

At this stage, I can extend my PhD in many directions, I will only present the four I am interested in.

1. *Creating novel visualizations and interaction techniques.* This would especially be usefull to reorganize the high number of controls required to manipulate a network. I can imagine integrating them directly in the visualizations or design smart interaction techniques to replace them.
2. *Guiding the exploration.* Allowing analysts to visualize their previous analysis, annotating it and providing some indicators to guiding the next steps of the exploration would help them in their work and help us to understand the exploration process.
3. *Releasing a stable system.* This is mandatory to run a longitudinal study. I would integrate and instrument all prototypes to log users' actions.
4. *Providing support for collaboration and communication.* This extension would help analysts to work together and to present their findings.

I briefly presented my research work on visual and interactive exploration of social networks. I am looking forward to the Doctoral Consortium to gather feedback on my research work and discuss the future direction of my PhD.

[4] `http://www.infovis-wiki.net/index.php/Social_Network_Generation`

References

[1] Borgatti, S., Everett, M., Freeman, L.: UCINET V user's guide. Analytic Technologies, Natick, MA (1999)

[2] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2006) ISBN 3-900051-07-0.

[3] de Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek. Structural Analysis in the Social Sciences. Cambridge Univ. Press (2005)

[4] Adar, E.: Guess: a language and interface for graph exploration. In: CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, ACM Press (2006) 791–800

[5] Shneiderman, B.: The Eyes Have It: A Task by Data Taxonomy for Information Visualization. Visual Languages (1996) 336–343

[6] Ahlberg, C., Williamson, C., Shneiderman, B.: Dynamic queries for information exploration: An implementation and evaluation. Proceedings of the ACM CHI'92: Human Factors in Computing Systems (1992) 619–626

[7] Perer, A., Shneiderman, B.: Balancing Systematic and Flexible Exploration of Social Networks. IEEE TVCG (Infovis'06 proceedings) **12**(5) (2006) 693–700

[8] Heer, J., Boyd, D.: Vizster: Visualizing Online Social Networks. In: Proceedings of the IEEE Symposium on Information Visualization. (2005) 5

[9] Lee, B., Parr, C.S., Plaisant, C., Bederson, B.B., Veksler, V.D., Gray, W.D., Kotfila, C.: Treeplus: Interactive exploration of networks with enhanced tree layouts. IEEE TVCG (Infovis'06 proceedings) **12**(6) (2006) 1414–1426

[10] Wattenberg, M.: Visual exploration of multivariate graphs. In: Proceedings of the CHI conference, Montréal, Québec, Canada, ACM Press (2006) 811–819

[11] Kang, H., Plaisant, C., Lee, B., Bederson, B.B.: Netlens: Iterative exploration of content-actor network data. Proceeding of IEEE Symposium on Visual Analytics Science and Technology (VAST) (2006) 91–98

[12] Tukey, J.: Exploratory Data Analysis. Addison-Wesley (1977)

[13] Ghoniem, M., Fekete, J.D., Castagliola, P.: On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Information Visualization **4**(2) (2005) 114–135

[14] Henry, N., Fekete, J.D.: Matrixexplorer: Un système pour l'analyse exploratoire de réseaux sociaux. Proceedings of IHM2006, International Conference Proceedings Series (September 2006) 67–74

[15] Henry, N., Fekete, J.D.: MatrixExplorer: a Dual-Representation System to Explore Social Networks. IEEE TVCG (Infovis'06 proceedings) **12**(5) (2006) 677–684

[16] Henry, N., Fekete, J.D.: Matlink: Enhanced matrix visualization for analyzing social networks. Proceedings of Interact (to be published) (2007)

[17] Fekete, J.D., Elmqvist, N., Do, T.N., Goodell, H., Henry, N.: Navigating wikipedia with the zoomable adjacency matrix explorer. INRIA Tech. Report (April 2007)

[18] Henry, N., Fekete, J.D., McGuffin, M.: Nodetrix: Hybrid representation for analyzing social networks. INRIA Tech. Report (April 2007)

[19] Henry, N., Fekete, J.D.: Evaluating visual table data understanding. In: BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV'06), Venice, Italy, ACM Press (2006)

[20] Plaisant, C., Lee, B., Parr, C.S., Fekete, J.D., Henry, N.: Task taxonomy for graph visualization. In: BELIV'06 workshop, Venice, Italy, ACM Press (2006) 82–86

[21] Plaisant, C.: The challenge of information visualization evaluation. In: Proceedings of the AVI Conference, Gallipoli, Italy, ACM Press (2004) 109–116