RC22564 (W0209-072) September 18, 2002
Computer Science

IBM Research Report

Visualizing Egocentric Social Networks from Document Collections

Douglas N. Gordin
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Visualizing Egocentric Social Networks from Document Collections

Douglas N. Gordin
IBM T.J. Watson Research Center
dgordin@us.ibm.com

Abstract

Social relations provide a primary index for document collections. Yet, the analysis and visualization of these relations are not well supported by digital libraries or document collection systems. Data extraction and visualization techniques are presented that powerfully illustrate an individual's social network, thereby aiding document collection navigation and discovery.

1. Introduction

Social relations provide a primary index for document collections. Yet, the analysis and visualization of these relations are not well supported by digital libraries or document collection systems. The utility of social relations for categorizing documents is well documented by social scientists that have identified the crucial role that communities play in the diffusion and creation of ideas. For example, if a specialist is given the title of a document, she might seek to gain further understanding of its perspective by finding out who the authors have studied with or what organization sponsored the work. In general, the formal elaborations of linkages like these are called social networks.

The goal here is to show how a set of three visualizations can powerfully illustrate an individual's social network, based purely on extracting data from a document collection. The individual that will be illustrated in this way is Jock Mackinlay, a pioneer in the field of information visualization. This choice was motivated by his compelling keynote presentation at SmartGraphics 2002 where he gave a two-decade retrospective of his work in information visualization along with his Xerox Parc colleagues. I wondered what I could learn about the underlying social network by analyzing Mackinlay's publications. Using the Raisin system [1], I analyzed the US patents of Mackinlay and his colleagues and produced a suite of three visualizations that I argue provide a compelling index into the work of an influential information visualization creator and his immediate colleagues. Before going through these figures, it is useful to establish some social network terminology.

There are different sorts of social networks with varying utility. Egocentric social networks focus an individual's network. Typically, they have a circular organization with the focal person at the center with their familiars arcing out from them (see Figure 1). In contrast, a relational social network shows the relationship between groups of people (e.g., see Figures 2 and 3). Another distinction is between one-mode and two-mode social networks. One-mode networks are made up solely of people and their relationships (e.g., see Figure 2 and 3). Two-mode networks contain people and some other type such as events or documents and show the relationship between these two types (e.g., see Figures 1). In graph terms, two-mode networks are bipartite since arcs only occur between nodes of different types. I focus here on egocentric social networks, that is, understanding a single author's social network. Accordingly, all the networks are either egocentric or derived from an egocentric network.

2. Extended Example

A suite of three visualizations illustrate Mackinlay's egocentric social network. Figure 1 shows a synoptic visualization of Mackinlay (blue), his patents (green), his co-authors (yellow), and his co-authors patents (green). Visualizations that trace out a person's influence in this way are called "degrees of separation" networks (as in the popular Six Degrees of Separation drama). Each level of indirection is a degree of separation. However, these are usually one-mode graphs (i.e., including only people). In contrast, Figure 1 is a two-mode graph including people and patents. This leads me to call this a one-and-a-half degree of separation graph where a half a degree is used for the inclusion of the second mode (documents or patents in this case).

The layout is based on a multi-dimensional scaling algorithm that positions the nodes so the Euclidean distance is similar to the graph distance (i.e., number of hops between nodes). This layout provides a graphical clustering of the nodes. For example, contrast the group of nodes labeled 'A' with the nodes labeled 'B'. Both sets show documents authored by Mackinlay. However, the former (i.e., 'A') are centrally placed since their co-authors are people with whom Mackinlay collaborated extensively. In contrast, the later (i.e., 'B') are placed more to the periphery since their co-authors were ones he collaborated with much less.

Figure 2 shows Mackinlay's social network, though with Mackinlay himself removed. This graph was produced by modifying Figure 1 as follows: (A) Arcs are added that connect all co-authors of each patent; (B) All patent nodes are removed; (C) the node for Mackinlay is removed. The value of this last step is to simplify the graph (Mackinlay would be connected to every other node), thereby allowing second order relations to emerge. These second order effects are meaningful because it takes into account all of the patents these people have written, not just ones they co-authored with Mackinlay. Again the multi-dimensional scaling layout is helpful to create visual clusters. In particular, the cluster of authors at the center are Mackinlay's primary collaborators and, readily identifiable as core contributors to the field of information visualization.

Figure 3 shows a lattice that was inferred from Figure 1. A lattice is a partial order, that is, it shows a non-strict ordering relation. The ordering relation here is based on co-authorship, where author A is greater than author B, if author A is a co-author on everything that author B wrote, but author A also wrote other documents as well. In the figure the "greatest" authors are at the top and the authors they "dominate" or "subsume" shown below them. Clearly, no definitive pecking order can be inferred, rather visualizations of this sort suggest how intellectual work was partitioned amongst a social group. As such they are useful for finding related papers and tracking the diffusion of ideas.

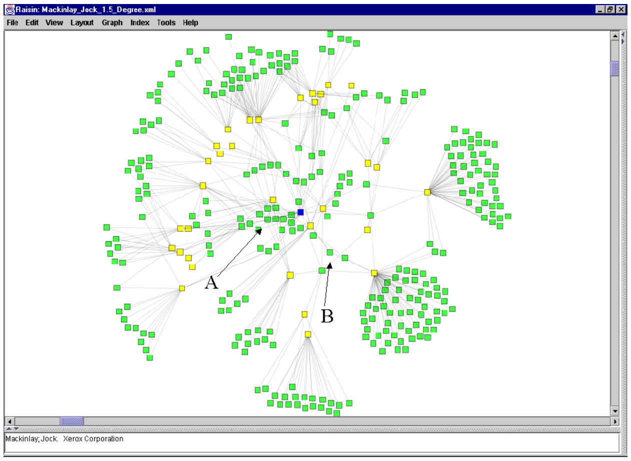
3. Summary and Conclusion

Document collections can improve their navigation and search capabilities by providing the means to visualize and analyze egocentric social networks. This point is shown here by demonstration using the Raisin system on the US Patent document collection to analyze the egocentric network of Jock Mackinlay. Other systems (notably Referral Web [2]) have shown social networks based on co-authorship, but differ in that only one-mode social networks are shown, they do not provide specialized facilities for egocentric networks, and do not provide social network analyses like partial orders. In future work I would like to investigate the means to flexibly focus on single authors or groups of authors, including finding those individuals that serve as primary linkages between contrasting communities of practice.

4. References

[1] Gordin, D.N. & Farrell, R.G. Visualizing the Patent Document Collection as a Graph of Inventors and Their Inventions. IEEE Symposium on Information Visualization 2001, October, San Diego, CA.

[2] H. Kautz, B. Selman, M. Shah, "Referral Web: Combining Social Networks and Collaborative Filtering." *Communications of the ACM, Vol 40,* No 3. March 1997.


Raisin: Mackinlay_Jock_1.5_Degree.xml
File Edit View Layout Graph Index Tools Help
A
B

Mackinlay;Jock. Xerox Corporation

Figure 1: Two-mode social network with 1.5º degrees of separation.

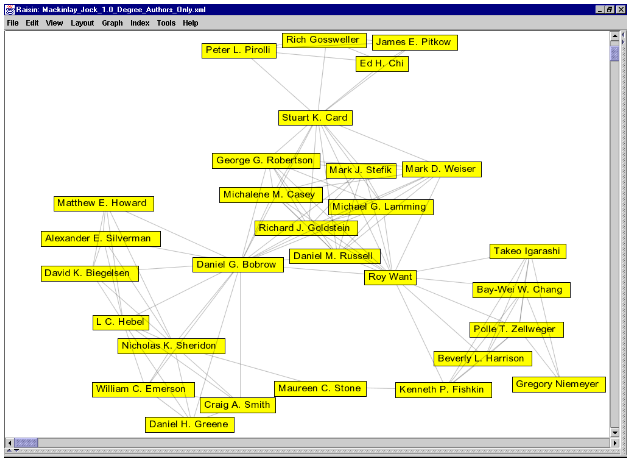


Figure 2: One-mode social network without focal author.

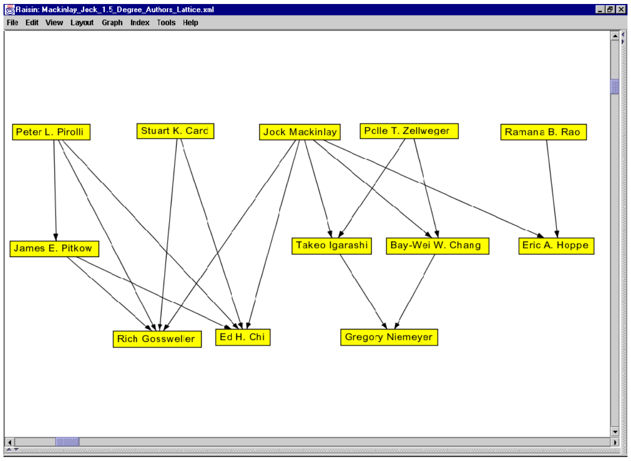


Figure 3: One-mode social network showing partial order between authors of documents.