Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis

Peter A. Gloor, Yan Zhao
MIT Center for Coordination Science & Center for Digital Strategies at Tuck at Dartmouth,
iQuest Analytics
pgloor@mit.edu, yzhao@iquestglobal.com

Abstract

iQuest is a novel software system to improve understanding of organizational phenomena with greater precision, clarity, and granularity than has previously been possible. It permits to gain new insights into organizational behavior, addressing issues such as tracking information while respecting privacy, comparing different interaction channels, network membership, and correlating organizational performance and creativity. It extends automatic visualization of social networks by mining communication archives such as e-mail and blogs through including analysis of the contents of those archives.

Keywords—semantic social network analysis, SNA, vector space information retrieval, temporal social surface, automatic communication analysis

1. Introduction

As we shift from an information-based society to one based on knowledge, it is critical that we have a kind of "grammar" that allows us to understand patterns of communication and, more important, to visualize new patterns. This paper introduces a novel integrated software system allowing organizations to drill down into communication processes by different criteria and to visualize the communication flow and contents, making it easy to understand and explain its communication patterns. The resulting communication "maps" provide a basis for further discussions about how to design organizations and their spaces in an optimal manner.

iQuest's goal is to visualize and analyze communication activity, content and unstructured text, and interactions in a practical and unobtrusive way while still respecting individual privacy.

2. Related Work

Recently visualization and analysis of social networks has become an active topic of research. There are many tools, which take as input actors and the strength of ties between actors to analyze and visualize all sorts of social network analysis metrics [Was92, Borg92, Bat98]. For example, EmailNet [Van04] has been used to gather data on knowledge-worker productivity in executive search firms. TeCFlow, a predecessor of iQuest [Glo04] has been used to analyze behavior of teenage cell phone users, managers of a global services firm, and project team members of two software companies. Vizster [Hee05] has been used to identify communication pockets and gaps in Enron data, by mining the enron e-mail archive and visualizing their social network. Communications Garden [Zhu01] has been used by the Arizona police department to predict links in criminal social networks, based on semantic analysis of communication contents.

3. iQuest Overview

Much can be learned from static and dynamic visualizations of the communication networks that show the evolution over time [Glo03]. A vastly extended version of predecessor tool TeCFlow [Glo04], iQuest takes as input any type of communication archives such as e-mail, phone records, blogs, Weblinks, or chat sessions. An interactive movie shows the evolution of social networks over time, displays active relationships in a sliding time window, and calculates and plots the evolution of group betweenness centrality and density over time to discover interesting events in the lifetime of a virtual team and different phases in the life cycle of virtual communities [Was94].

As its main graph layout algorithm iQuest uses the Fruchterman-Rheingold spring-embedder layout algorithm [Fru91]. Actors are depicted as nodes, ties between actors as connecting arcs. The stronger the

0-7695-2602-0/06 $20.00 © 2006 IEEE

relationship between two nodes, the shorter is their connecting arc.

We use Freeman's index for quantifying the overall level of betweenness in a set of actors BC_{group}, which summarizes the actor normalized betweenness indices:

$$BC_{group} = \frac{\sum_{i=1}^{g}[BC_*' - BC_i']}{(N-1)}$$

where BC_*' is the largest realized normalized actor betweenness index of the set of actors and N is the number of nodes in the network. DC_{group} is the normalized group degree centrality, where DC_* is the largest simple actor degree and N is the number of nodes in the network:

$$DC_{group} = \frac{\sum_{i=1}^{N}[DC_* - DC_i]}{[(N-1)(N-2)]}$$

Group density of a network is defined as the proportion of ties present in relation to all ties possible:

$$D = \frac{l}{N(N-1)/2}$$

l is the number of present edges; N is the number of nodes in the network.

For our dynamic visualization, we are using a sliding time frame algorithm, where we are looking at a time interval consisting of a flexibly chosen number of days.

We also look at the frequency with which individuals send and receive messages. We have defined a measure, which we call the "contribution index":

$$\frac{messages\_sent - messages\_received}{messages\_sent + messages\_received}$$

The contribution index is +1, if somebody only sends messages and does not receive any message. The contribution index is –1, if somebody only receives messages, and never sends any message. The contribution index is 0, if somebody has a totally balanced communication behavior, sending and receiving the same number of messages [Glo03]. We then plot the contribution index against the total number of messages sent and received of each participant. An example of those views is shown in figure 1.

4. Analyzing Social Networks by Contents

In addition to the social network view, in iQuest the same algorithms are also applied to textual content analysis. Free text documents are parsed, and a keyword index is computed, based on term frequency inverse document frequency tfidf, where the term frequency tf in a document is defined as $\mathrm{tf} = \frac{n_i}{\sum_k n_k}$ with n_i being the frequency of term i in a document, divided by the number of all terms in the document. This is calibrated with the frequency of term i in the document collection,

where N is the number of all documents, and D_i is the number of documents where term i occurs. tfidf is then calculated for each term in each document according to the classical formula [Sal89]:

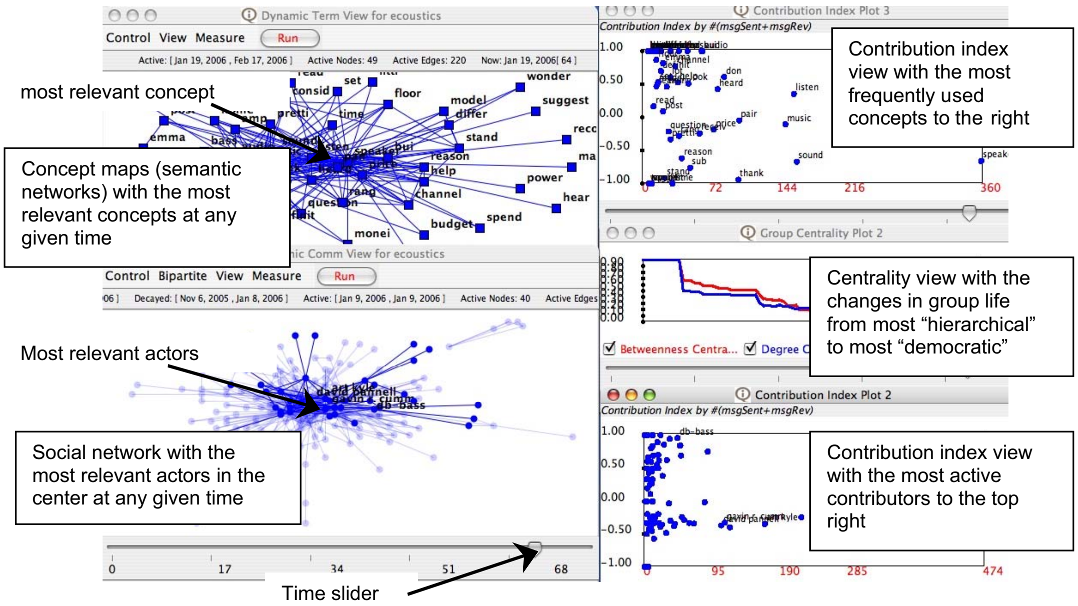
tfidf = tf * log (N/ D_i).

This means that the more documents two terms share, the stronger is their link. The attraction between two terms is based on the summation of all weights of the common terms.

The main advantage of iQuest is its capability to visualize concept maps and social networks over time, based on a sliding time frame algorithm [Glo04], where the new documents or e-mail messages added on the next day define what new actors and terms will appear in the concept and social network view.

The right side of figure 1 illustrates those temporal elements. The contribution index view positions people based on the number of messages they exchanged, and terms based on the number of new documents in which they appear. The centrality view plots a continuous curve of the changes in group betweenness and degree centrality as well as graph density of both social network view and concept view.

iQuest fully respects privacy of users. They always have full control about how much information they want to include into their analysis, e.g. excluding contents, e-mail headings, or anonymizing e-mail addresses. If desired, they can analyze entire communities without including any individual actor information at all.


Dynamic Term View for ecoustics
Control View Measure Run
Active: [Jan 19, 2006 , Feb 17, 2006] Active Nodes: 49 Active Edges: 220 Now: Jan 19, 2006[64]
Contribution Index Plot 3
Contribution Index by #(msgSent+msgRev)
Group Centrality Plot 2
Betweenness Centra... Degree C
Contribution Index Plot 2
Contribution Index by #(msgSent+msgRev)
Control Bipartite View Measure Run
Decayed: [Nov 6, 2005 , Jan 8, 2006] Active: [Jan 9, 2006 , Jan 9, 2006] Active Nodes: 40 Active Edges
0 17 34 51 68
0 72 144 216 360
0 95 190 285 474
1.00 0.50 0.00 -0.50 -1.00

most relevant concept

Concept maps (semantic networks) with the most relevant concepts at any given time

Most relevant actors

Social network with the most relevant actors in the center at any given time

Time slider

Contribution index view with the most frequently used concepts to the right

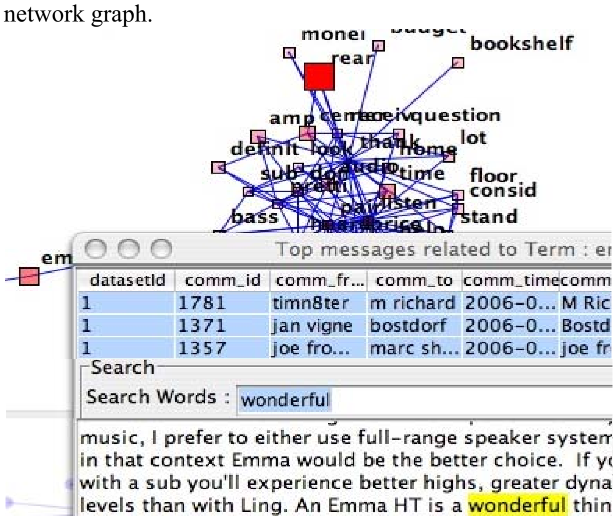
Centrality view with the changes in group life from most "hierarchical" to most "democratic"

Contribution index view with the most active contributors to the top right

Figure 1 Different Views of iQuest

4.1 Searching Concept Networks by Similarity

The concept view also permits to do *similarity searches*. For example, doing a similarity search in a bulletin board on loud speakers for term "wonderful" finds the terms having the strongest association with search string "wonderful" by traversing the concept network graph.



datasetId	comm_id	comm_fr...	comm_to	comm_time	comm
1	1781	timn8ter	m richard	2006-0...	M Ric
1	1371	jan vigne	bostdorf	2006-0...	Bostd
1	1357	joe fro...	marc sh...	2006-0...	joe fr

Search

Search Words : wonderful

music, I prefer to either use full-range speaker system
in that context Emma would be the better choice. If yo
with a sub you'll experience better highs, greater dyna
levels than with Ling. An Emma HT is a wonderful thin

Figure 2. Similarity search for "wonderful'

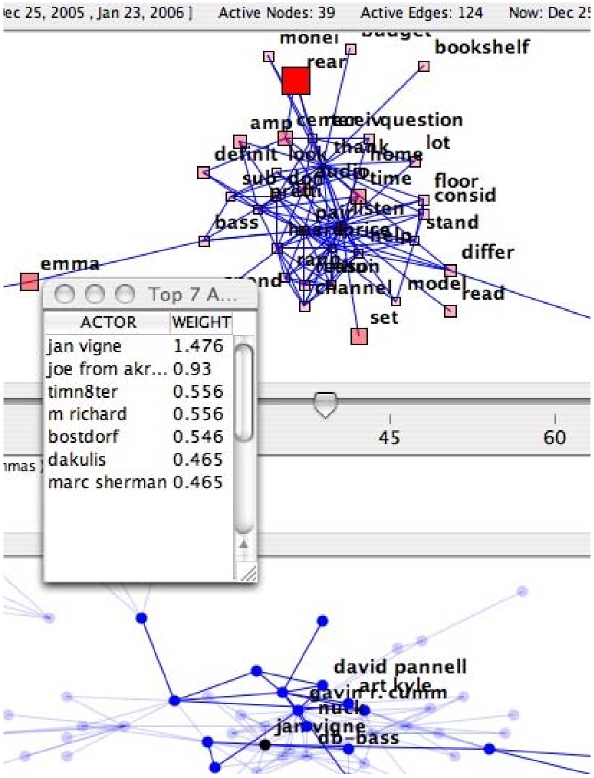
In figure 2, a similarity search for term "wonderful" shows the terms "rear" and "emma" to be the most significant in the concept map. This means that the terms "rear" and "emma" are the most significant in the messages also containing term "wonderful". (The messages are talking about how "wonderful "rear" surround speakers and the "emma" type speakers are).

4.2 Identifying Topics and Main Actors in Communications

iQuest permits to discover which actor is talking about what topic at what point in time. Figure 3 illustrates one way of combining content and social networks. The bottom of figure 3 shows the main actors talking about "emma" such as Jan Vigne, as well as Jan Vigne's central network position in the online speaker discussion forum.

The top of figure 3 shows the concept map automatically generated from the text content of the bulletin board. After thematic concept clusters have been created, the content of the users' mail can be categorized and users can become members of thematic groups according to the nature of their messages. For example, right-clicking on term "emma" shows the most important actors discussing this concept (box in the center of figure 3). In this way, it is possible to determine which users have similar interests and, therefore, help them to better collaborate and communicate. iQuest can be used to identify the most important thematic fields. This allows for collecting information about the issues that matter most to the people who exchange messages.

Combining temporal textual analysis with temporal social network analysis opens up wide opportunities for new applications. Using temporal analysis of e-mail content makes it possible to keep track of changes of the thematic context of a user's mailbox or to monitor changes in the thematic context of a mailing list or conversational thread. This functionality can, e.g., be used to analyze existing mail archives from different company departments, open source mailing lists, etc., to draw conclusions regarding specific patterns of temporal movement of context from one organizational unit to the next.



ec 25, 2005 , Jan 23, 2006]   Active Nodes: 39   Active Edges: 124   Now: Dec 2!

Top 7 A...

ACTOR	WEIGHT
jan vigne	1.476
joe from akr...	0.93
timn8ter	0.556
m richard	0.556
bostdorf	0.546
dakulis	0.465
marc sherman	0.465

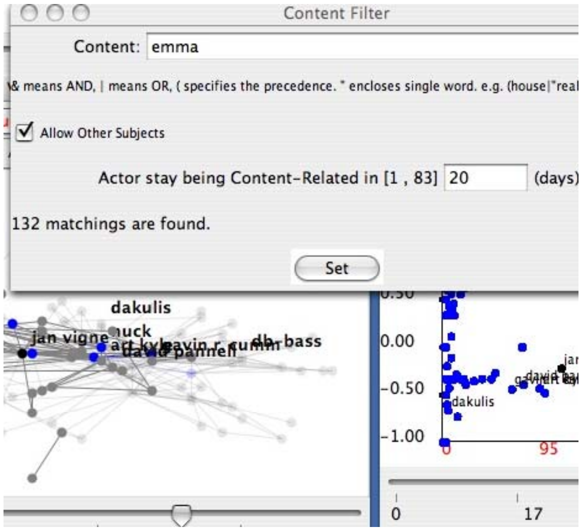
Figure 3. Combining Actor and Content view

Comparing actor and concept information permits to locate the sources of valuable information to identify which groups create what knowledge. This way, readers get additional cues on which information is important to read and which can be ignored.

A user's mailbox can also be compared to the context of a general category topic. This functionality can be used to watch personal communication that is irrelevant to a given topic. This permits to discover new discussion threads evolving out of originally unrelated contents, locating the origins of new ideas. Topic categories can be constructed after parsing collections of documents. Users can also look out for different conversation threads on the same topic. This can be useful, for example, for members of an organization that may have the same interests, but are unaware of each other's activities.

4.3 Tracking Information Flow in Social Networks

Our system also easily permits to maintain a visual history of the information flow among actors within a discussion. Maintaining a history of the information flow makes it possible to visually track the exchange of messages between the users over time.



Content Filter

Content: emma

& means AND, | means OR, (specifies the precedence. " encloses single word. e.g. (house|"real

☑ Allow Other Subjects

Actor stay being Content-Related in [1 , 83] 20 (days)

132 matchings are found.

Figure 4. Visualizing Information Flow in online forum

Figure 4 identifies the most central players in an online forum of hundreds of participants, discussing the respective advantages of "emma." Whenever participants talk about "emma", they are colored in blue. After a specified time interval, they are grayed out again. In figure 4, overall most active participants "jan vigne" and "nuck" are shown conversing about "emma", embedded in all the actors within the speaker forum.

4.4 Combining Communication Links and Content Similarity

We have also developed a combined view, where the calculation of ties between actors is a weighted average of the number of messages exchanged and the number of common terms in the messages sent and received between two actors. In addition to the links computed based on number of messages between two actors, another link is calculated based on the number of common terms collected from all the messages that actor A and actor B send or receive from any other actor. To add additional flexibility, a link weight is defined ranging from 0 to 1, and a term weight inversely ranging from 1 to 0. The view can then be dynamically shifted between placing emphasis on common communication links or common terms.

Figure 5 shows the same network as a link network and as a content network. The top of figure 5 shows the social network of an online community, as calculated based on e-mails exchanged between the members of the mailing list. The bottom of figure 5 shows the same view, with the term slider dragged to the right: additional emphasis is placed on calculating ties based on similar terms in messages exchanged between actors.

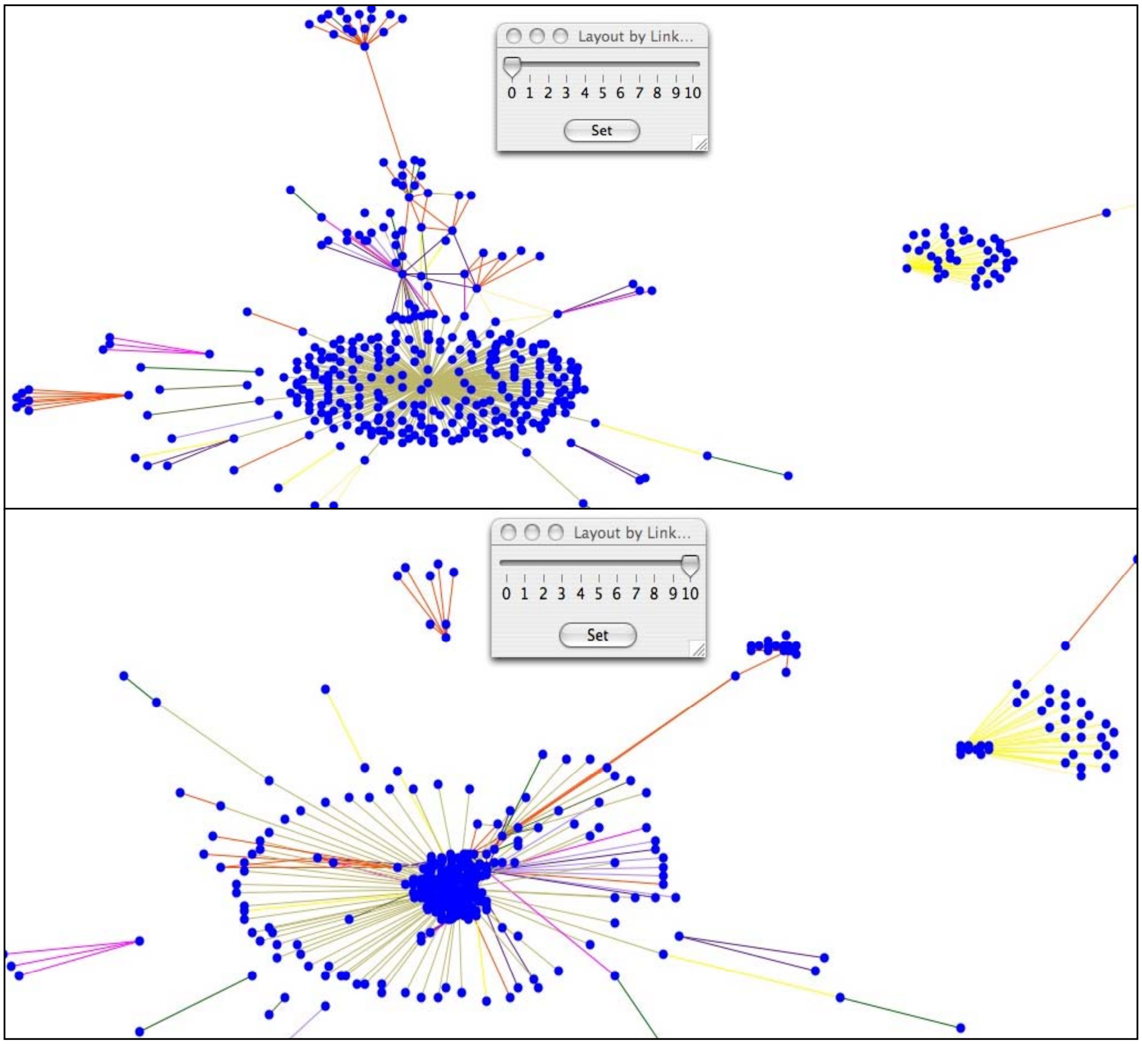


Figure 5. Link-term view with term weight=0 (top) and term weight=1 (bottom)

5. Sample Application – Monitoring Positions of Web Sites in Cyberspace

It is straightforward to extend iQuest to visualize the daily changing positions of Web sites in cyberspace. The same algorithm that animates the exchange of e-mails based on the triple (from_address, to_address, timestamp) also works for Weblinks, as they have the same temporal structure. Based on the simple insight "you are not to whom you link", but "who links to you" iQuest collects all the links up to n degrees of separation pointing to a source URL. This application is useful for positioning brand-aware Web sites and for positioning Blogs, as it indicates their relative standing in cyberspace at a glance.

It even permits discovering Web trends by looking at changes in link patterns between blogs, such as e.g. tracking the centrality of a certain blog within the entire blogosphere.

Figure 6 shows the top 10 Web sites returned to the query "avian flu drug", as well as the top 10 Web sites pointing to each of the top 10 Web sites returned by the query, and the top 10 Web sites pointing to the previous Web sites, i.e. at most 10*10*10=1000 Web sites. Note that in fact we only have 227 Web sites, because Web sites come up repeatedly at each subsequent level of separation from the original 10 Web sites. Those multiply connected Web sites of high betweeness centrality, such as e.g www.nature.com are the ones most deserving further investigation.

Figure 6 shows the link structure of those Web sites also including analysis of their content, i.e. similarity by common terms is factored in. A cluster of closely related Web sites is emerging near the black rectangle in the middle of the picture representing the original query. Www.nature.com becomes one of the most central Web sites, further pointing out the relevance of its contents for the query "avian flu drug".

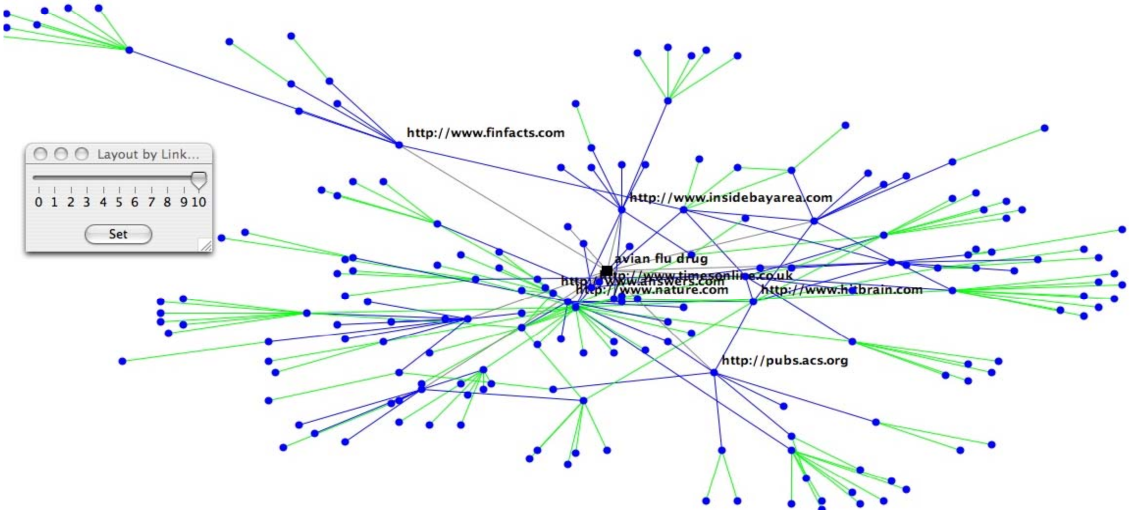


Figure 6. Visualizing the relationships between the top Google search results to query "avian flu drug"

6. Conclusions

iQuest is currently used to further explore open research questions analyzing hidden relationships and their context. It assists in comparing different technologies of interaction: Do social networks depend on the interaction technology? In other words, does the same group of people exhibit different network attributes when interacting via telephone, email, face-to-face or other? iQuest can also analyze network membership: How do social networks change over time? Is it possible to detect new patterns and correlate them to external events? How do collections of ego networks correlate to group networks? Another research issue where our tool is helpful is to discover correlations between communication structures and performance and productivity: Can we correlate network attributes either statistically or visually with performance criteria, for example, creativity, cohesion, information overload, or productivity? Papers describing first results of these projects can be found on our project Web site http://www.ickn.org/html/ckn_publications.htm.

Our tool permits organizations to develop working, communication, process improvement, and trust-building practices. iQuest also offers novel individual knowledge worker productivity enhancement functionality for better individual communication behavior. The use of iQuest provides organizations with information about their current communication structures and possible problem areas and support needs. By recognizing their previously hidden internal innovation networks, and discovering trends and trendsetters organizations can mobilize links and synergies of which they were not aware.

Acknowledgements

The authors thank Tom Allen, Hans Brechbuhl, Marius Cramer, Scott Dynes, Kai Fischbach, Francesca Grippa, Myank Jain, Eric Johnson, Christine Kohlert, Casper Lassenius, Rob Laubacher, Tom Malone. Maria Paasivaara, John Quimby, Ornit Raz, Josh Rosenthal, Detlef Schoder, Thorsten Seehawer, and Antonio Zilli for helpful discussions, suggestions, and invaluable feedback as early testers of our software.

References

[1] Batagelj , V. and A. Mrvar. Pajek—Program for Large Network Analysis. *Connections* 21 (2): 47-57. 1998

[2] Borgatti, S., M. Everett and L. C. Freeman, L.C. 1992. UCINET IV, Version 1.0, Columbia: Analytic Technologies.

[3] Fruchterman, T.M.J Reingold, E M. Graph drawing by force directed placement. Software: Practice and Experience, 21(11), 1991.

[4] Gloor, P. Laubacher, R. Dynes, S. Zhao, Y. Visualization of Communication Patterns in Collaborative Innovation Networks - Analysis of some W3C working groups, Proc. ACM CKIM, New Orleans, Nov 3-8, 2003.

[5] Gloor, P. Zhao, Y. TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis, ACM CSCW Workshop on Social Networks. ACM CSCW Conference, Chicago, Nov. 6. 2004.

[6] Heer J. Boyd, D. Vizster: Visualizing Online Social Networks. InfoVis 2005, IEEE Symposium on Information Visualization. Available at http://jheer.org/vizster.

[7] Salton, G. Automatic Text Processing. Addison-Wesley, Reading, MA, 1989.

[8] Tyler, J. Wilkinson, D. Huberman, B. A. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. HP Laboratories, 2003.

[9] Van Alstyne, M. and Zhang, J, EmailNet: A Tool for Capturing Anonymous Email, University of Michigan working paper., 2004.

[10] Wasserman , S., Faust, K. 1994. *Social Network Analysis : Methods and Applications*. Cambridge University Press.

[11] Zhu, B. and Chen, H., Social Visualization for Computer-mediated Communication: A Knowledge Management Perspective, Proc. WITS'01, 23-28, 2001.