Identifying Potential Suspects by Temporal Link Analysis

Peter A. Gloor
MIT CCS and iQuest Analytics
pgloor@mit.edu

Sebastian Niepel, Ye L
University of Cologne

Abstract

This paper describes the application of temporal link and content analysis to the well-known Enron e-mail dataset. TeCFlow permits to extract dynamic movies of the evolution of social networks, identifying gatekeepers and other central actors, as well as to generate temporally correlated cluster maps of e-mail content. Our approach combining social network analysis with information mining permits intelligence analysts to easily identify patterns of potentially suspicious actors and activities in large e-mail and other communication archives.

Keywords

network dynamics , Enron, e-mail analysis , temporal visualization , content map.

Introduction

We have extended TeCFlow [Glo04], a tool for the temporal visualization and analysis of social networks, to visualize content in its temporal context. TeCFlow takes as input outlook mailboxes, Eudora mailboxes, Web mailing lists and online forums, Web links, and flat files. It parses those documents and incrementally stores them in a MySQL database. It then offers the option to visualize and analyze this data in manifold and visual ways. TeCFlow creates interactive dynamic movies of relationships. It combines visualization and analysis of the evolution of social networks over time with visualization and analysis of the evolution of semantic networks (concept maps [Glo91]). It shows synchronized changes in central positions of social actors and core concepts over time.

The TeCFlow term view analyzes temporal text to identify the leading themes or concepts based on the vector space information retrieval model [Li898]. It identifies themes by clustering previously unconnected statements and documents. Active relationships between concepts are displayed in a sliding time window, with inactive relationships decaying over time. TeCFlow also calculates and plots the evolution of group betweenness centrality, density, and contribution index of actors [Was93], be it people or concepts over time to discover

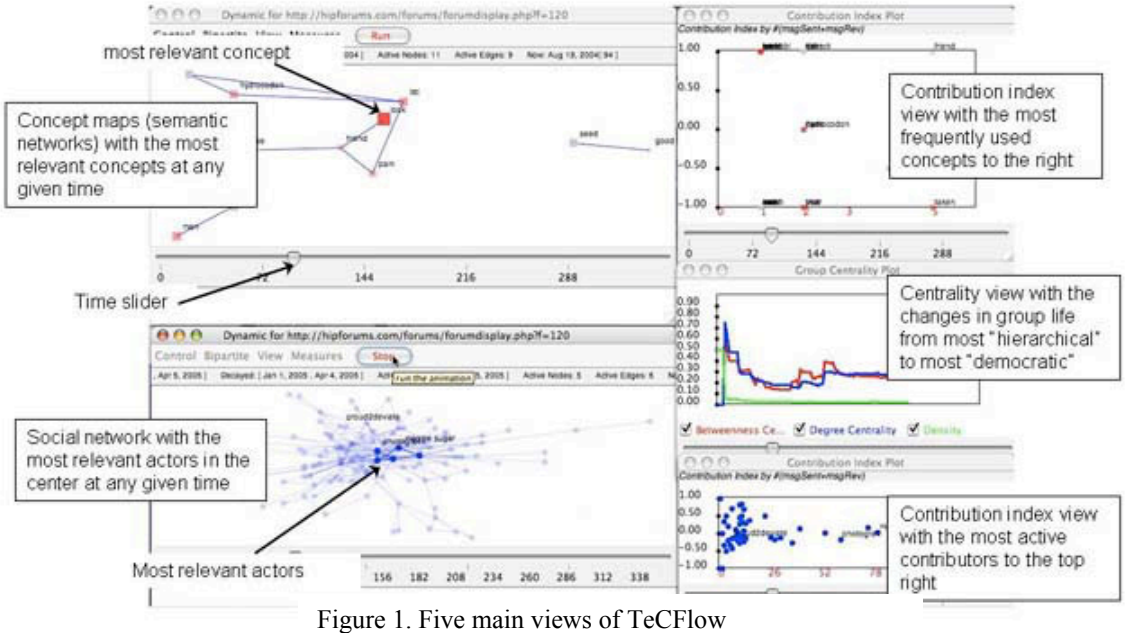

Figure 1. Five main views of TeCFlow

changes in interesting concepts in the lifetime of an actor. TeCFlow takes as input structured and unstructured data such as the Web, Google search results, email logs, phone archives, Intranets, and plain documents. TeCFlow can track the evolution of relationships between people and ideas in real time as events unfold, or in retrospect as history. Through social network analysis, TeCFlow can identify key gatekeepers, influencers, innovators, leaders and communicators in a network of individuals and, based on communications patterns, predict who is likely to talk to whom about a specific theme or concept [Glo06].It allows, for example, combining tracking the changes in social structures in an e-mail network with visualizing the central concepts discussed in the e-mails. Another application is combining the link structure of Web documents, where changes in the network of Web links are visualized over time, with tracking changes in contents of the Web documents. TeCFlow offers unique capabilities to display and identify unfolding relationships, be they people, words or concepts.

In the remainder of this paper we illustrate the use of TeCFlow for the visualization of analysis of information flows in organizations. Based on the Enron e-mail dataset we show how TeCFlow helps discover collusion and fraud by revealing hidden links and relationships within Enron employees.

Discovering Suspicious Activity in the Enron e-Mail Dataset

Enron was one of the world's leading energy, commodities and services companies. The company marketed electricity and natural gas, delivered energy and other physical commodities, and provided financial and risk management services to customers around the world.

Among other misdeeds, Enron employees in 2001 artificially introduced an energy shortage and subsequently overcharged Californian energy users.

Dave Delainey	Chairman and CEO, Enron Energy
Greg Whalley	former President and COO
Jeffrey K.Skilling	former CEO
John Lavorato	President and CEO, Enron Americas
Kenneth Lay	former Chairman, President and CEO
Louise Kitchen	President Enron Online
Mark Haedicke	Managing Dir. and General Counsel, Enron Wholesale
Richard Causey	Exec.VP,Chief Accounting Officer
Richard Shapiro	Vice President Regulatory Affairs
Sally Beck	Chief Operating Officer
Steven J. Kean	VP and Chief of Staff

Table 1 Some Actors of Enron (source Wikipedia)[1]

We imported the Enron email dataset into a TeCFlow database. It consists of 517,431 messages that belong to 150 mailboxes. We used three methods to identify potential suspects: (1) filtering out messages with potentially suspicious contents, and then focusing on the social network created by those messages, (2) doing a large-scale social network analysis, judging actors by their closeness to suspicious people, and (3) searching for clusters of suspicious activity, by looking for what we term “collaborative innovation networks” (COINs).

Filtering by Keywords

One possibility to find potential suspects is to filter the communication data in a way that all “good” actors are filtered out and only the “evil” ones remain. The goal is to raise the participation level of “evil” actors and lower the level of participation of the “good” ones in the filtered communication network.

A simple way to filter emails is to filter them by special words. Our hypothesis was that emails used to plan or coordinate criminal actions contain some special keywords. Actors who are talking about illegal actions do not call them by name. They disguise terms like “bombs” using other words such as “package”. If a criminal context

[1] That a name appears in this list does not mean that this person has been convicted. It only means that persons in this list had a central role at Enron in the critical period when the Californian Energy crisis was happening.

is known, it is possible to search and filter for emails containing words belonging to that particular context. In addition the relationships between these words can be analyzed to further focus on the context.

An advantage for an analysis of the Enron mailbox is that the context in which the criminal actions were taken is well known. We therefore used a combination of the following words for filtering:

- affair (criminals don't use clear words)
- FERC (Federal Energy Regulatory Commission)
- devastating (what is coming up, and they know)
- investigation (dangerous thing)
- disclosure (dangerous thing)
- bonus (most important thing)

Figure 2 shows the resulting static view of all communications in 2001, only including messages containing a combination of the above terms.

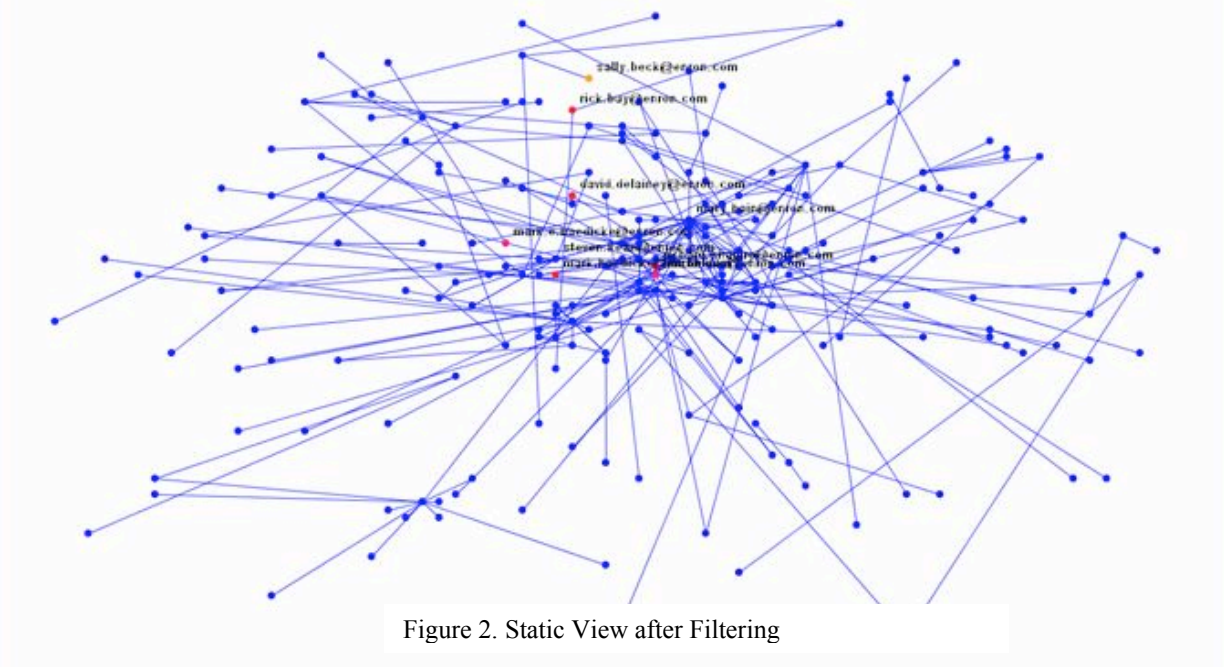

Figure 2. Static View after Filtering

The static view in figure 2 shows that the actors, mentioned in table 1 become more central. In the filtered view these potentially suspicious actors are close together, but there are still too many other actors.

In the next step we therefore created a concept map of terms used in the filtered dataset and looked at how they are used together in email messages (figure 3).

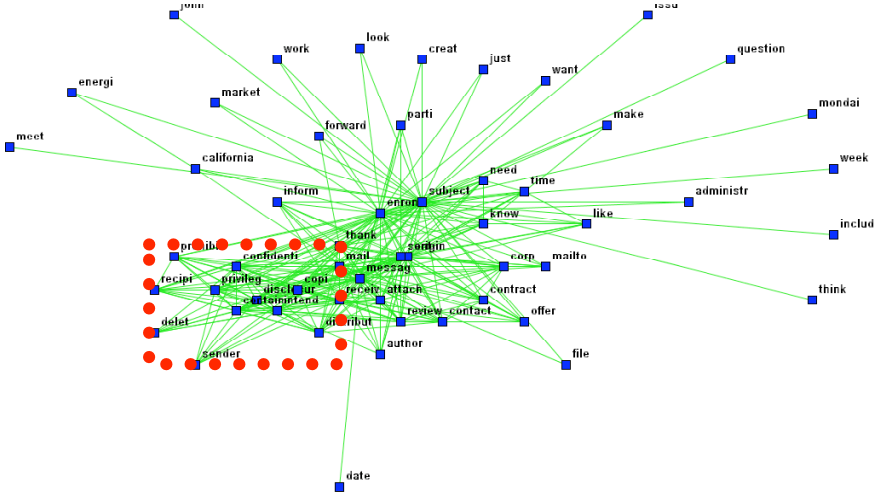

Figure 3. TeCFlow Term View (All Terms)

The section marked in red in figure 3 contains words that could potentially be interesting in a forensic context. Figure 4 contains an enlargement of the red section in figure 3.

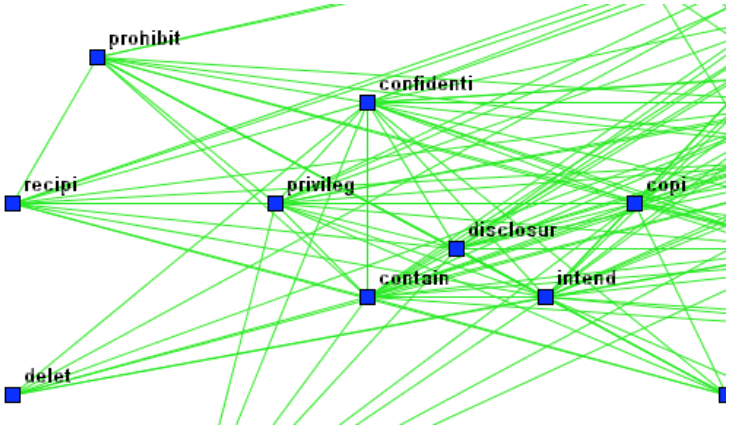

Figure 4. TeCFlow Term View of interesting terms

The short distance between words like “confidential”, “prohibition” and “delete” allows us to now zoom into the messages containing those terms and again look at the social network of the actors using those terms.

Birds of a Feather Flock Together

Another way to discover suspicious people and activities is to start with known suspects, and then look at their interactions with common friends. In our analysis, a relationship between two actors exists if there is direct communication between them. The more they communicate, the more intensive is their relationships.

Figure 5 illustrates the “common friends” and gatekeepers between Kenneth Lay and Tim Belden. Both actors played a central role at Enron during the 2001 Californian Energy crisis.

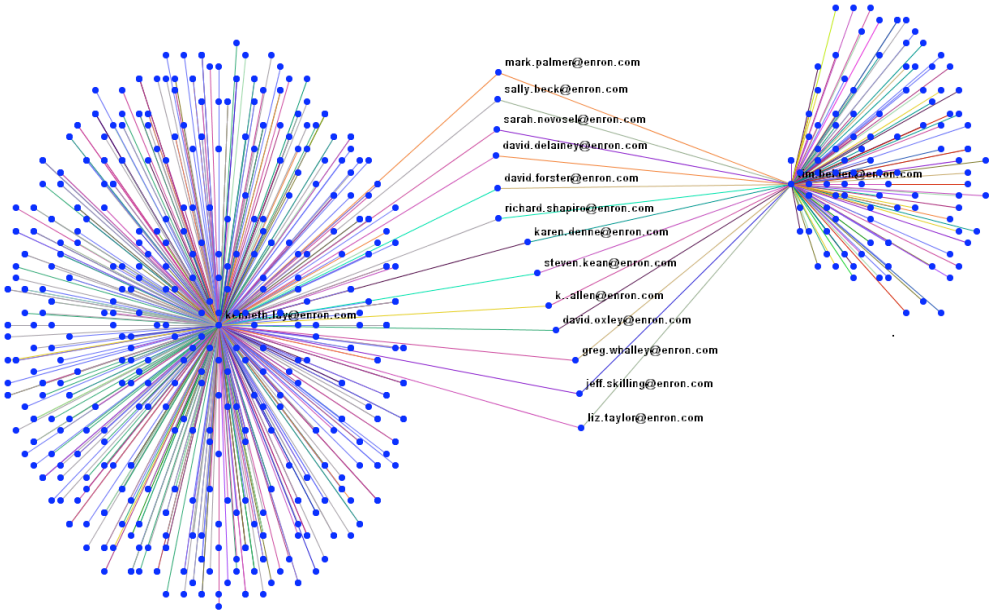

Figure 5. “Gatekeepers” between Lay and Belden

There is no direct connection between the two actors, but 13 common communication partners: David Forster, David Oxley, Jeff Skilling, Karen Denne, Liz Taylor, Mark Palmer, Philipp Allen, Richard Shapiro, Sally Beck, Sarah Novosel, and Steven Kean, out of which 6, namely David Delainey, Greg Whalley, Jeff Skilling, Richard Shapiro, Sally Beck, and Steven Kean appear in our Enron main actor list (table 1).

This method is very simple to use and can easily be applied to large datasets. The starting point for this method are the mailboxes of suspicious persons, by combining them we can then extract the “common friends” and gatekeepers. In a next step we could then look at the contents of the discussion of the suspicious people.

Searching for Innovation Structures

In earlier work we have introduced the concept of “Collaborative Innovation Networks” or COINs [Glo06]. In our framework three types of communities work together to form an ecosystem of interconnected communities, consisting of:

- COINs (Collaborative Innovation Networks)
- CLNs (Collaborative Learning Networks)
- CINs (Collaborative Interest Networks)

COINs (Collaborative Innovation Networks) develop around a small core group of people over time. Around the core team, there are people linked to only one or two of the core team members. A COIN has high density and relatively low group betweeness centrality.

Figure 6. A typical visualization of a COIN

In a CLN (Collaborative Learning Network), a small group of subject matter experts talking among themselves is developing around the coordinator in the center of the graph, who builds a learning network. The communication activities are arranged around them, communicating with a large group of other community members, who are not communicating among themselves.

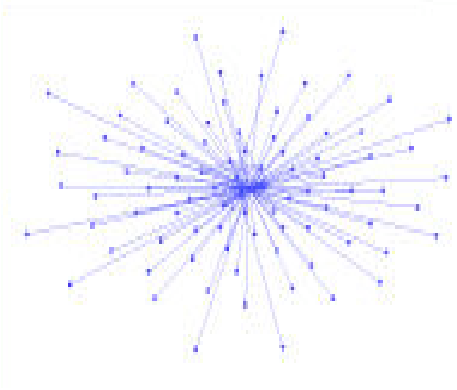

Figure 7. A typical visualisation of a CLN

In a CIN (Collaborative Interest Network) there are different small teams, operating as isolated islands. Over time, the structural holes are filling up, until the network is almost fully connected. There is no clear center in this graph, different people are acting as local hubs.

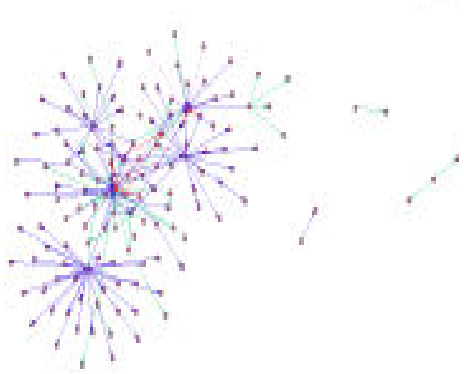

Figure 8. A typical visualisation of a CIN

If the email data is as large as the Enron dataset, it can be quite difficult to get a clear overview of the structure of the communication using conventional visualization of social networks. This problem can be mitigated if we use a temporal visualization, only focusing on a few days at time. We can further narrow down our research to the most interesting persons. Figure 9 illustrates this approach. We can identify three communities with Jeffrey Skilling positioned in the center of one community as the leader of this group.

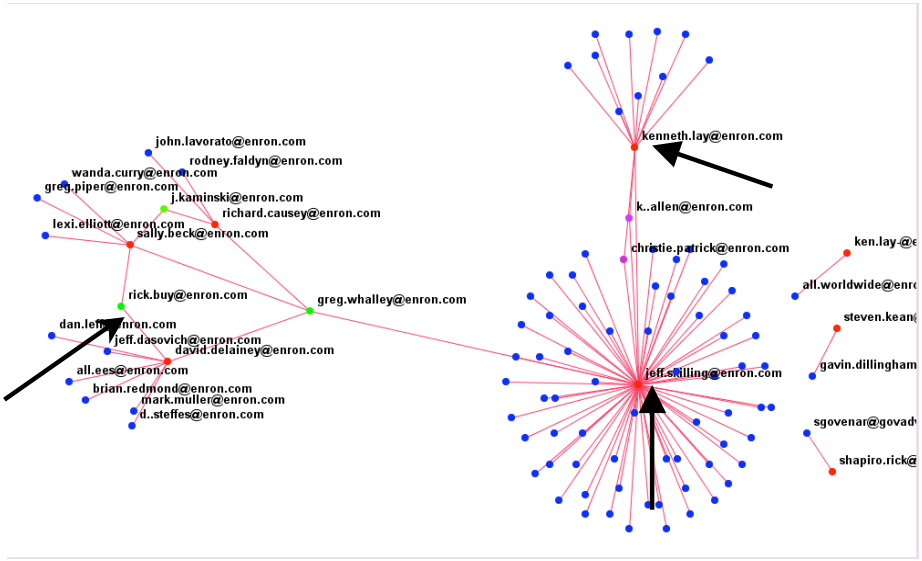

Figure 9. Static view of communities and its leaderships

As innovation can be done for good or for bad, discovering potential COINs is most interesting for gathering intelligence. In figure 10 we have found two COINs. Almost all the suspicious people are appearing in the bigger of the two COINs.

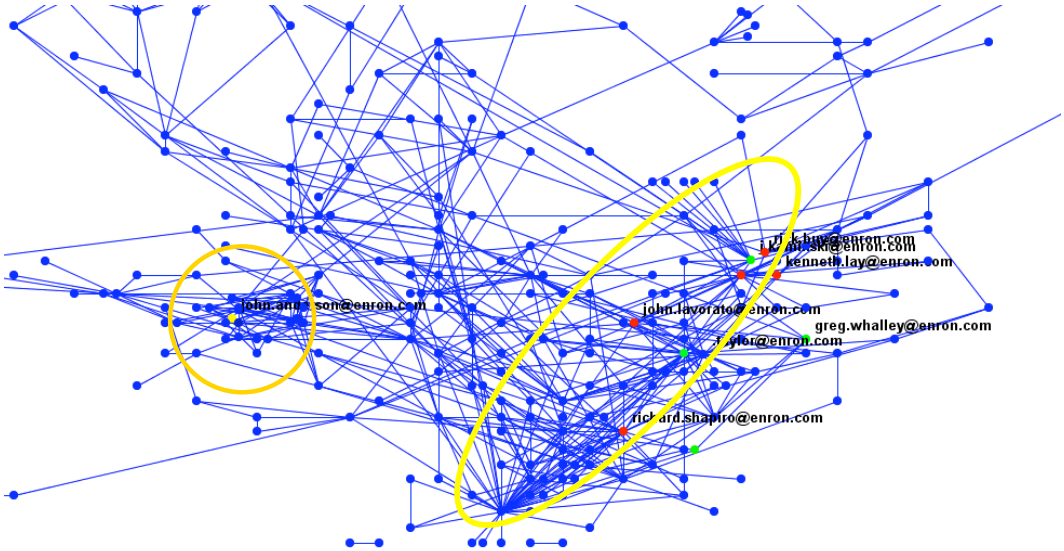

Figure 10. Two COINs discovered in dynamic view

Actors can have certain roles within their respective communities, e.g. as a gatekeeper, who connects at least two communities, or a leader, or a knowledge expert. Besides identifying those roles visually in the social network graph, they can also be found by calculating their contribution index [Glo03]. Roles of different actors can be obtained by measuring differences in their contribution frequency (measured in the numbers of messages sent), and the extent to which their communication is balanced between sending and receiving messages, which we measured via a simple contribution index:

$$\frac{\text{messages sent} - \text{messages received}}{\text{total of messages sent and received}}$$

This index is –1 for somebody who only receives messages, 0 for somebody who sends and receives the same number of messages, and +1 for somebody who only sends messages.

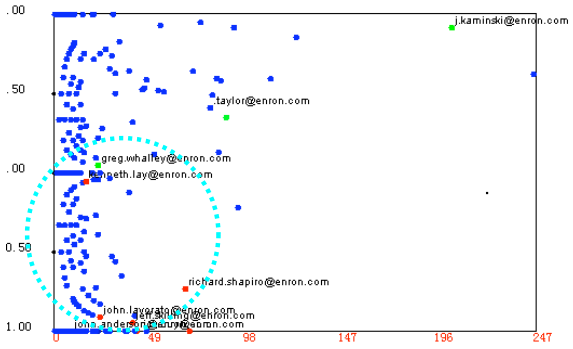

Figure 11. Differences in Contribution Index

Figure 11 illustrates the unfiltered contribution index. We find that most suspects involved in the COIN structure above are now located in the blue circle. It is straightforward to recognize communities and their leaders.

Conclusions

This brief walk-through has only given a glimpse of how temporal analysis of social networks and communication content can be used for forensic investigations. In the past, we have used similar temporal analysis to correlate creativity and performance in teams of open source developers with temporal communication structure [Kid05]. We anticipate applying a similar approach to correlate suspicious activity with temporal communication structure. This way, compliance analysts and members of the legal community can track emerging trends in email conversations – who is writing to whom about what, when and where they are writing from. Government, Intelligence and Law Enforcement professionals will be able to analyze, in near-real time "conversations" and links between email traffic, blogs, message boards and other network-based communication, discovering hidden links and emerging trends.

Acknowledgements

We are indebted to our fellow Enron analysis team member Kirsi Ziegler from the joint Cologne/Helsinki fall 2005 seminar „Collaborative Innovation Networks".

References

[Glo91] Gloor, P. CYBERMAP — yet another way of navigating in hyperspace. In Proceedings of ACM Hypertext '91, San Antonio, Texas, Dec 15-18, 1991.

[Glo03] Gloor, P. Laubacher, R. Dynes, S. Zhao, Y., 2003, "Visualization of Communication Patterns in Collaborative Innovation Networks: Analysis of some W3C working groups". *Proc. ACM CKIM International Conference on Information and Knowledge Management, New Orleans, Nov 3-8, 2003*

[Glo04]] 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.

[Glo06] Gloor, P. *Swarm Creativity, Competitive Advantage Through Collaborative Innovation Networks*. Oxford University Press, 2006

[Hee04] Herr, J. Exploring Enron: Visualizing ANLP Results in Applied Natural Language Processing. InfoSys 290-2. University of California, Berkeley, 2004. http://jheer.org/enron/

[Kid05] Kidane, Y. Gloor, P. Correlating Temporal Communication Patterns of the Eclipse Open Source Community with Performance and Creativity, NAACSOS Conference, June 26 - 28, Notre Dame IN, North American Association for Computational Social and Organizational Science, 2005.

[Li98] Li, Y.H. Jain, A.K. Classification of text documents, The Computer Journal, Volume 48, No 8, pp. 537-546. 1998

[Was94] Wasserman, S., Faust, K, 1994, "Social Network Analysis, Methods and Applications", *Cambridge University Press. 1994.*