EMAILTIME: VISUALIZATION AND ANALYSIS OF EMAIL DATASET

by

Minoo Erfani Joorabchi
Bachelor of Engineering, Tehran University 2003-2007

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
School of Interactive Art and Technology

SIMON FRASER UNIVERSITY

Fall 2010

APPROVAL

Name: **Minoo Erfani Joorabchi**

Degree: **Master of Science**

Title of Thesis: ***EmailTime*: Visualization and Analysis of Email Dataset**

Examining Committee:

Chair: **Dr. John Bowes**
Professor

Dr. Christopher D. Shaw
Senior Supervisor
Associate Professor

Dr. John Dill
Supervisor
Professor Emeritus

Dr. Lyn Bartram
Internal Examiner
Assistant Professor

Date Defended/Approved: 25th November, 2010

ABSTRACT

Although the discovery and analysis of communication patterns in large complex email datasets is a difficult task, it can be a valuable source of information. We describe the design and visualization technique of *EmailTime*, a tool for visual analysis of email correspondence patterns over the course of time that interactively portrays personal and interpersonal networks. *EmailTime* helps email dataset explorers interpret archived messages by providing interactions, visualizing histograms and measuring centrality (To, Cc and Sent) and frequency (sent and received). We performed case studies on the *Enron* dataset to discover impacts of executive position on the email behaviour of organizational workers using a series of metrics e.g. number of sent and received emails as determined by *From*:, *To:* and *Cc:* fields, recipient counts of sent emails. In addition, we evaluated the visualization through pilot and user studies to find out whether users were able to recognize the selected capabilities.

Keywords: Email visualization, email correspondence, *Enron* case study, *EmailTime*, usability study, knowledge visualization, information visualization.

DEDICATION

I would like to dedicate my Master of Science thesis to my wonderful family who offered me unlimited love and support throughout the course of this thesis.

ACKNOWLEDGEMENTS

I am delighted to offer my blessings to all of those who supported me during the completion of this thesis, which has not been possible without their time and cooperation.

I owe a debt of gratitude to my senior supervisor, Christopher D. Shaw for his careful guidance and invaluable support from the very beginning of this thesis, to its end.

Special regards as well to my other committee members, John Dill and Lyn Bartram for their support, dedicated reading time and feedback.

I extend my deepest appreciation to my wonderful colleague Ji-Dong Yim. Ji-Dong, your great advice, insightful criticisms, and constant encouragement from the initial to the final level enable me in the writing of this thesis.

Thanks go to my dear colleague, Jeffrey Guenther, for his help in editing and correcting notes.

I would also like to thank those who agreed to be participated in my studies.

Finally, I would be remiss without mentioning my dear sister, Mona, my wonderful friends and my great HVI and VA lab colleagues, who always provide me with an encouraging research atmosphere.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1: INTRODUCTION

Email datasets are an interesting subject of study as the nature of data is private and they contain a large amount of information about peoples' correspondents, thoughts and activities. They consist of personal records of people's past interactions including work interactions, relationships with family members, friends, etc. Because of the interesting information embedded within datasets (e.g. peoples' correspondences and correspondents), visualization and visual analysis of these datasets would be valuable in order to gain insight and understanding.

Finding and working with real email datasets has been a challenge because of the private nature of the email data. Moreover, they are usually large and time consuming to read. Clearly, some understanding is gained by looking at the data from a different perspective, such as viewing email messages distributed over time, or organized other specific ways provided by information visualization techniques. In these situations, interactive techniques of information visualizations can enable the users to make sense of their data and make new observations about the datasets.

According to Donath from MIT Media Lab [26], "visualizations can provide some of the missing context by revealing the data and patterns that are hidden within the email datasets". Based on the type of visualization, visualizations can reveal different attributes. For example, at the individual level, the visualizations can show when the owner of the dataset is active, who his/her main correspondents are, what the connections are between them, major shifts in his/her contacts, etc. At the organizational level, visualizations can show who works together, for how long, who is the link between groups of people, how new employees are integrated into the company, etc [26].

The general purpose of this thesis is to:

- Improve our understanding of the visualization and analysis of email datasets,
- Design and build a system (called *EmailTime*) for visualization and analysis of this type of data and,
- Evaluate *EmailTime* using different types of experiments:
    - Run case studies on the *Enron* dataset to investigate the impacts of executive (organizational) position on the email behaviour of organizational workers using a series of metric. Metrics are the number of sent emails as determined by the *From:* field, number of received emails as determined by the *To:* and *Cc:* fields (for *Form:*, *To:* and *Cc:* fields see Figure 3-1), recipient counts of sent emails, number of email addresses, and number of created folders.
    - Run usability studies on *EmailTime*'s capabilities to investigate (through the visualization) users are able to use the system's capabilities in order to find:

- Changes of activities over time (e.g. switching from one email address to another one),
- Correspondence patterns between email users over time (such as the most frequent correspondents and types of their correspondences; general or private message) and,
- Role of the owners of email addresses in an event (e.g. secretary or leader in a biweekly meeting in an organization dataset).

1.1 Background

Because of the interest in this topic, different tools and techniques have been introduced in order to support the visualization and/or analysis. In this section, I briefly explain some notable work, specifically those I was inspired by for my thesis.

Heer from Stanford University has made novel visualization techniques for exploring data; software tools that simplify visualization creation, customization, and collaborative analysis. *Vizster* [2] (See Figure 6-2) and *Enronic* (Exporting *Enron*) [3] (See Figure 6-3) are his remarkable works which are discussed in chapter 2: Related Work. He led the design of the *Prefuse* toolkit [25] which our system *EmailTime* uses for part of its interaction. For example, basic interaction techniques such as zooming and panning are inherited from the *Prefuse* toolkit.

Schneiderman and *Perer* from University of Maryland have created novel Social network analysis (SNA) tools such as *SocialAction* [8] (See Figure 6-7) which inspired me to explore this topic. *SocialAction* is a social network analysis tool that integrates visualization and statistics to improve the analytical process. It focuses on visualizing the network by graphing. In *EmailTime,* we aim to have the same approach, a combination of statistic and visual analysis, except our visualization type is an XY scatter plot.

Gloor also did remarkable work in Social Network Analysis such as *TeCFlow (Condor)* [5] (See Figure 6-6)*. TeCFlow* tool visualizes the temporal evolution of communication patterns among groups of people to analyze the email dataset. In part of the case study that we explained in chapter 4.1, we benefit from the concept of Contribution Index (1) that was introduced in [5] to specify the role (sender, receiver or both) of email address. This index is near to –1 for the receivers and +1 for the senders. In this formula they used *received* emails whereas we expanded it to *received emails as determined by To: field* (To-CI) and *received emails as determined by Cc: field* (Cc-CI) to see the impact of *To:* and *Cc:* fields separately.

$$\frac{\text{emails sent} - \text{emails received}}{\text{total of emails sent and received}} \qquad (1)\ [5]$$

Viégas's work such as *Social Network Fragments* (*SNF*), *PostHistory* and *Themail* [17, 18, 19, and 20] (See Figure 6-10 and Figure 6-11) focuses on the social, collaborative, and artistic aspects of information visualization*.* They visualize the emails by graph and chronologically by plot with a social and artistic perspective (e.g. see the evolution of users' relationships over the years and the overall picture of their past communications).

There are many other notable works such as *Rohall*'s *ReMail* [21] (See Figure 6-8) and *Kerr*'s *ThreadArc* [22] (See Figure 6-16). They concentrate on visualizing other aspects of email, e.g. email thread, conversations and navigation.

1.2 Thesis Contribution

The main contribution of this thesis is visualizing:

- Changes of activities over time,
- Correspondence patterns between email users over time and,
- Role of the owners of email addresses in an event.

The general contribution is the development of *EmailTime* and the results of experiments on it. We have designed and implemented *EmailTime*, a visual analysis of email correspondence patterns that visualizes the relationships between individual messages and correspondents over the course of time. *EmailTime* provides visual and interactive access to the electronic mail archives. Our interest is to visualize the email dataset of individuals and groups in order to examine the patterns in the email correspondences and compare the behaviour pattern of his/her different email addresses, activity level and sent/received email frequency of email addresses.

Moreover, we were interested in evaluating *EmailTime* system. In order to do so, we performed the following experiments:

- Case studies on *Enron* dataset, to investigate the impacts of the executive (organizational) positions on the email behaviour of organizational workers with respect to a series of metrics.

- Pilot and user studies on *EmailTime*'s capabilities to investigate whether users are able to discover interpersonal social activities in email datasets (such as changes of activities, roles of the owners of email addresses in an event and the correspondence patterns between email users) using the selected capabilities.

Details of the experiments are presented in the next subsection.

1.3 Purpose and Result

1.3.1 Case Study

1.3.1.1 Purpose and Research Question

In the case study we are interested in the impact of organizational positions on the email behaviour of organizational workers (using several metrics). The dataset is from the *Enron* Corporation in between January 2000 and December 2001. It includes 101 *Enron* workers in seven executive positions: CEO, President, Vice President, Manager, Director, Employee and Trader (we refer to these as "*organizational positions*"). The metrics are the number of sent emails as determined by the *From:* field, number of received emails as determined by the *To:* and *Cc:* fields, number of the email addresses that a person owned, number of his/her own created folders and recipient count of the sent emails. Recipient count of sent emails is the number of recipients (in *To:*

and *Cc:* fields) of the email. For the data analysis, the statistical tool SPSS [31] was used to analyze the numerical results.

1.3.1.2 Results

From the analysis on the activity levels of organizational positions (with respect to the number of sent and received emails as determined by the *From*:, *To:* and *Cc:* fields), solely based on observation and diagram we recognized three categories of activity and divided the organizational positions into Inactive, Moderate and Active. Managers and Employees were Active, Traders and Directors were Inactive, and the rest were Moderate.

From the analysis on the role (sender, receiver or both) of email addresses of seven organizational positions, we realized that managerial groups (such as CEO and President) tend to be receivers more often than the staff in lower positions.

Analysis on the recipient count of sent emails (number of recipients in *To:* and *Cc:* fields) shows a statistically significant difference between CEOs and other groups in sending emails with the Large number of recipients. Traders and then Managers sent emails with Medium number of recipients more than any other groups.

According to the results, since no relationship between the number of created folders and organizational positions was found in our dataset, we believe a user’s choice in the number of created folders is subjective. More details are presented in section 4.1.

1.3.2 Usability Study

1.3.2.1 Purpose and Hypothesis

As we used *EmailTime* to explore the *Enron* email dataset, we recognized some capabilities with the *EmailTime* system. In the usability study, we hypothesized that *EmailTime* visualization enables users/analysts to find interpersonal social activity in email datasets through visualization including:

- Changes of activities over time (e.g. switching from one email address to another),
- Correspondence patterns between email users over time (such as the most frequent correspondents and types of their correspondences; general or private messages) and,
- Role of the owners of email addresses in an event (e.g. secretary or leader in a biweekly meeting in an organization dataset)

These capabilities form a basis for interpreting data that is visualized by *EmailTime*, such as:

1. *Time Comparison*: Compare different time periods to each other and recognize their differences with respect to the crowded eras, large gaps (no activity), sent emails with Large number of recipients, etc.
2. *Most Frequent Correspondents*: Find the most frequent correspondents of a person and types of their correspondences (private or general messages based on the recipient count of sent

emails).

3. *Email Address Comparison*: Compare different email addresses to each other with respect to the duration, and activity level and role (sender, receiver or both) of each email address, and discover which email addresses were switched.

The focus was on investigating the dataset of individuals. Twenty-three graduate students from SIAT, SFU participated in the one to two hour testing sessions (four participants in Pilot Study I, six participants in Pilot Study II and thirteen participants in User Study).

1.3.2.2 Results

The scenarios in the user study contain inferential and deductive tasks. Therefore, the tasks were not easy, as the users need to do inference and draw conclusions. The majority of the participants were able to complete the tasks but some of them were confused in the deductive part of the scenarios and asked the observer (me). We expected that the participants would accomplish the scenarios between 7 to 10 minutes. It appears that users accomplished easier tasks faster.

Generally most of the participants mentioned that they are able to recognize similar scenarios. More than 80% of them added that the introduction was necessary to get the concept of the visualization and working with system. Half of them agreed that they needed to concentrate in working with the system and answering the tasks. From the likes and dislikes of the visualization plot and control panel in the post questionnaire, the participants' comments were constructive in terms of how to improve the interactivity of the system. More details are presented in section 4.2.

1.4 Thesis Outline

Chapter 1: gives the reader an introduction of the study.

Chapter 2: reviews the related works on interacting with the email datasets. As we explored them, we recognized some works have similar approaches (e.g. thread-based, graph-based, etc.). As a result, one method of classification is to group these tools based on their focus into different categories. Therefore at the beginning of this chapter, we propose a categorization of these works in Table 2-1.

Chapter 3: gives an overview of the *EmailTime* system. It first describes the visualization design. Then we explain system functionalities including basic interactions such as zooming and panning; visibility filters which is a *node type selector* applied to the three types of email node - *Sent*, *To*, and *Cc*; search options; some statistic measurements namely frequency, centrality and histogram views of sent and received emails. Finally, we present some capabilities of the *EmailTime* system through several examples. We specified these capabilities while we investigated *Enron* email dataset.

Chapter 4: explains the experiments (research question and methodology) that we have done on the *Enron* email dataset as our benchmark. We then present the result of the studies and a discussion on that.

Chapter 5: concludes with suggestions for the future work.

2: RELATED WORK

Different tools and techniques have been introduced in order to support visualization and analysis of email datasets.

2.1 Proposed Category

As we explored different tools and techniques for analyzing email datasets, we recognized some work has similar approaches. Therefore, I have chosen to categorize these tools based on their *main* approach (see Table 2-1).

Table 2-1. Categories of the visual interactive tools for the email dataset.

Focus		**Description**
Thread-based	**Graph-based**	*Category 1*: Visualizing email thread and reply chain by node-link diagrams.
	Statistic-based	*Category 2*: Visualizing email thread and reply chain by statistical techniques and diagrams (plot, histogram, etc.).
	Other	-
Non thread-based	**Graph-based**	*Category 3*: Visualizing network, communication, etc. by node-link diagrams.
	Statistic-based	*Category 4*: Visualizing network, communication, etc. by statistical techniques and diagrams (plot, histogram, etc.).
	Other	*Category 5*: Visualizing network, communication, etc. by icon, mountain map, etc.

At the high level, we identified two different perspectives for approaching the visualization of email datasets. The focus of the first aspect is on email thread and reply chain (Thread-based category) whereas the focus of the second aspect is on network and communication (Non thread-based category). In terms of the visualization techniques, we identified two main approaches which we grouped those into Graph-based category and Statistic-based category. Therefore, we grouped the works that use different kinds of trees, graphs, graphic metrics, etc. into the Graph-based category and the ones that use different kinds of timelines, plots, statistic metrics, etc. into the Statistic-based category. “Other” in Table 2-1 refers to the works that visualize the email datasets using other visualization techniques such as Icon, Mountain [29], etc. The reason for this classification is there is much work that used Graph-based and Statistic-based approaches. Following we mention some of the works in each area.

Work in category 1 concentrates on visualizing email’s thread and conversations by node-link diagrams and graphic metrics. *Rohall* et al [21] (See Figure 6-16) visualized message threads over time along with message content to display the relationship among messages. *Kerr* introduced *Thread Arcs* [22] (See Figure 6-16) that represents the threads of email conversations as a sequence of nodes (messages) along a line, with semicircular arcs linking an email to its reply. The chronology of the thread is coded by position so it gives a visual summary of how the conversation has progressed over time. Seven key qualities of *Thread Arc* are chronology, relationships, stability, compactness, attribute highlighting, scale and interpretation/sense in comparison to other techniques such as Tree Diagram and Tree Table in [22]. *Venolia* et al. [4] (See Figure 6-17) investigated the character of email conversations by examining and visualizing conversation patterns. These were visualized using a tree-based approach.

Category 2 visualizes the email's thread and conversations by statistical techniques (e.g. plot, histogram, bar-chart, and statistical metrics). *Perer* and *Shneiderman* [24] (See Figure 6-18) presented threading messages by common subject lines and reply-chain information in email headers. This allowed the interpretation of archived messages by providing access to the full scope of discussions that stretch beyond the thread.

Category 3 concentrates on visualizing the communication network with node-link diagrams and graph metrics. In this category, *Heer* presented *Enronic* [3] (See Figure 6-3) that integrates information visualization techniques with various algorithms to explore the email document, including ANLP (Applied Natural Language Processing), social network inference, message categorization & community analysis. Nodes are small coloured pie charts denoting email content categories (e.g. company policies, regulations) while edges represent direct email messages between people, so that clusters show patterns of social networks & community structures.

Xiaoyan Yu et al developed *VisPEAM* [1] (See Figure 6-1), which is based on *Vizster* [2] and *Prefuse* [25]. It enables the user to examine emails, display the frequency of exchanged messages for a particular topic, manage their email collections and search emails by different search criteria.

Gloor et al introduced *TeCFlow* [5] (See Figure 6-6) - a Temporal Communication Flow Visualizer for Social Network Analysis for analyzing the email dataset. They were interested in discovering suspicious activity in Enron email dataset using filtering, term view map and type of networks (COINs (Collaborative Innovation Networks), CLNs (Collaborative Learning Networks) and CINs (Collaborative Interest Networks)-See [5] for details). Then they find about group betweenness centrality, density, and contribution index for measuring the activity of an individual as a sender or receiver.

Chapanond et al [6] developed directed and undirected email graph and computed and studied several graph metrics such as degree distribution, average distance ratio, clustering coefficient and compactness to discover the properties of Enron email graph. They also mentioned that for creating a benchmark pre-processing of data has a significant influence on the results.

Diesner and *Carley* [7] investigated structural properties of the networks in *Enron* to recognize the key players over time. They found that during the *Enron* fall the network had been denser and more connected than during normal times.

Perer and *Shneiderman* presented *SocialAction* [8] (See Figure 6-7), a tool to effectively understand social networks. It uses attribute ranking and coordinated views (node-link diagrams, ordered list, etc.) to help users systematically examine numerous Social Network Analysis (SNA) measures. Users can filter nodes and find outliers, aggregate nodes and find cohesive subgroups and communities, find patterns by viewing different edge types, etc.

By considering an email network as a social network, *Fisher* and *Dourish* [4] described two types of systems supporting everyday collaboration, displaying ways to represent the temporal and social structures of online activity. They developed *Soylent* (See Figure 6-4) to find social and temporal structures and elements in interaction of electronic records of activity. Then introduced *TellMeAbout* (See Figure 6-5), an awareness tools based on structural information. It is an initial client that uses the *Soylent* infrastructure to provide end users with an understanding of the structures within which their work is embedded.

Heer and *Boyd* introduced *Vizster* [2] (See Figure 6-2) to support visual exploration and identify the community structures. *Vizster* is an interactive visualization tool for online social networks such as friendship, forming an undirected graph in which users are the nodes and friendship links are the edges. It supports a range of exploratory search features, users' profiles, linkage view, connectivity highlighting, and community structures visualization. Public installation and controlled studies of the system demonstrate the system's usability and potential for engaged social activity.

Category 4 visualizes the network, communication, by statistical techniques (e.g. plot, histogram, barchart, and other statistical metrics). As we visualize the network and correspondents over time (by plot, histogram), our own work, *EmailTime* fits to category 4.

Viégas et al developed *Themail* [17] (See Figure 6-13), which visualizes the conversational history between the owner of the email address and one of her email contacts. It displays a series of columns of keywords in the exchanged messages over time. By this visualization they can answer questions like “what sorts of things do the owner of the archive talk about with each of her email contacts?” and “how do her email conversations with one person differ from those with other people?” From the user study, participants were quite excited to use *Themail* to look back at their email archives, see the evolution of their relationships and gain a new perspective on that.

Email Mining Toolkit (EMT) developed by *Stolfo* et al [11, 12, 13, 14] (See Figure 6-10 and Figure 6-11), is a data mining system that visualizes the details of emails and computes “behaviour profiles or models” of user email accounts. It analyzes email archives by graphical display to explore relationships between users and the chronological flow of an email message. *EMT* includes different features such as User Clique, Enclave Clique, Email Flow, Message Table, Similar Users, Usage Histogram and Recipient Frequency. It has security applications, including virus and spam detection, as well as security policy violations.

Leuski et al represents *eArchivarius* [16] (See Figure 6-12) that visualizes the relationships between messages, people, and events. It combines ranked retrieval with cluster-based and time-based navigation. They used a clustering technique to minimize the distance between nodes that have a high frequency of email traffic between them.

Viégas and *Donath* [18, 19, 20] (See Figure 6-14 and Figure 6-15) presented two visualizations of email, *Social Network Fragments* (*SNF*) which displays a network graph with the email contacts as nodes (category 1) and *PostHistory* which displays the chronological patterns of communication between two individual contacts (category 3).

Category 5 focuses on visualizing the network, communication, by other techniques (e.g. Icon, Mountain [29], etc.). *Mandic* and *Kerne* developed *faMailiar* [9, 10] (See Figure 6-8 and Figure 6-9), an intimacy-based email visualization. They defined intimacy computationally for an email as a combination of two metrics: contact intimacy category and message intimacy weight (See [9, 10] for more details). It uses brightness, hue and iconography to visualize intimacy over time to analyze the email's personal and social role. Results from the user study shows remembering past activities and contacts involved, the time of the day/month that any specific contact would next email them and how long it would take a specific contact to respond after receiving an email from them.

There are other various ways to classify the visual interactive tools. *Perer* et al categorized these tools into six categories based on the type of data (archived/online) and the creation location (individual/organizational/social) [15]. Regarding to this classification, our work fits within category 4 and 5, as we present new techniques for exploring the archived email of an individual and groups of people. Depending on the memory features, it also can explore the dataset of more people such as an organization.

Table 2-2. *Perer* et al explored types of interactions with email collections [15].

Current	*Category 1*. Controlling an individual user's current inbox
	Category 2. Controlling current email within an organization
	Category 3. Controlling current conversations in a social space
Archived	*Category 4*. Analyzing an archive of an individual's messages
	Category 5. Analyzing an archive of an organization's messages
	Category 6. Analyzing an archive of a social space

3: EMAILTIME OVERVIEW

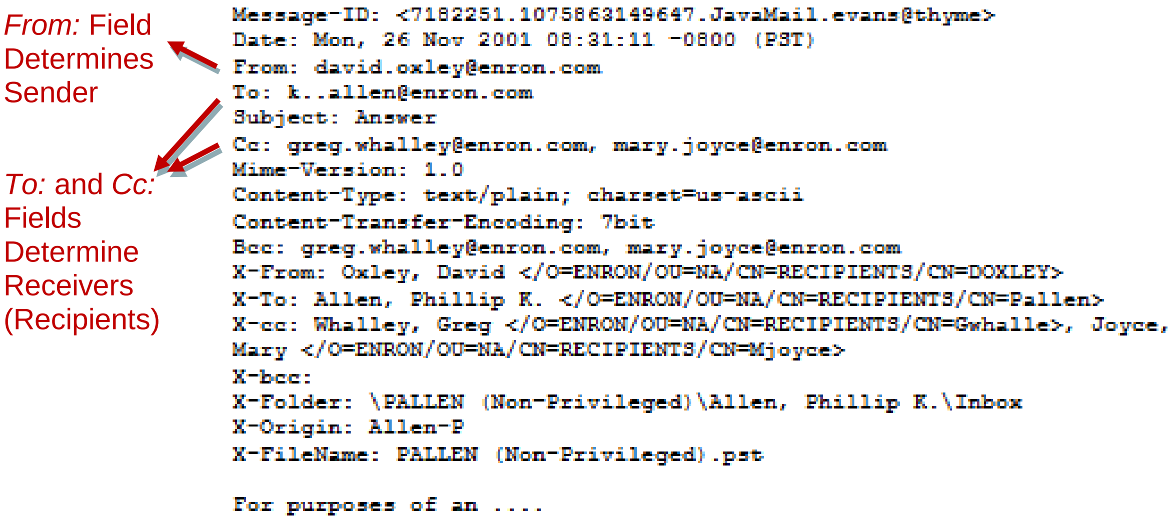
Our system, *EmailTime*, visualizes the communication activities found in a collection of emails for a period of time. *EmailTime* provides a visual analysis of email correspondence patterns over the course of time that interactively portrays personal and interpersonal networks using the correspondence in the email dataset. Our approach is to make time as a primary variable of interest, and plot emails along a timeline. *EmailTime* helps email dataset explorers interpret archived messages by providing zooming, panning, filtering, highlighting, displaying message content, controlling Y-axis and Time-axis. To support analysis, it also measures ent and Received frequency, Sent, To and Cc centrality on the communication graph and visualizes histograms.

The original email dataset is from the *Enron* Email Corpus [27]. We explain this dataset in more detail in section 4.1.2, where we used it as our benchmark for the case study. To apply our visualization to the archive, we inserted the data into the system from either one or more users' datasets, and filtered out the emails in which we were interested.

3.1 Visualization Design

In order to have a common terminology, *sender* is an email address of a person who sends an email as determined by the *From:* field of the email.

Receiver/Recipient is an email address of a person who receives an email as determined by the *To:* or *Cc:* fields the email (See Figure 3-1).



From: Field Determines Sender
To: and *Cc:* Fields Determine Receivers (Recipients)

```
Message-ID: <7182251.1075863149647.JavaMail.evans@thyme>
Date: Mon, 26 Nov 2001 08:31:11 -0800 (PST)
From: david.oxley@enron.com
To: k..allen@enron.com
Subject: Answer
Cc: greg.whalley@enron.com, mary.joyce@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: greg.whalley@enron.com, mary.joyce@enron.com
X-From: Oxley, David </O=ENRON/OU=NA/CN=RECIPIENTS/CN=DOXLEY>
X-To: Allen, Phillip K. </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Pallen>
X-cc: Whalley, Greg </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Gwhalle>, Joyce,
Mary </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Mjoyce>
X-bcc:
X-Folder: \PALLEN (Non-Privileged)\Allen, Phillip K.\Inbox
X-Origin: Allen-P
X-FileName: PALLEN (Non-Privileged).pst

For purposes of an ....
```

Figure 3-1. A sample email (message) that is processed by *EmailTime*. *From:*, *To:* and *Cc:* fields are specified.

In the visualization, the horizontal dimension represents (sent) time. The vertical dimension can be assigned to different attributes of email. One email can have multiple circles (nodes) in three different colors. A black circle (node) for the sender to indicate the sent email (the email address in the *From:* field), a blue circle (node) for *To:* recipient and a green circle (node) for *Cc:* recipient (See Figure 3-1 and Figure 3-3). Size of the black nodes represents the number of recipients (*To* + *Cc*) of the email. Bigger sent nodes have more recipients.

A common representation for email correspondence is shown in Figure 3-2. It is a graph representation of a small network. In this graph, each node is an email address. An arrowed links show the sending direction. For example, Chris and Beth exchanged emails whereas Chris and David did not exchange any email. The graph view is good in showing the connectivity but limited in displaying temporal properties and types of correspondences in a network. Figure 3-3 displays an *EmailTime* plot of the same network for Aaron's dataset (all the emails that Aaron was involved as either sender or receiver).

As mentioned above a message can draw multiple circles in three different colors. For example, when Aaron sent the Message #1 to Beth; *EmailTime* plots a black circle on aaron@a.org's line to indicate sent email and a blue circle on beth@b.org's line to indicate received email as determined by the *To:* field. The glyphs are placed on the same imaginary vertical line to indicate that both glyphs are at the same time (this time is specified by the "Date" field in the email file in Figure 3-1); in fact, both are the same message. When Beth sent the Message #2 to Aaron and copied it to Chris and David; *EmailTime* plots a bigger black circle on beth@b.org's line, a blue circle on aaron@a.org's line and a green circle on chris@c.org's and david@d.org's lines for the Message #2 (on the same imaginary vertical line). Therefore, the black circle gets larger as the number of recipients increased.

Thus, we redundantly plot one circle for each *From*:, *To*: or *Cc*: field in an email. The purpose of such redundancy is to allow the viewer to infer patterns of correspondence without actually drawing marks for links. In addition, as you may realize the plot representation contains larger amount of details.

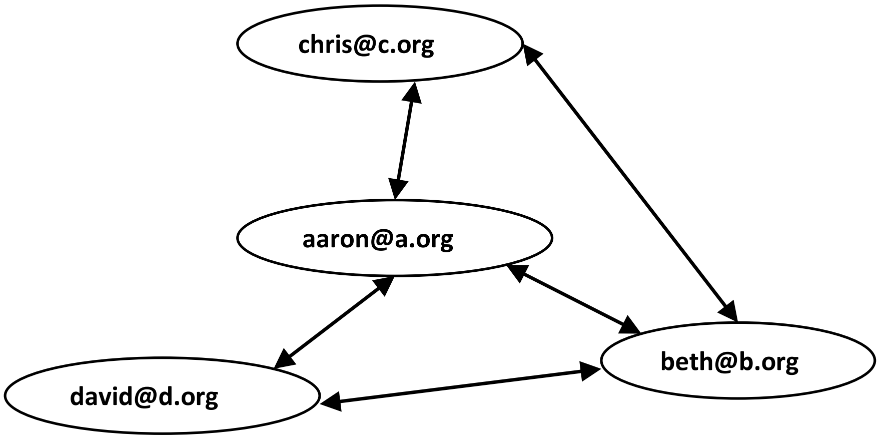


Figure 3-2. Graph view of a small network.

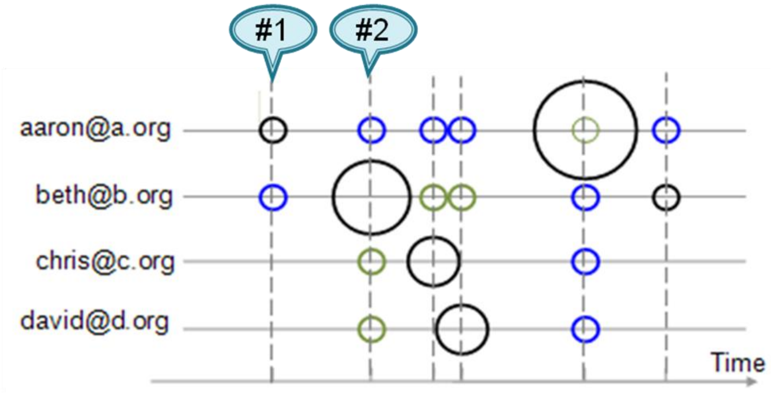


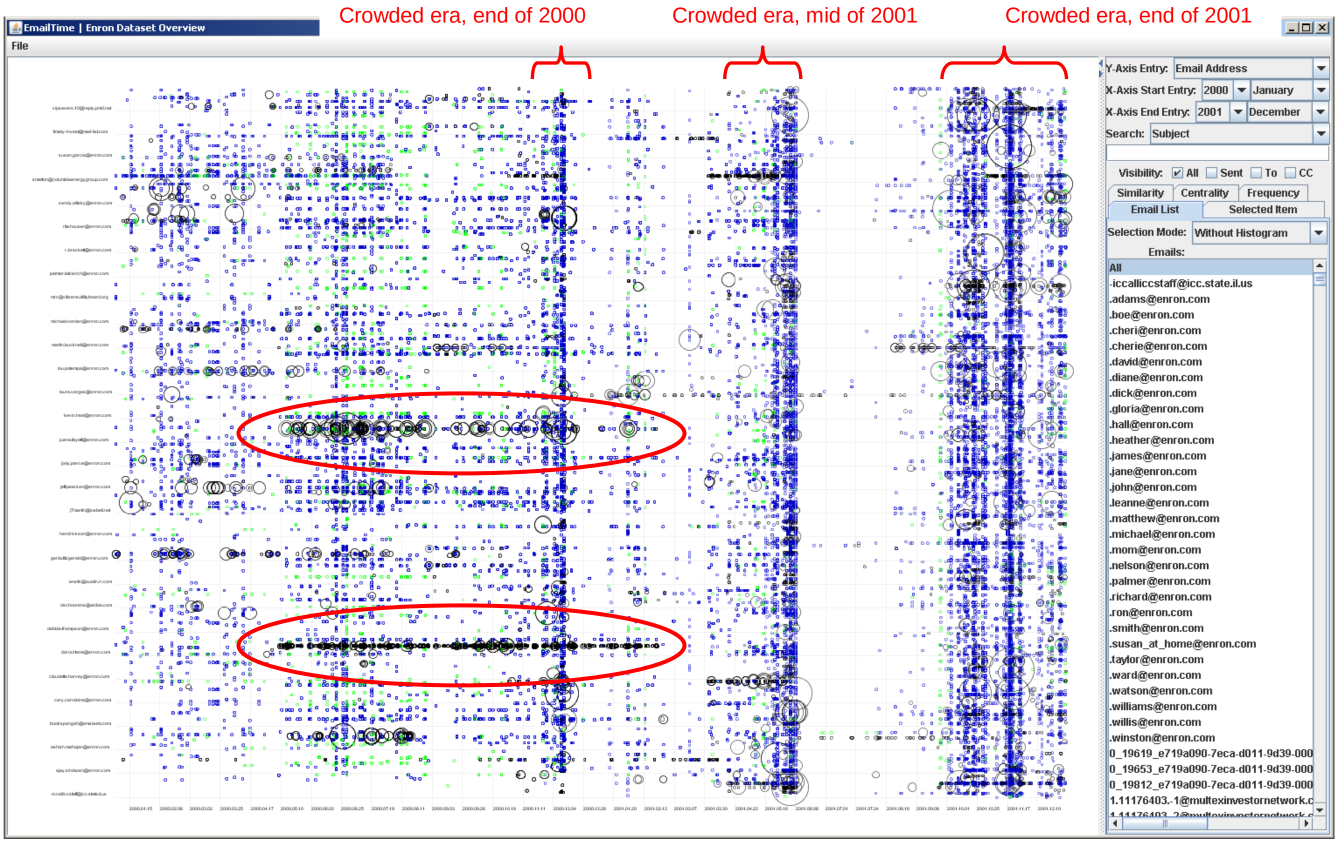
Figure 3-3. Plot view of the same network for aaron@a.org by *EmailTime*. A message can draw multiple circles in three different colours; black for sent email as determined by *From*: field, blue for received email as determined by *To:* field, and green for received email as determined by *Cc:* field. The size of a sent node represents the number of recipients. (e.g. the Message #2 is sent by Beth to Aaron, and copied it to Chris and David.)

Figure 3-4 is a snapshot of the *EmailTime* visualization. The right side is the control panel. The Left side displays a visualization of collection of emails from datasets of six *Enron* people with different organizational positions: David Delaney (CEO), Stanley Horton (president), Thomas Martin (vice president), Andrew Lewis (director), Martin Cuilla (manager), and Albert Meyers (employee).

In the visualization, the X-axis depicts the period of time from January 2000 to December 2001, and the Y-axis is assigned to email addresses. Such a dimensional combination helps the user find visual patterns of email correspondence over time. For example, we can clearly see that received emails as determined by *To*: field (blue circles) are very crowded at the end of 2000, in mid 2001 and at the end of 2001 (See Figure 3-4, red arcs). Seeing several or tens of large black circles in each period, we can say that the high density of recipients is caused by a limited number of actors who sent out announcement emails.

Two (or three) low density gaps then follow the first and the second high density periods, which might mean that most of the *Enron* workers suddenly stopped using emails or the dataset was lost when it is collected or part of the dataset was intentionally erased before making it public. The latter seems to make more sense, considering that the *Enron Corporation* collapsed in December 2000. So it might be the case that many emails were filtered because they include private and critical contents related to the company’s tragedy.

Two active senders are apparent around the second half of 2000 (two rows of black nodes that are determined by the red ellipses in Figure 3-4), where the average number of recipients of their emails look different. The details may be not clear due to the crowded nature of the visualization. Those details become obvious when zoom and visibility filters are applied to the visualization. In the following sections, we describe how *EmailTime* highlights the visual patterns with its capabilities.



Y-Axis Entry: Email Address
X-Axis Start Entry: 2000 January
X-Axis End Entry: 2001 December
Search: Subject

Visibility: ☑ All ☐ Sent ☐ To ☐ CC

Selection Mode: Without Histogram

Emails:

All
-iccalliccstaff@icc.state.il.us
.adams@enron.com
.boe@enron.com
.cheri@enron.com
.cherie@enron.com
.david@enron.com
.diane@enron.com
.dick@enron.com
.gloria@enron.com
.hall@enron.com
.heather@enron.com
.james@enron.com
.jane@enron.com
.john@enron.com
.leanne@enron.com
.matthew@enron.com
.michael@enron.com
.mom@enron.com
.nelson@enron.com
.palmer@enron.com
.richard@enron.com
.ron@enron.com
.smith@enron.com
.susan_at_home@enron.com
.taylor@enron.com
.ward@enron.com
.watson@enron.com
.williams@enron.com
.willis@enron.com
.winston@enron.com
0_19619_e719a090-7eca-d011-9d39-000
0_19653_e719a090-7eca-d011-9d39-000
0_19812_e719a090-7eca-d011-9d39-000
1.11176403.-1@multexinvestornetwork.c

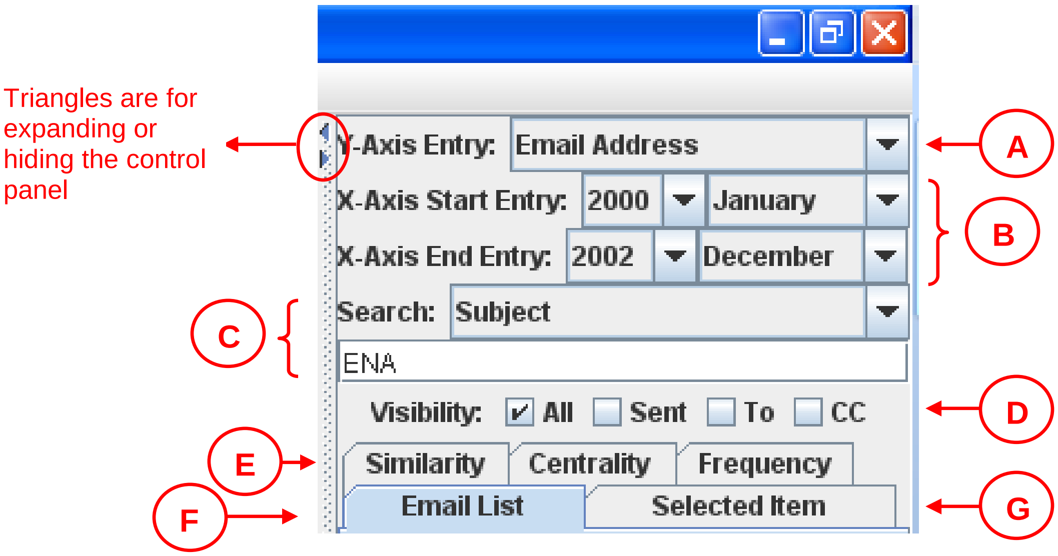
Figure 3-4. The *EmailTime* visualization. The left side presents a visualization displaying a collection of emails from datasets of six *Enron* workers. The activities of email addresses (Y-axis) are plotted over time (X-axis). On the right side is the control panel that provides axes controls, keyword search, visibility filters, centrality and frequency analysis tabs, and more option tabs.

3.2 *EmailTime* Interactions

As *EmailTime* visualization is written in Java based on the *Prefuse* toolkit [25], basic interaction styles such as zooming and panning are inherited from the toolkit. In addition, *EmailTime* highlights selected node sets by colour and pops up details of a message (message subject and owner of the message) in a tooltip when the mouse rolls over each node. The *EmailTime* control panel in Figure 3-5 also provides other features. The Y-axis control assigns different attributes to the vertical dimension of the canvas (See Figure 3-5-A); attributes such as Email

Address, Message Subject, Message Type, Message Length, Message Date/Time, Email Address shows the most interesting visualization result. X-axis (Time-axis) control adjusts the horizontal dimension to a time period specified by the start and end year and month (See Figure 3-5-B). By clicking on each node, the content of the email node is displayed in the "Selected Item" tab in the control panel (See Figure 3-5-G).

Analyst/user can make new findings in the visualization plot using system's interactions. We explained it in detail in section 3.4 Discussion on System Capabilities through examples.



Triangles are for expanding or hiding the control panel

Y-Axis Entry: Email Address

X-Axis Start Entry: 2000 January

X-Axis End Entry: 2002 December

Search: Subject

ENA

Visibility: ☑ All □ Sent □ To □ CC

Figure 3-5. The *EmailTime* control panel. A) Y-axis list. B) Start and end year and month lists of X-axis (Time-axis). C) Search option. D) Visibility filters. E) Function tabs F) Email List tab contains the list of all email addresses G) Selected Item tab displays the content of a selected message.

3.2.1 Visibility Filters

Since our visualization displays a large number of emails and draws multiple glyphs for a message, the very basic and important filter in the system is the *Node Type Selector* (See Figure 3-5-D) applied to the three types of email node (glyph) – *Sent* (black nodes), *To* (blue nodes), and *Cc* (green nodes).

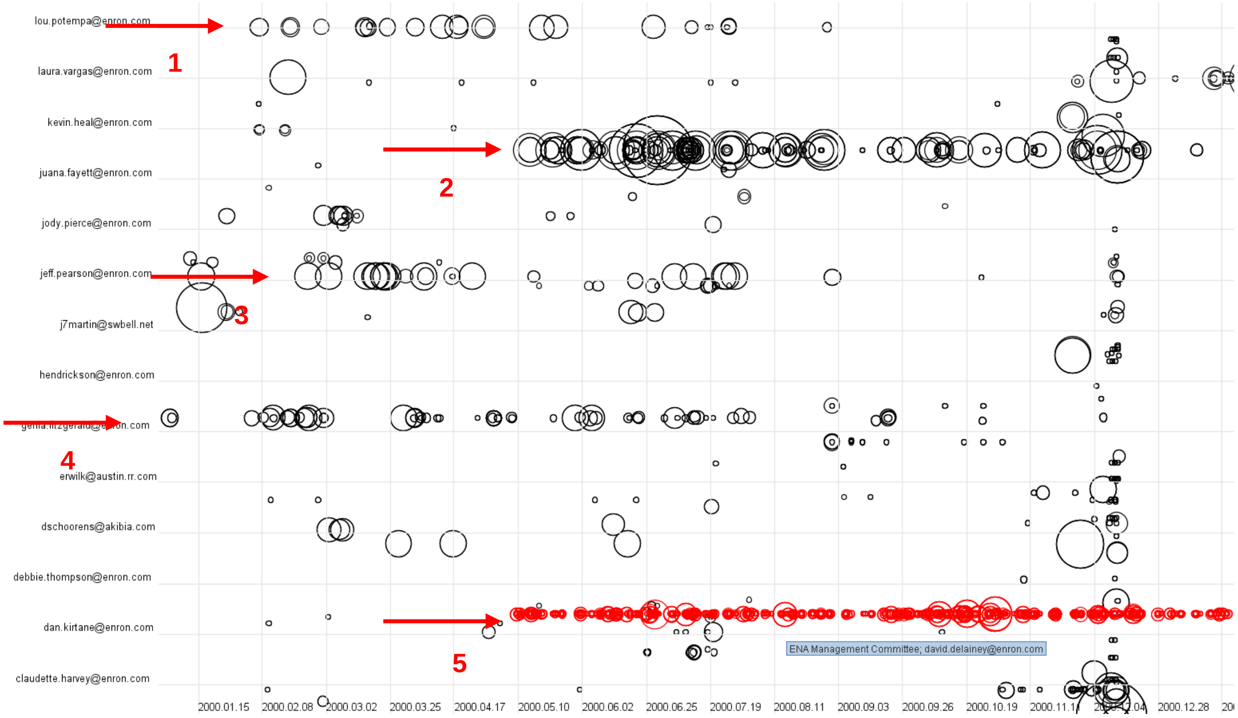


Figure 3-6. A subset of Sent emails (black nodes) within the example dataset in year 2000. The highlighted nodes at the bottom (in red) are the messages that the actor david.delainey@enron.com sent.

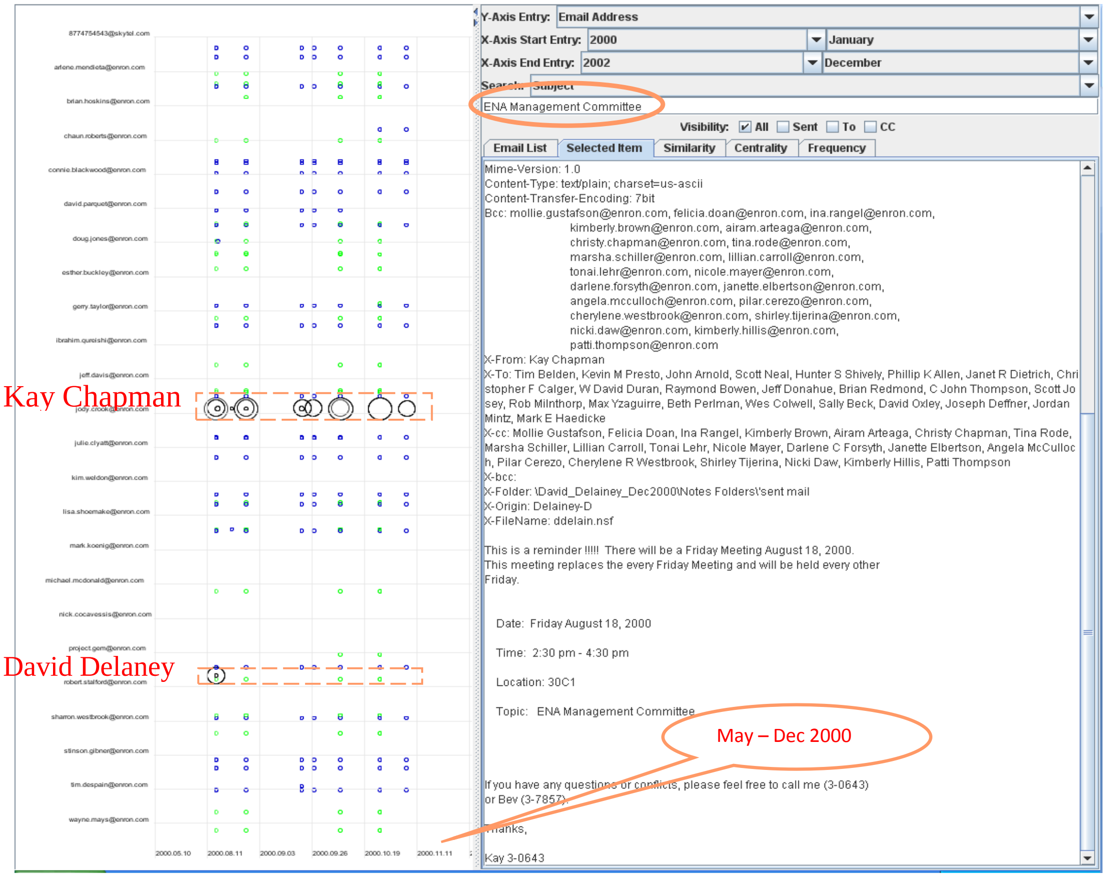
In Figure 3-6, where nodes for received email (as determined by the *To:* and *Cc:* fields) are filtered out, we can clearly see different patterns of sending activities. We can recognize five active email senders (the red arrows – rows of black nodes) who were not obvious in Figure 3-4. For example, Sender 2 (Kay Chapman) frequently spread emails to many people, while Sender 5 (David Delaney) usually talked to a small group of people as indicated by the size of sent nodes. In similar ways, received emails make interesting patterns each.

These patterns are able to help the analyst in discovering new findings or deeper investigations on social phenomena in a group of email users, e.g. temporal communication patterns, social roles, etc (See section 3.4 System Capabilities for more details).

3.2.2 Search Option

Another filter is based on the text search function in the *EmailTime* control panel. It helps analysts, especially when they are expert on the dataset, to find an event (e.g. a biweekly meeting) by searching the related keywords and learn more about those by reading the content of emails. The search options of *EmailTime* are coupled to *Subject*, *From*, *To*, or *Cc* field, and filter a particular group of messages in the visualization (See Figure 3-5-C).

For example in Figure 3-7, when exploring the dataset of David Delaney (*Enron* CEO) we discover david.w.delainey@enron.com and kay.chapman@enron.com exchanged several emails related to ENA (apparently an association in the *Enron* Company) during May to December 2000. When we search for a particular subject such as “ENA Management Committee”, we found that David Delaney sent a message to a group of people about having a meeting every second Friday afternoon (acting as a leader of the meeting, who setup the meeting). Then Kay Chapman sent out several messages with this subject frequently to remind the participants of having a meeting in the following Friday with the mentioned time and location of the meeting (acting as a secretary). The announcement emails were usually sent on Wednesday or Thursdays. As you can see, the blue and green nodes show the recipients of the emails sent by Kay Chapman.



Y-Axis Entry: Email Address
X-Axis Start Entry: 2000 January
X-Axis End Entry: 2002 December
Search: Subject
ENA Management Committee
Visibility: ☑ All □ Sent □ To □ CC

```
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: mollie.gustafson@enron.com, felicia.doan@enron.com, ina.rangel@enron.com,
                    kimberly.brown@enron.com, airam.arteaga@enron.com,
                    christy.chapman@enron.com, tina.rode@enron.com,
                    marsha.schiller@enron.com, lillian.carroll@enron.com,
                    tonai.lehr@enron.com, nicole.mayer@enron.com,
                    darlene.forsyth@enron.com, janette.elbertson@enron.com,
                    angela.mcculloch@enron.com, pilar.cerezo@enron.com,
                    cherylene.westbrook@enron.com, shirley.tijerina@enron.com,
                    nicki.daw@enron.com, kimberly.hillis@enron.com,
                    patti.thompson@enron.com
X-From: Kay Chapman
X-To: Tim Belden, Kevin M Presto, John Arnold, Scott Neal, Hunter S Shively, Phillip K Allen, Janet R Dietrich, Chri
stopher F Calger, W David Duran, Raymond Bowen, Jeff Donahue, Brian Redmond, C John Thompson, Scott Jo
sey, Rob Milnthorp, Max Yzaguirre, Beth Perlman, Wes Colwell, Sally Beck, David Oxley, Joseph Deffner, Jordan
Mintz, Mark E Haedicke
X-cc: Mollie Gustafson, Felicia Doan, Ina Rangel, Kimberly Brown, Airam Arteaga, Christy Chapman, Tina Rode,
Marsha Schiller, Lillian Carroll, Tonai Lehr, Nicole Mayer, Darlene C Forsyth, Janette Elbertson, Angela McCulloc
h, Pilar Cerezo, Cherylene R Westbrook, Shirley Tijerina, Nicki Daw, Kimberly Hillis, Patti Thompson
X-bcc:
X-Folder: \David_Delainey_Dec2000\Notes Folders\'sent mail
X-Origin: Delainey-D
X-FileName: ddelain.nsf

This is a reminder !!!!!  There will be a Friday Meeting August 18, 2000.
This meeting replaces the every Friday Meeting and will be held every other
Friday.


  Date:  Friday August 18, 2000

  Time:  2:30 pm - 4:30 pm

  Location: 30C1

  Topic:   ENA Management Committee



If you have any questions or conflicts, please feel free to call me (3-0643)
or Bev (3-7857).

Thanks,

Kay 3-0643
```

Figure 3-7. Search result of “ENA Management Committee” in David Delaney’s dataset.

3.3 Statistic Measures and Histogram Views

3.3.1 Frequency

EmailTime calculates the *sent frequency (count)* and *received frequency (count)*. For sent frequency, it counts all the number of messages a given email address sent over a specified time period. In other words, it counts all the recipients of emails that a selected email address has sent over a specified time period. For received frequency, it counts all the number of messages a given email address received over a specified time period. The time period is automatically updated by the X-axis (Time-axis) control in the control panel.

As an example, we calculated the sent and received frequency for two email addresses in Figure 3-3 in Table 3-1 and Table 3-2. The tables show data over the entire timeline.

Table 3-1. Sent and received frequency for aaron@a.org in Figure 3-3.

Row #	Email Address	Sent Freq.	Received Freq.
1	aaron@a.org	1	1
2	beth@b.org	2	2
3	chris@c.org	1	1
4	david@d.org	1	1

From Row #1 of Table 3-1, Aaron sent an email to himself and received an email (as determined by the *Cc:* field) from himself. From Row #2 of Table 3-1, Aaron sent two emails to Beth and received two emails (as determined by the *To:* field) from her and so on.

Table 3-2. Sent and received frequency for beth@b.org in Figure 3-3.

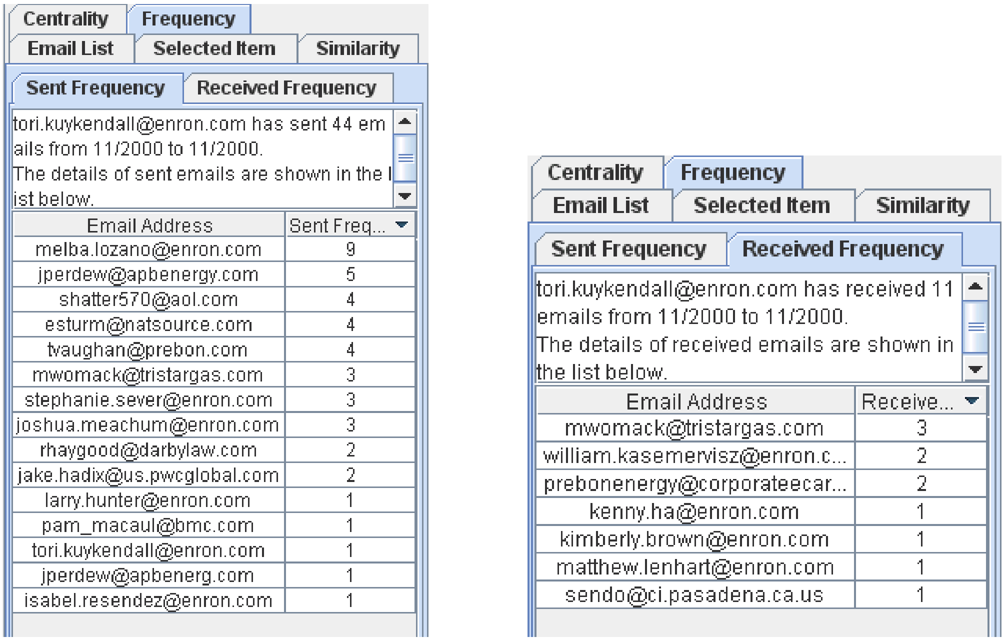
Row #	Email Address	Sent Freq.	Received Freq.
1	aaron@a.org	2	2
2	chris@c.org	1	1
3	david@d.org	1	1

Table 3-3. Total sent and received frequency for each email address in Figure 3-3.

Email Address	Aaron	Beth	Chris	David
Total Sent Frequency	5	4	2	2
Total Received Frequency	5	4	2	2

From Table 3-3, Aaron sent emails to 5 email addresses (not necessary different email addresses) and received emails from 5 email addresses (not necessary different email addresses) overall. Beth sent 4 emails and received 4 emails overall and so on.

Figure 3-8 displays the frequency for Tori Kuykendall (*Enron*'s Trader) in November 2000. His sent frequency is 44 during this month (See Figure 3-8-A) and his received frequency is 11 (See Figure 3-8-B).



tori.kuykendall@enron.com has sent 44 em ails from 11/2000 to 11/2000. The details of sent emails are shown in the l ist below.

Email Address	Sent Freq...
melba.lozano@enron.com	9
jperdew@apbenergy.com	5
shatter570@aol.com	4
esturm@natsource.com	4
tvaughan@prebon.com	4
mwomack@tristargas.com	3
stephanie.sever@enron.com	3
joshua.meachum@enron.com	3
rhaygood@darbylaw.com	2
jake.hadix@us.pwcglobal.com	2
larry.hunter@enron.com	1
pam_macaul@bmc.com	1
tori.kuykendall@enron.com	1
jperdew@apbenerg.com	1
isabel.resendez@enron.com	1

A.

tori.kuykendall@enron.com has received 11 emails from 11/2000 to 11/2000. The details of received emails are shown in the list below.

Email Address	Receive...
mwomack@tristargas.com	3
william.kasemervisz@enron.c...	2
prebonenergy@corporateecar...	2
kenny.ha@enron.com	1
kimberly.brown@enron.com	1
matthew.lenhart@enron.com	1
sendo@ci.pasadena.ca.us	1

B.

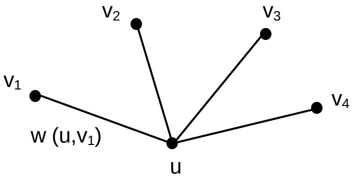
Figure 3-8. Displaying the frequency for Tori Kuykendall. A) Sent frequency. B) Received frequency.

3.3.2 Centrality

The idea behind calculating the centrality is to compare the activity level of members in a group of people and answering questions such as who is (are) the most active member(s) in a group of people based on the number of Sent-emails/To-emails/Cc-emails or what is the activity level of each person in a group of people in comparison to each other based on the number of Sent-emails/To-emails/Cc-emails.

In graph theory, the *degree centrality* of a node *u* is defined as the sum of weights of edges incident to *u*. Weight is the number of emails that two people send to each other or receive from each other. In the following equation, *(u, v_i)* is an edge from node u to node *v*. *$w\ (u, v_i)$* is the weight of that edge. *E* is the edge set.

$$C_p(u) = \sum_{(u,v_i)\in E} w(u,v_i).$$

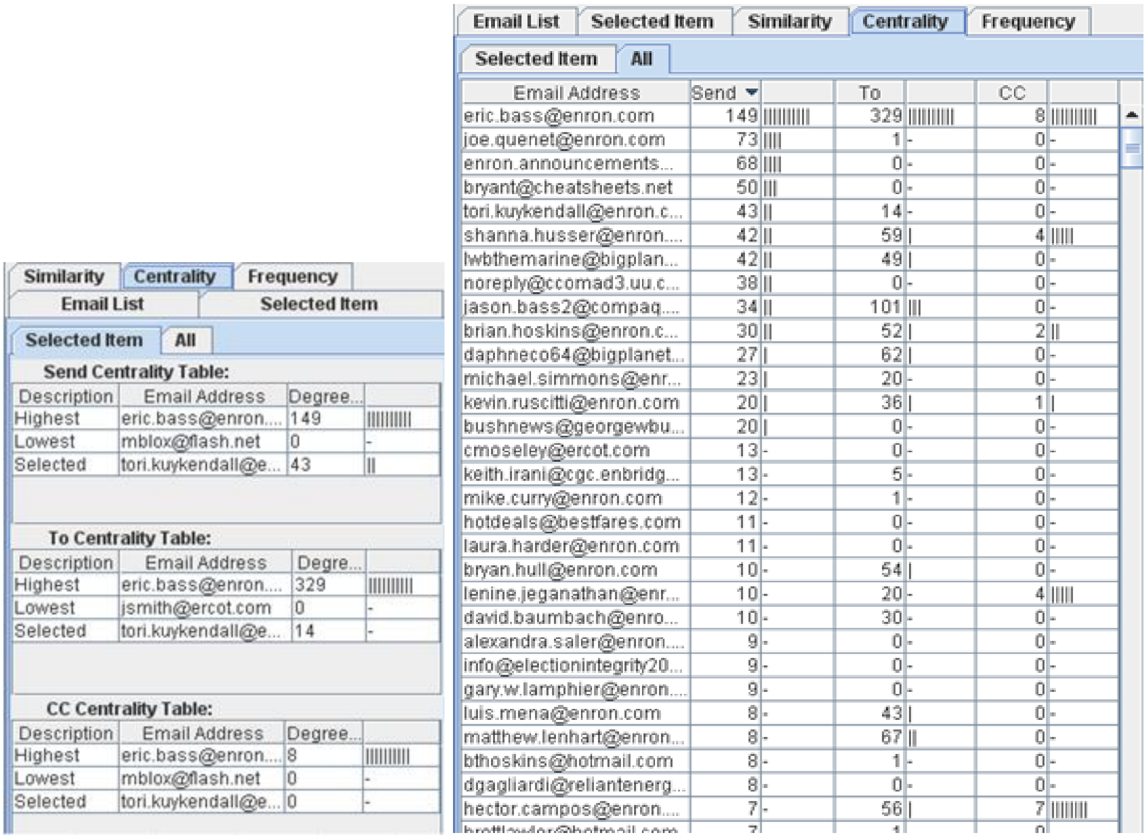


EmailTime specifies the *degree centrality* for an email address in the network. As our network is directed, *EmailTime* displays the three separate measures of the *degree centrality*; namely in-degree (*To centrality* and *Cc centrality*) of *To*, *Cc* emails and out-degree (Sent *centrality*) of sent emails for the owner of the selected message in the numerical and graphical representations. Table 3-4 displays the Sent centrality, To centrality and Cc centrality for each email address in Figure 3-3. Therefore Aaron and Beth are the most active participants in terms of sending message in this group. Aaron is the most active person in terms of receiving email as determined by the *To:* field and Beth is the most active one in terms of receiving the email as determined by the *Cc:* field. The table shows data over the entire time period.

Table 3-4. Sent, To and Cc centrality for each email address in Figure 3-3.

Email Address	Aaron	Beth	Chris	David
Sent Centrality	2	2	1	1
To Centrality	4	2	1	1
Cc Centrality	1	2	1	1

To have a better comparison, *EmailTime* displays the highest and the lowest in each centrality measures (Figure 3-9-A). It represents the three measures of *degree centrality* for all the email addresses in a separate list. Each column is sort-able by clicking on its header (Figure 3-9-B). Figure 3-9 displays the centrality for Tori Kuykendall.



Send Centrality Table:

Description	Email Address	Degree...	
Highest	eric.bass@enron....	149	||||||||||
Lowest	mblox@flash.net	0	-
Selected	tori.kuykendall@e...	43	||

To Centrality Table:

Description	Email Address	Degre...	
Highest	eric.bass@enron....	329	||||||||||
Lowest	jsmith@ercot.com	0	-
Selected	tori.kuykendall@e...	14	-

CC Centrality Table:

Description	Email Address	Degree...	
Highest	eric.bass@enron....	8	||||||||||
Lowest	mblox@flash.net	0	-
Selected	tori.kuykendall@e...	0	-

Email Address	Send ▾		To		CC	
eric.bass@enron.com	149	||||||||||	329	||||||||||	8	||||||||||
joe.quenet@enron.com	73	||||	1	-	0	-
enron.announcements...	68	||||	0	-	0	-
bryant@cheatsheets.net	50	|||	0	-	0	-
tori.kuykendall@enron.c...	43	||	14	-	0	-
shanna.husser@enron....	42	||	59	|	4	|||||
lwbthemarine@bigplan...	42	||	49	|	0	-
noreply@ccomad3.uu.c...	38	||	0	-	0	-
jason.bass2@compaq....	34	||	101	|||	0	-
brian.hoskins@enron.c...	30	||	52	|	2	||
daphneco64@bigplanet...	27	|	62	|	0	-
michael.simmons@enr...	23	|	20	-	0	-
kevin.ruscitti@enron.com	20	|	36	|	1	|
bushnews@georgewbu...	20	|	0	-	0	-
cmoseley@ercot.com	13	-	0	-	0	-
keith.irani@cgc.enbridg...	13	-	5	-	0	-
mike.curry@enron.com	12	-	1	-	0	-
hotdeals@bestfares.com	11	-	0	-	0	-
laura.harder@enron.com	11	-	0	-	0	-
bryan.hull@enron.com	10	-	54	|	0	-
lenine.jeganathan@enr...	10	-	20	-	4	|||||
david.baumbach@enro...	10	-	30	-	0	-
alexandra.saler@enron....	9	-	0	-	0	-
info@electionintegrity20...	9	-	0	-	0	-
gary.w.lamphier@enron....	9	-	0	-	0	-
luis.mena@enron.com	8	-	43	|	0	-
matthew.lenhart@enron...	8	-	67	||	0	-
bthoskins@hotmail.com	8	-	1	-	0	-
dgagliardi@reliantenerg...	8	-	0	-	0	-
hector.campos@enron....	7	-	56	|	7	||||||||
brettlawler@hotmail.com	7		1		0	

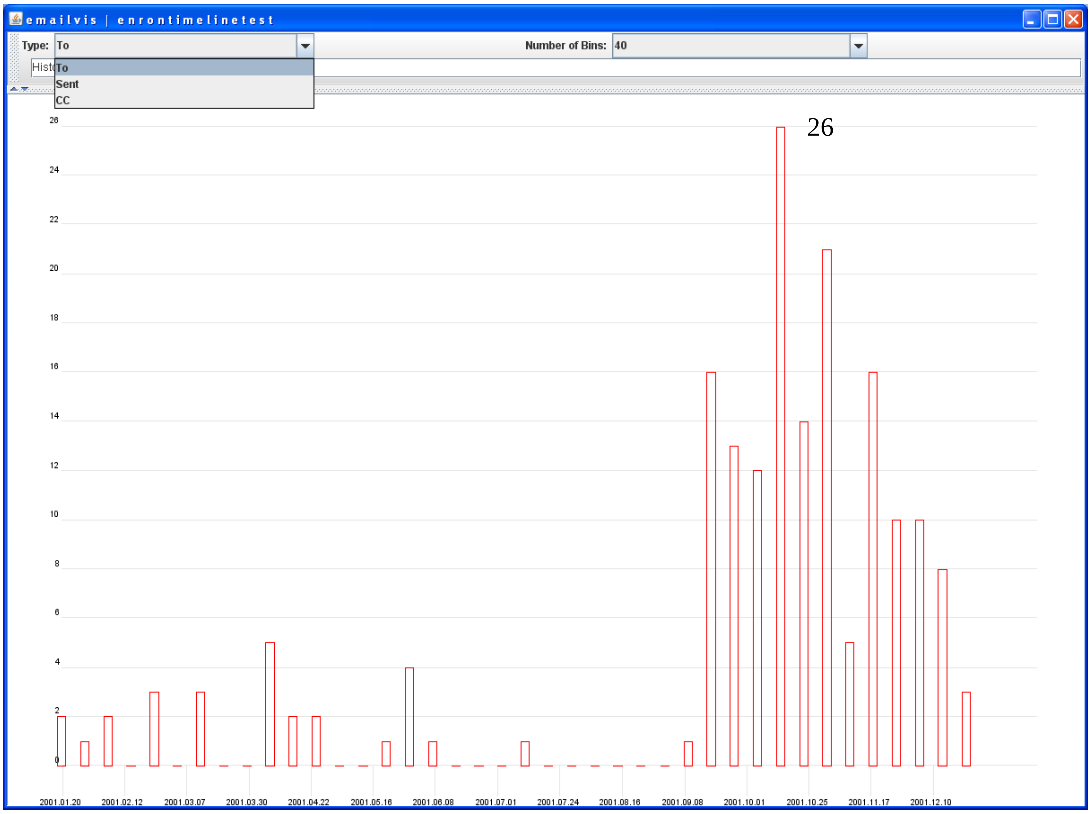
A. Tori Kuykendall.

B. All.

Figure 3-9. Displaying the centrality for Tori Kuykendall. A) Tori Kuykendall. B) All. Sorted based on Sent centrality.

3.3.3 Histogram View

We can create a histogram for each selected email address in the email list to display sent emails as determined by the *From*: field, received emails as determined by the *To:* field or received emails as determined by the *Cc:* field. In terms of its implementation, we used some parts of [32].



Type: To

Number of Bins: 40

Figure 3-10. Received emails as determined by the *To:* field for Eric Saibi (*Enron* Trader) with 40 intervals in 2001. There is a peak of received emails in October.

User can set the number of intervals to 3, 15, 45, 75 or 100. We provided a different range of numbers in order to enable the users to switch between small and large number of intervals based on their own preferences. The X-axis depicts time and user can set the Y-axis scale in order to display the number of sent and received emails (See Figure 3-10). It means that the whole time period is divided into the number of intervals (equal slices) and the number of sent or received emails in each time slice is calculated.

3.4 Discussion on System Capabilities

EmailTime visualization enables the analysts to infer underlying communication patterns by observation and interaction. Using the system analysts/users can:

- Compare different email addresses to each other with respect to the duration, activity level and role (sender, receiver, both) of each email address. (Some changes such as discovering the switches between email addresses)
- Recognize the most frequent correspondents of a given mailbox (person) and types of their correspondence (general or private message).
- Compare the network in different time periods with respect to the temporal gaps, crowded eras, number of senders, sent emails with large number of recipients.

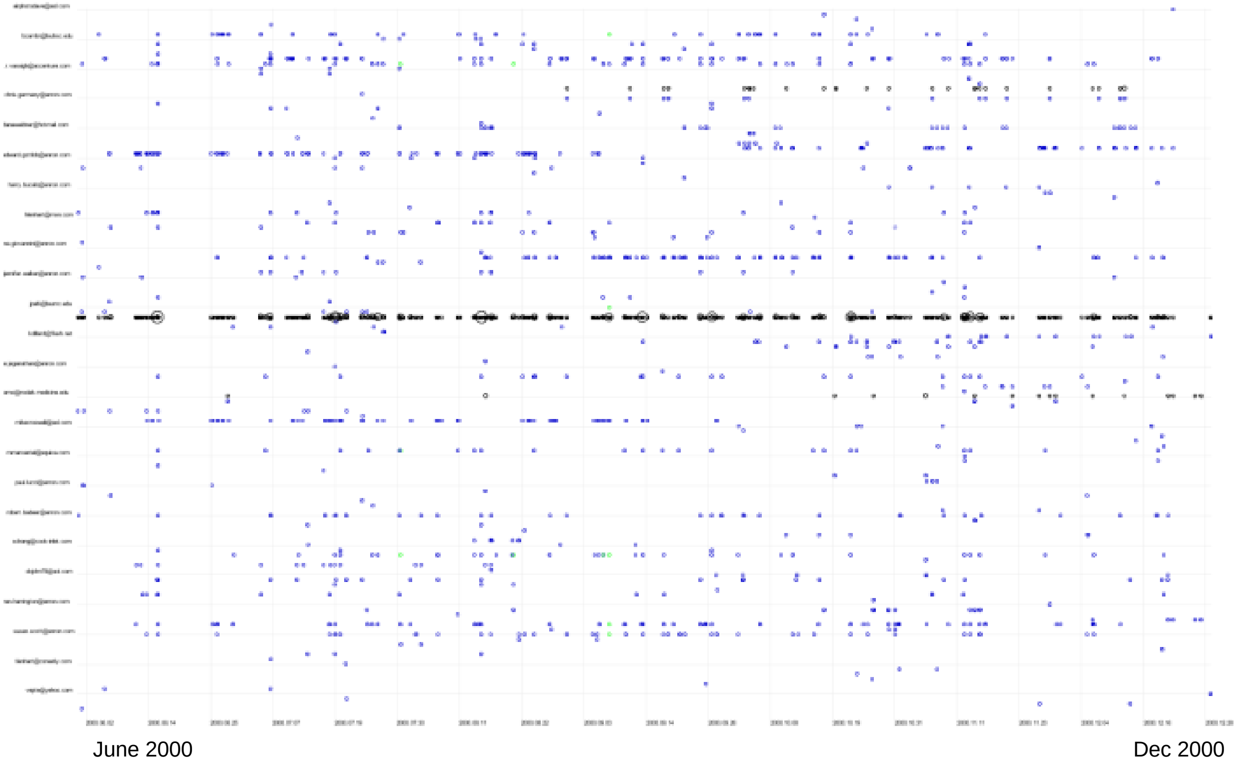
3.4.1 Examples

- *Compare the activity of one or more email addresses in the particular periods of time (See* Figure 3-11 (A, B, C and D)*).*
- *Temporal gaps in email addresses' activities are obvious. Therefore, if analysts/users know some events they can easily relate those gaps to the events (e.g. holidays, trips…, See* Figure 3-11 (A, B, C and D)*).*

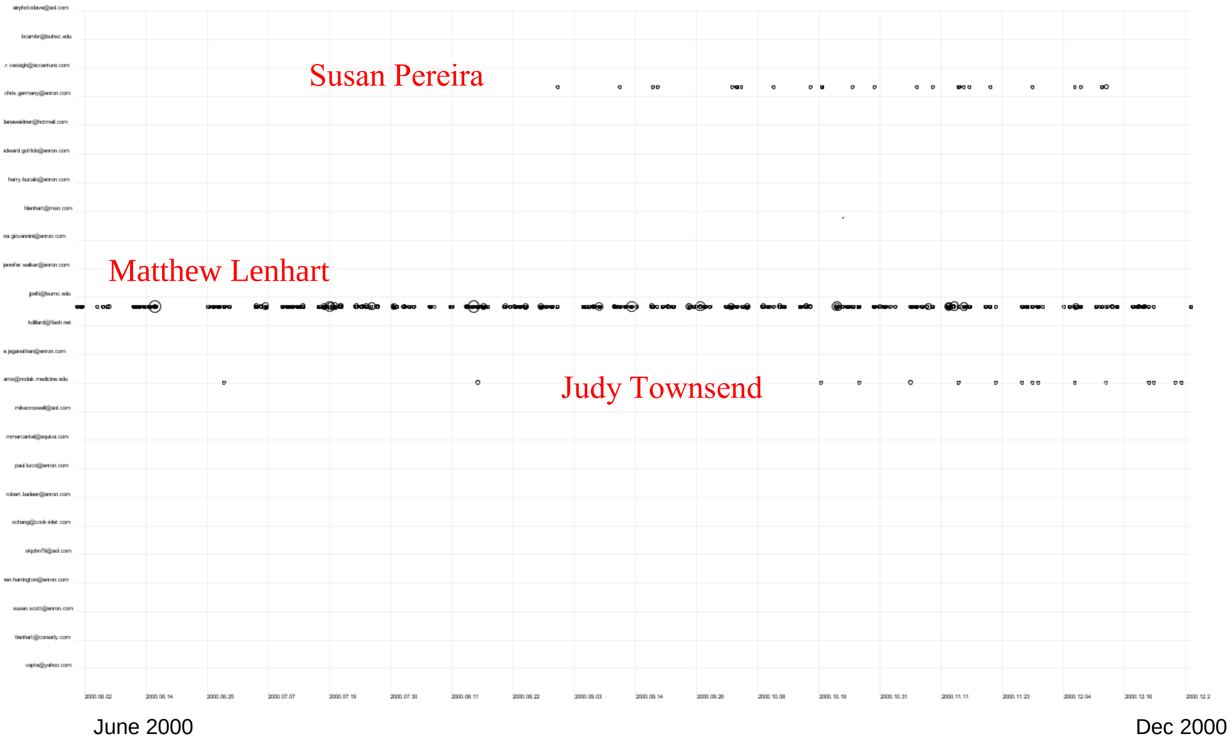
Figure 3-11 displays the datasets of four randomly selected *Enron* employees namely; Albert Meyers, Judy Townsend, Matthew Lenhart, Susan Pereira in two different time periods, June-Dec 2000 and June-Dec 2001. As it is shown these two time periods are completely different. The first time period (See

Figure 3-11-A) is more uniform than the second time period whereas the second one (See Figure 3-11-C) has the gap period, Large sent nodes and crowded part at the end of 2001 which can be the result of *Enron*'s fall.

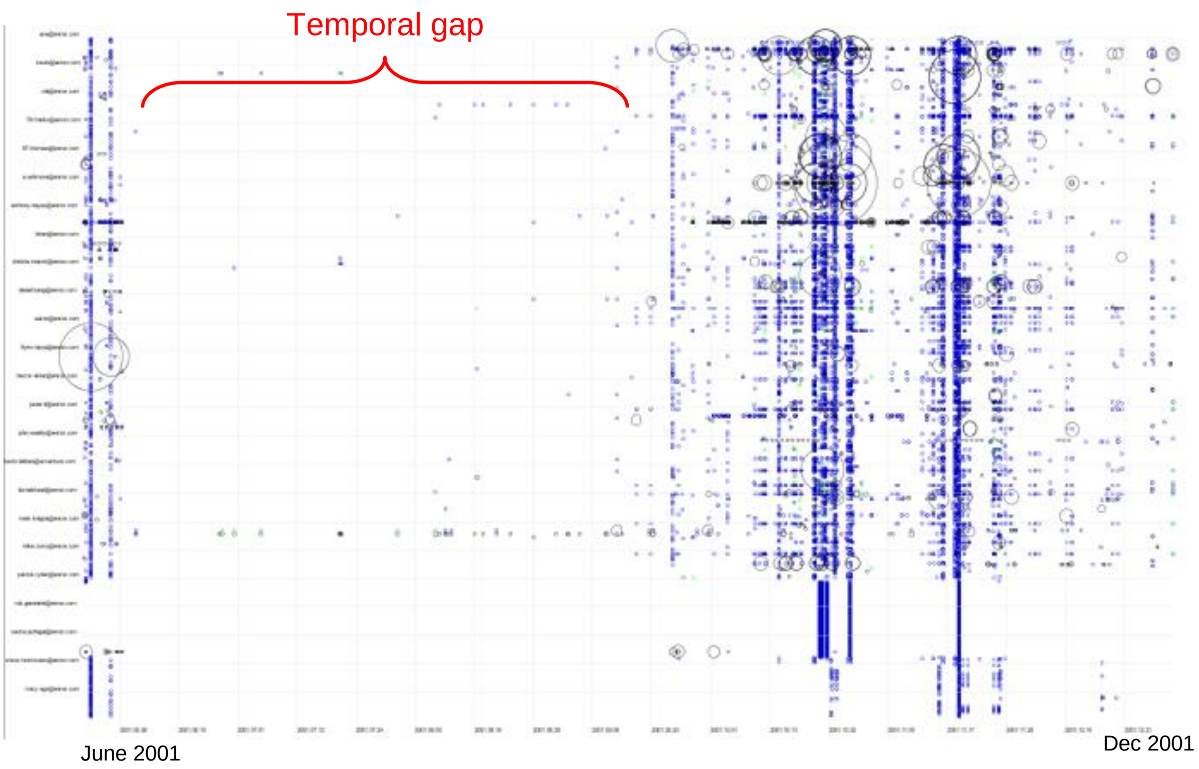
In Figure 3-11-B, when we filtered out received emails as determined by the *To:* and *Cc:* fields (blue and green nodes), we realized that in the first period only three employees sent emails with different activity levels (e.g. Matthew Lenhart was the most active sender among them whereas Albert Meyers had no activity at that time). In Figure 3-11-D, many others sent emails in the second period as well. Judy Townsend had no activity as a sender at that time.



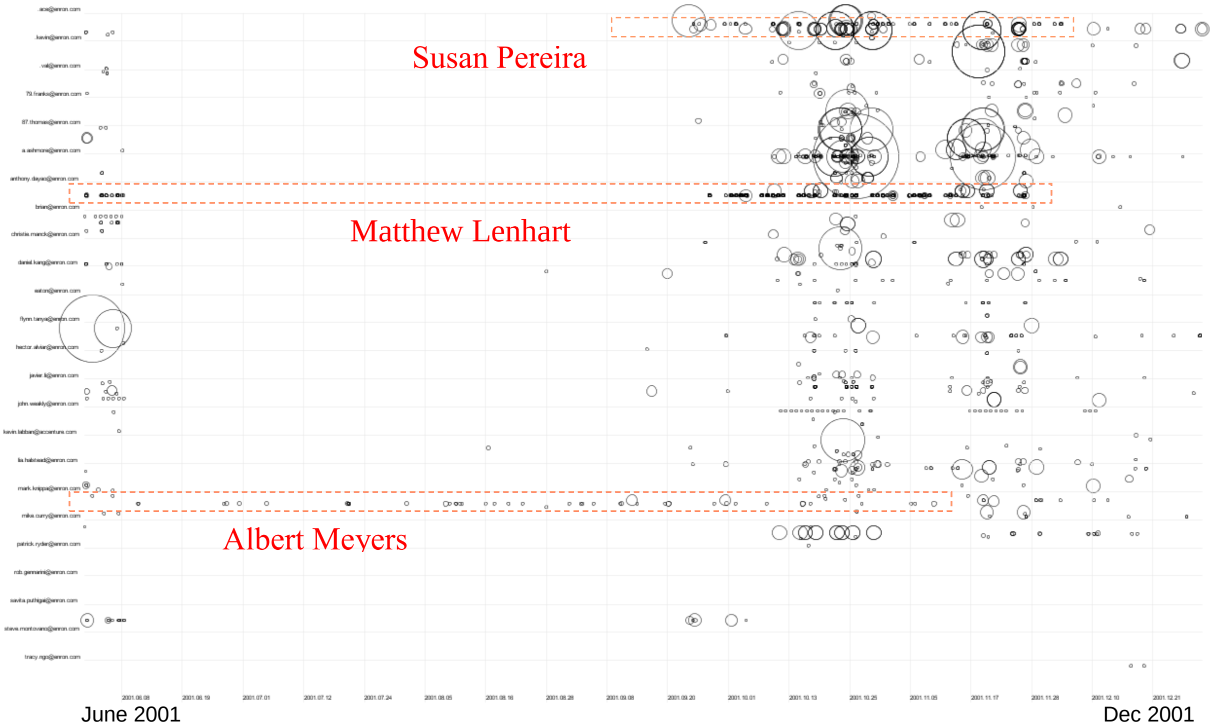
A. The datasets of four randomly selected *Enron* employees. First time period, June-Dec 2000; No filtering.



B. The datasets of four randomly selected *Enron* employees. First time period, June-Dec 2000; Filtered by the sent emails (black nodes).



C. The datasets of four randomly selected *Enron* employees. Second time period, June-Dec 2001; No filtering.



D. The datasets of four randomly selected *Enron* employees. Second time period, June-Dec 2001; Filtered by the sent emails (black nodes).

Figure 3-11. The datasets of four randomly selected *Enron* employees. A) June-Dec 2000; no filtering. B) June-Dec 2000; with filtering out the received emails. C) June-Dec 2001; no filtering. D) June-Dec 2001; with filtering out the received emails.

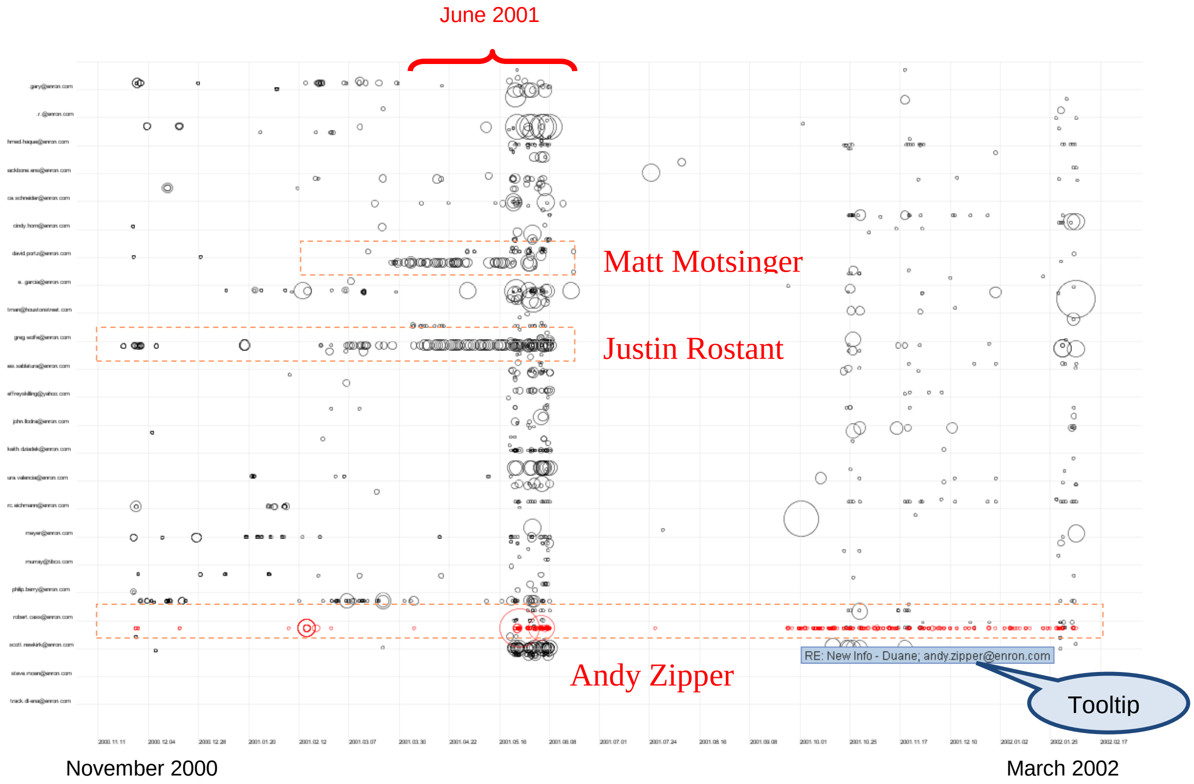
- *The most frequent correspondents who have sent/received a relatively large number of emails to/from the owner are easily distinguished – a row of black circles (sent emails) (See* Figure 3-12*).*

- *Type of correspondences (general or private messages regarding to the size of black circles; See* Figure 3-12*).*

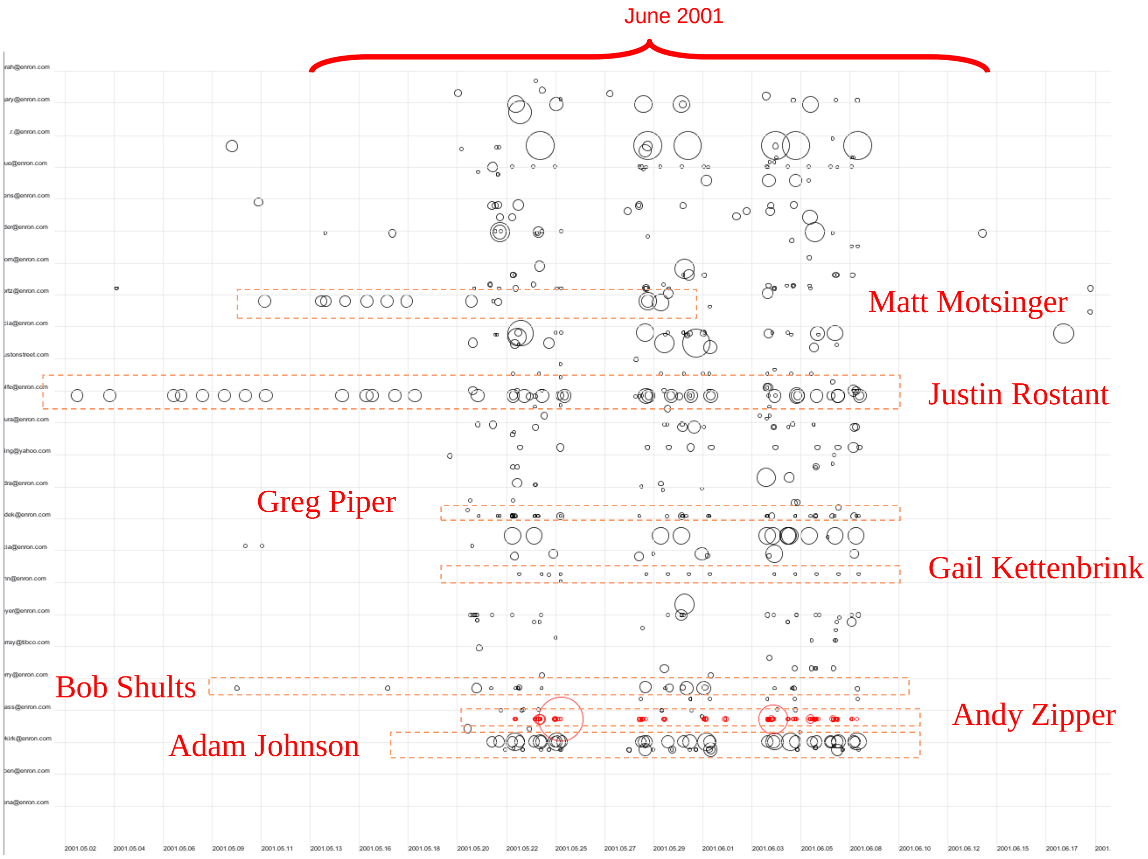
Figure 3-12 displays the zoomed-out and zoomed-in views of the sent messages discovered from the dataset of Andy Zipper (*Enron* vice president). His dataset contains more than 2000 messages from November 2000 to March 2002. In order to filter the plot to display only the sent messages, we selected the

"Sent" option from the *Node Type Selector* (See Figure 3-5-D) in the control panel. Therefore, we can identify who sent frequently emails to him in June 2001 in Figure 3-12. Each row represents the data for an email address. We specified his most frequent correspondents who sent several emails to him by the red dashed rectangles.

We can also recognize the types of his correspondents based on the size of the sent messages. For example in Figure 3-12 Justin Rostant sent general messages (with Medium number of recipients) to Andy whereas Greg Piper sent private messages (with Small number of recipients) to him.



A. Zoomed-out view.



B. Zoomed-in view.

Figure 3-12. Dataset of Andy Zipper (*Enron* vice president) filtered by the sent emails. A) Zoomed-out view. B) Zoomed-in view.

- *Compare different email addresses (user can select one or more email addresses from the email list to filter the plot in order to display their related emails, See* Figure 3-13*).*
- *Changes in activities such as switching from one email address to another one can be recognized (See* Figure 3-13*).*
- *Role (sender, receiver, both) of email address and other information such as duration, activity level of each email address can be recognized (See* Figure 3-13*).*

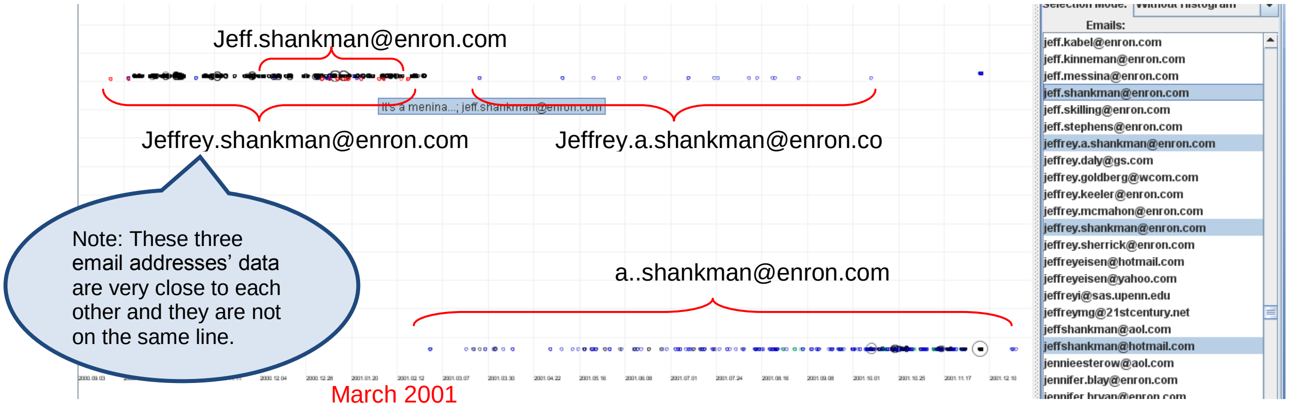


Figure 3-13. Emails sent or received by the five email addresses of Jeffrey Shankman (*Enron* president).

The *Email List Tab* in the control panel contains the list of all email addresses in the dataset. We can filter the plot visualization to only display the emails on the selected email addresses' lines. In Figure 3-13 we can see Jeffrey Shankman had multiple email addresses. His first email address was jeffrey.shankman@enron.com and switched to a..shankman@enron.com on March 2001. He didn't actively use other email addresses (Jeffrey.shankman@enron.com and Jeffrey.a.shankman@enron.com).

4: EXPERIMENT

4.1 Case Study

By exploring the *Enron* case study, we described *email behavior* through metrics using *EmailTime*. The metrics were:

- Number of sent emails as determined by *From:* field and received emails as determined by *To:* or *Cc:* fields (for *Form:*, *To:* and *Cc:* fields see Figure 3-1),
- Number of email addresses,
- Number of created folders (by the owner of the mailbox) and
- Recipient count of sent emails (determined by *To:* and *Cc:* fields of the email).

We present the mean and standard deviation of the mentioned metrics for each category in Appendix 2: Tables and Figures of Case Study. As you can see, the standard deviations for several categories are highly distributed (See Table 4-9).

4.1.1 Research Question

The general research question is what impact the executive positions (in this case *Enron* Corporation dataset) have on the email behaviour of people in an organization for the mentioned metrics.

4.1.2 Benchmark: *Enron* Email Dataset

Finding a real world email benchmark has been a challenge because of the private nature of the email data. Email datasets of individuals and organizations are good examples of this private type of data; therefore, we used the public *Enron* email archive which is a unique large dataset that contains around 517,431 emails [28]. The dataset used in this case study is selected from a two year time span between January 2000 and December 2001.

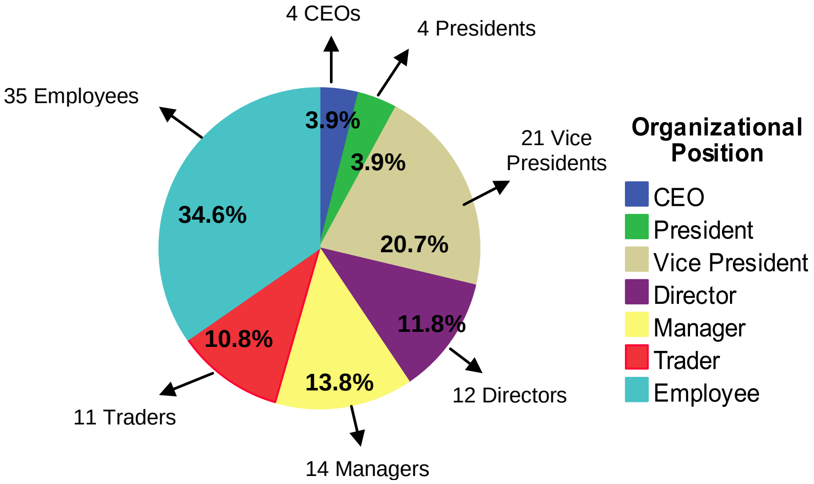


Figure 4-1. Composition of *Enron* organizational positions.

We grouped the email users in the archive into the seven categories identified in the public dataset [27], including CEO, President, Vice president, Director, Manager, Trader, and Employee (See Figure 4-1 for the number of workers in each category). We found the executive position of each *Enron* worker (in the public dataset [27]) using [30]. In the next section, we detail the differences of email behaviors within organizational positions.

This study is a between subject design as each subject participates in one and only one category. The number of subjects in each category is different.

4.1.3 Methodology, Analysis and Results

For the data collection and data analysis of case study, we used quantitative research methods (e.g. we measured mean, standard deviation, ANOVA, Post-hoc, Tukey and Games-Howell) as we gathered and investigated the quantitative properties of *Enron* email dataset. In the data collection, the number of sent emails as determined by *From:* field, number of received emails as determined by *To:* field, number of received emails as determined by *Cc:* field and, recipient count of the sent emails were provided by *EmailTime*. The number of the email addresses and number of the created folders were counted manually. For the data analysis, the statistical tool SPSS [31] was used to analyze the numerical results. We performed one-factor ANOVA and Post-hoc analysis (Tukey and Games-Howell test). The independent variable was the organizational position, which has seven levels (CEO, President, Vice President, Manager, Director, Employee and Trader). The dependent variables were:

- Number of sent and received emails,
- Contribution Index [5] (CI: number of sent emails as determined by *From:* field minus number of received emails divided by the total number of emails, To-CI: number of sent emails as determined by *From:* field minus number of received emails as determined by *To:* field divided by the total number of emails, Cc-CI: number of sent emails as determined by *From:* field minus

number of received emails as determined by *Cc:* field divided by the total number of emails),

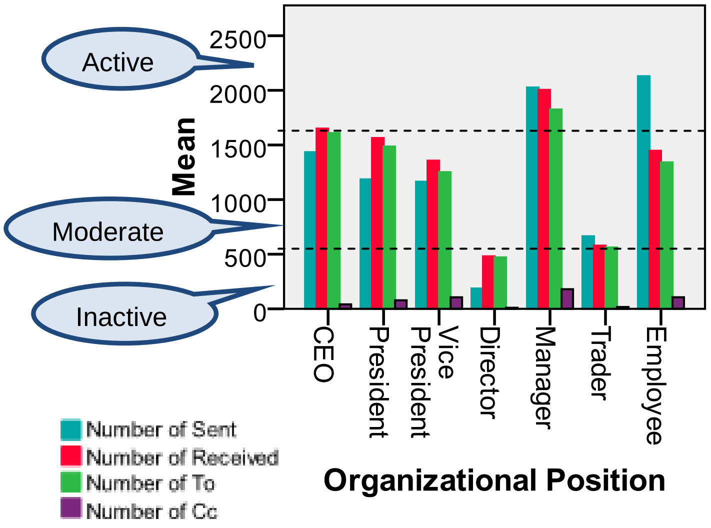
- Count of sent emails with single recipient, Small (2-9 recipients) number of recipients, Medium (10-29 recipients) number of recipients, and Large (30 and up recipients) number of recipients,
- Number of email addresses.

4.1.3.1 Analysis of the Activity Level

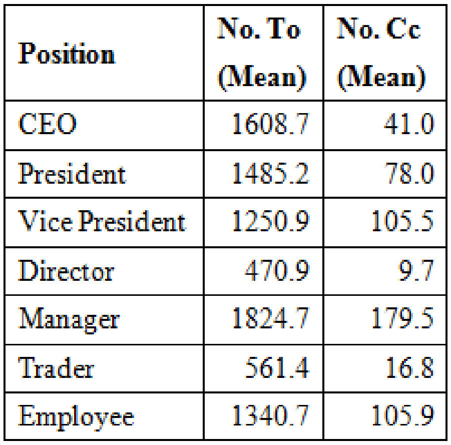
In this section, we are interested in finding out which category (organizational position) tends to be active. From comparing activity levels of organizational positions, some categories behaved similarly in terms of the number of sent and received emails. From Figure 4-2-A, based solely on observation, we recognized three categories of activity and divided them into Inactive, Moderate and Active (See Table 4-1). It graphically specifies the three categories as indicated by the dashed lines. Results identified Managers and Employees were Active, Traders and Directors were Inactive, and the rest were Moderate. Employees have the highest average number of sent emails, while Managers have the highest average number of received emails. These findings may be the result of the nature of their positions.

Table 4-1. Classification of organizational positions.

No. of Emails	Inactive	Moderate	Active
Exchanged (Sent + Received)	[<1500] Directors, Traders	[1500>-<3500] CEOs, Presidents, Vice Presidents, Employees	[>3500] Managers
Sent	[<1000] Directors, Traders	[1000>-<1500] CEOs, Presidents, Vice Presidents	[>1500] Employees, Managers
Received	[<1000] Directors, Traders	[1000>-<2000] CEOs, President, Vice Presidents, Employees	[>2000] Managers



A.



Position	No. To (Mean)	No. Cc (Mean)
CEO	1608.7	41.0
President	1485.2	78.0
Vice President	1250.9	105.5
Director	470.9	9.7
Manager	1824.7	179.5
Trader	561.4	16.8
Employee	1340.7	105.9

B.

Figure 4-2. Results of exchanged emails. A) Average number of sent and received (as determined by the *Cc:* and *To:* fields) emails for *Enron* organizational positions from 2000 to 2001. B) Average number of received emails as determined by the *To:* and *Cc:* fields for each organizational positions.

4.1.3.1.1 Results of ANOVA and Post-hoc Test

We performed one-factor ANOVA (See Table 4-2 (for the actual numbers) and Table 4-3 (for the normalized numbers)), where the dependent variables were actual and normalized number of sent emails as determined by *From*: field and received emails as determined by *To:* and *Cc:* fields. The independent variable was the organizational position with seven levels (CEO, President, Vice President, Manager, Director, Trader and Employee).

For the normalized (ratio) numbers, we calculated the percentage of each type of message (*From*, *To* and *Cc*) with respect to the total number of emails associated with the *Enron* worker (1). We did this for each *Enron* worker. For example the “normalized number of sent emails” is:

$$\frac{\text{count of sent emails}}{\text{total of emails}} * 100 \tag{1}$$

Then we calculated the average of each category (organizational position). No significant difference is found in the ANOVA results for the actual and the normalized numbers (See Table 4-2 and Table 4-3).

From the Post-hoc analysis on the normalized numbers, there are significant differences between:

- Employee and Director in the number of sent emails as determined by *From*: field and received emails as determined by *To:* field,
- Vice President and Director in the number of received emails as determined by *To:* field.

From the Post-hoc analysis on the actual numbers, there is a significant difference in the number of emails sent by Vice President and Director.

Table 4-2. Results of ANOVA on the actual number of sent and received emails.

Dependent Variable	F(6,94)	p	η2	power
Number of Sent	.883	.511	.053	.333
Number of Received (To + Cc)	.851	.534	.052	.322
Number of Received (To)	.85	.535	.051	.321
Number of Received (Cc)	.864	.526	.052	.326

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

Table 4-3. Results of ANOVA on the normalized number of sent and received emails.

Dependent Variable	F(6,94)	p	η2	power
Normalized Number of Sent	1.476	.195	.086	.548
Normalized Number of Received (To + Cc)	1.624	.149	.094	.677
Normalized Number of Received (To)	1.898	.089	.108	.194
Normalized Number of Received (Cc)	.501	.806	.031	.596

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

4.1.3.2 Analysis of the Roles

Another approach to interpret Figure 4-2 is to compare the number of sent and received emails to determine the role (sender, receiver or both) of email address for different organizational positions. We were interested to discover if certain categories tend to be a specific role. Table 4-4 specifies the results.

Table 4-4. Organizational positions' roles (sender, receiver or both).

Role	Organizational Position
Receiver (#Sent < #Received)	President, Director, CEO and Vice President
Both (#Sent = #Received)	Manager
Sender (#Sent > #Received)	Employee and Trader

Email addresses have different roles for different periods of time. *Gloor* et al. [5] defined Contribution Index (CI) to specify the role of email addresses (2).

$$\frac{\text{emails sent} - \text{emails received}}{\text{total of emails sent and received}} \tag{2}$$

This index is near to –1 for the receivers and +1 for the senders. We expanded this formula to To-CI and Cc-CI to see the impact of *To:* and *Cc:* fields separately. To-CI (3) is the number of sent emails as determined by *From:* field minus the number of received emails as determined by *To:* field divided by the total number of sent and received emails as determined by *To:* field.

$$\frac{\text{emails sent} - \text{emails received as determined by } \textit{To:} \text{ field}}{\text{total of sent and received emails as determined by } \textit{To:} \text{ field}} \tag{3}$$

Cc-CI (4) is the number of sent emails as determined by *From:* field minus the number of received emails as determined by *Cc:* field divided by the total number of sent and received emails as determined by *Cc:* field.

$$\frac{\text{emails sent} - \text{emails received as determined by } \textit{Cc:} \text{ field}}{\text{total of sent and received emails as determined by } \textit{Cc:} \text{ field}} \tag{4}$$

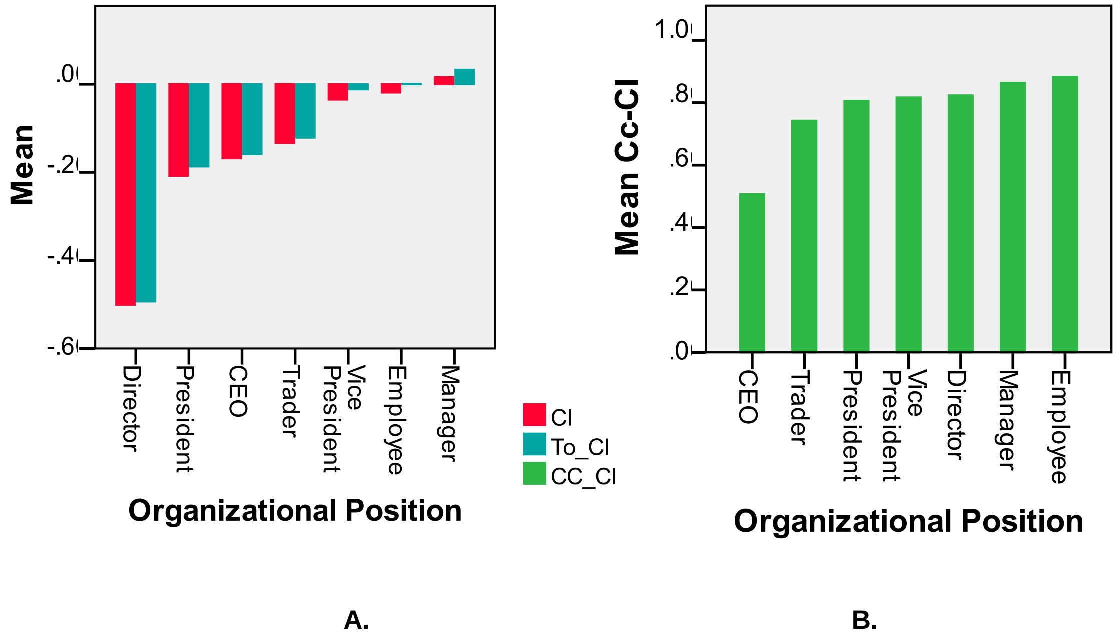


Figure 4-3. Contribution Index (CI) for organizational positions. A) Average of CI and To-CI. B) Average of Cc-CI.

Figure 4-3 shows that the average Contribution Index of administrative groups (such as CEO, President and Vice President) tends to be lower than staff in lower positions. It may be the result of the nature of their positions (e.g. employees ask and report whereas administrators are reported to and make orders). In addition, CI and To-CI follow a same trend while Cc-CI has a different trend. Figure 4-3-A shows that CIs of Employees, Traders and Managers are near zero, which means that they had same amount of sent and received emails on average. Then there is a jump to Trader, CEO, and President where the three groups behaved as weak receivers. Finally, there is a jump to Director where the executive officers behaved as strong receiver.

On the other hand, Figure 4-3-B shows Cc-CI for organizational positions where most of them except CEO were near +1. This means that they had received few emails as determined by *Cc*: field.

4.1.3.2.1 Results of ANOVA and Post-hoc Test

We performed one-factor ANOVA (See Table 4-5), where the dependent variables were CI, To-CI and Cc-CI. The independent variable was the organizational position with seven levels (CEO, President, Vice President, Manager, Director, Trader and Employee).

No significant difference is found in the ANOVA results in CI, To-CI and Cc-CI. From the Post-hoc analysis, there is a significant difference between Employee and Director in To-CI and CI.

Table 4-5. Results of ANOVA on CI, to-CI and Cc-CI.

Dependent Variable	F(6,94)	p	η2	power
CI	1.676	.135	.097	.613
To-CI	1.761	.116	.101	.6
Cc-CI	1.351	.243	.079	.505

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

4.1.3.3 Analysis of the Recipient Count of Sent Emails

We divided the sent emails into four categories based on the number of recipients:

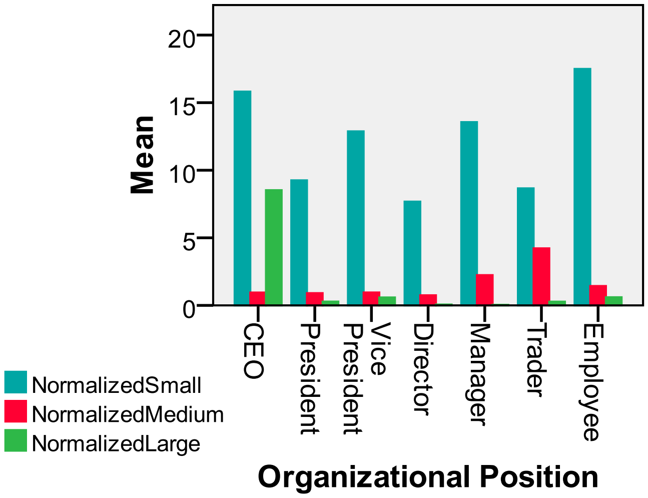
- Single recipient (including only one recipient),

- *Small* number of recipients (2-9 recipients),
- *Medium* number of recipients (10-29 recipients) and,
- *Large* number of recipients (30 and up recipients).

We then calculated the normalized number of sent emails with Small, Medium and Large number of recipients for each *Enron* worker. For normalized (ratio) numbers, we calculated the percentage of each type of sent message (with Small, Medium and Large number of recipients) with respect to the total number of sent emails associated with the *Enron* worker (5). We did this for each *Enron* worker. For example the “normalized number of sent emails with small number of recipients” is:

$$\frac{\text{count of sent emails with Small number of recipients}}{\text{total number of sent emails}} * 100 \qquad (5)$$

Then we calculated the average of each category (organizational position) (See Figure 4-4). Comparing the actual and normalized graph, we figured out the normalized graph contribute more in terms of presenting the habits of different positions in sending emails to group of people. It showed a statistically significant difference (from the graph and ANOVA results) between CEOs and other groups in sending emails with Large number of recipients. Traders and then Managers sent emails with Medium number of recipients more than any other groups.



A.

Organizational Position	Node Size	
	Med	Large
CEO	82	24
President	30	8
Vice President	207	44
Director	16	4
Manager	128	6
Trader	323	3
Employee	461	97

B.

Figure 4-4. Results from recipient count of sent email. A) Average normalized number for sent emails with Small, Medium and Large number of recipients for each organizational position. B) Average number of sent emails with Medium and Large number of recipients for each organizational position.

4.1.3.3.1 Results of ANOVA and Post-hoc Test

We performed one-factor ANOVA (See Table 4-6 (for the actual numbers) and Table 4-7 (for the normalized numbers)), where the dependent variables were actual and normalized number of sent emails with *Small* number of recipients, Medium number of recipients and Large number of recipients. The independent variable was the organizational position with seven levels (CEO, President, Vice President, Manager, Director, Trader and Employee).

For the actual numbers, there is no significant difference in the ANOVA results. For the normalized number there is a significant difference between CEO and other groups in sending emails with Large number of recipients (See Table 4-6 and Table 4-7).

From the Post-hoc analysis, there are significant differences between:

- Employee and Director in sending emails with Small number of recipients,
- Employee and Trader in sending emails with Small number of recipients.

Table 4-6. Results of ANOVA on the number of sent emails with Small number of recipients, Medium number of recipients and Large number of recipients.

Dependent Variable	F(6,86)	p	η2	power
Number of Small	.958	.459	.063	.359
Number of Medium	.679	.667	.045	.256
Number of Large	.678	.668	.045	.256

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

Table 4-7. Results of ANOVA on the normalized number of sent emails with Small number of recipients, Medium number of recipients and Large number of recipients.

Dependent Variable	F(6,86)	p	η2	power
Normalized Number of Small	2.830	**.015**	.165	.864
Normalized Number of Medium	.954	.461	.062	.358
Normalized Number of Large	6.779	**.000**	.321	.999

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

4.1.4 Conclusion and Discussion

We presented a case study to analyze the activity level, type and recipient count of sent emails by *Enron* workers between January 2000 and December 2001. We found some groups behaved similarly on average and grouped them into three categories of Inactive, Moderate and Active. Managers and Employees were Active, Traders and Directors were Inactive, and the rest were Moderate.

In addition, administrative groups (such as CEO, President and Vice President) tend to be receivers more often than the staff in lower positions. Analysis on the recipient count of sent emails shows a statistically significant difference between CEOs and other groups in sending emails with Large number of recipients. Traders and then Managers sent emails with Medium number of recipients more than any other groups.

According to the results, since no relationship between the number of created folders and organizational positions was found in our dataset, we believe a user's choice in the number of created folders is subjective.

We also performed one-factor ANOVA (See Table 4-8), where the dependent variable was the number of email addresses. The independent variable was the organizational position with seven levels. There was a significant difference for number of Email Address between CEO and other groups, F(6,94) = 2.67, p < .05. 80% of the cases had the number of email addresses (with *Enron* domain) within the range of 2 to 6.

Table 4-8. Results of ANOVA on the number of email address.

Dependent Variable	F(6,94)	p	η2	power
Number of Email Address	2.673	**.019**	.146	.843

*P < .05. Significant effects are in bold.
Power is the ability to detect an effect (ranges: 0-1 where .95 means a 5% chance of failing to detect an effect that is there.)
Partial-Eta-squared (η2) is the proportion of total variability attributable to a factor.

4.1.4.1 Challenges

One of the challenges in this case study was the data was highly distributed (See Table 4-9 and Figure 4-5).

In terms of suggestion for the solution, we could:

- Exclude the extremes ones in each category,
- Exclude highly distributed categories like employee,
- Test it with other statistical tool (e.g. JMP) and so on.

Table 4-9. Mean and Standard deviation of numbers of sent and received emails. It shows that data is highly distributed.

Organizational Position	Count	No. Emails (Sent+Received)		No. Sent		No. Received (To+Cc)	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
CEO	4	3083.5	725.3	1433.7	1595.6	1649.7	1184.3
Director	12	666.92	348.4	186.2	174.9	480.6	274.7
Employee	35	3161.2	4593.8	2129.4	4176.5	1446.7	1878.6
Manager	14	4030.7	7802.1	2026.4	4014.5	2004.3	3968.6
President	4	2748.2	1310.6	1185.0	1167.9	1563.2	1148.3
Trader	11	1242.6	1868.8	664.3	1448.1	578.2	641.2
Vice President	21	2520.2	2734.9	1163.7	1222.6	1356.5	1906.7
Total	101	2623.7	4251.3	1458.9	3023.7	1308.5	2080.5

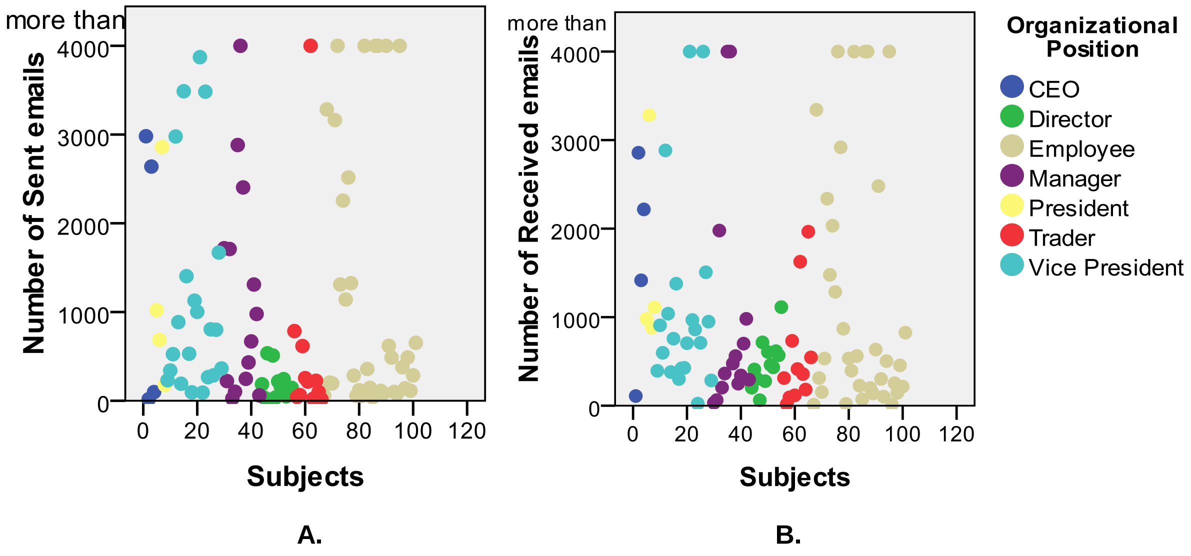


Figure 4-5. A. Number of sent emails. B. Number of received emails for Enron worker from Jan 2000 to Dec 2001. Colour specifies the *Enron* organizational positions.

4.2 Usability Testing

We conducted three separate studies. Each study consists of four or five scenarios (e.g. a scenario about time comparison or email address comparison). In each scenario users completed a number of tasks. Studies are:

- Pilot Study I (See section 4.2.1), to test the system and find the possible problems with the tasks in each scenario, pre and post questionnaires.

- Pilot Study II (See section 4.2.2), to find the strengths and weaknesses of the visualization and the control panel and to test the capabilities of *EmailTime*'s visualization in allowing participants to complete a number of scenarios.

- User Study (See section 4.2.3), is similar to pilot study II but it was more specific and detailed, in an effort to find out whether users are capable of accomplishing the scenarios in a specified time period using the system's capabilities.

The consent form and pre-questionnaire were the same for all the studies. Tasks were improved and became more specific after each study according to the user's comments and our observations. The post-questionnaire was changed for each study based on the user's comments and the purpose of study. The consent form, pre- and post-questionnaires for Pilot Study II and tasks for User Study are attached in Appendix 3: User Study Documents.

The lessons we learned and comments we received during the first two pilot studies helped us in redesigning the tasks for User Study and made the tasks clear. In order to quantify the results of tasks, we timed users and had written answers for User Study. According to Table 4-10 and Table 4-11, the tasks became more specific and clear after each pilot study. We decided to reduce the number of scenarios in order to prevent overloading users with too much information as learned in the first pilot study.

4.2.1 Pilot Study I

The aim of this pilot study was to test the quality of the tasks, pre and post questionnaires. After each testing session, we revised and edited the tasks and questionnaires based on the participant's performance and/or comments.

4.2.1.1 Participants

The pilot study's testing sessions for *EmailTime* visualization took place on July 27th and 28th, 2010 with four subjects (2 females and 2 males, n=4).

Participants were SIAT graduate students from SFU. They were between 24 and 25 years of age and none of them were color-deficient. They were familiar with the concept of visual analytics and visualization and had the experience of running user studies, analyzing results and developing visualization systems.

4.2.1.2 Procedure and Scenarios

Experiments were held in a quiet lab environment. Each participant had between one to two hours to perform the experiment as they didn’t complete the same number of scenarios. We started this pilot study with the scenarios in Table 4-10.

At the beginning of each experiment the participants were asked to sign a consent form and fill out a pre-questionnaire. Then we introduced them to the *EmailTime* visualization. In the introduction, we explained the plot visualization and the control panel using the dataset of *Enron* Trader, “Eric Bass”. Next, they were asked to perform some scenarios (each scenario is a number of tasks, see Table 4-10). Finally, they filled out a post-questionnaire.

Tasks and questionnaires had small changes after each testing session based on the user’s comments and our observation. Two participants completed five scenarios and two others accomplished four scenarios as we removed the fifth scenario (system’s functionalities) from the study due to overloading of participants with too much information. Think Aloud Protocol was used and participants explained their answers to the observer (me). Therefore, there was no written answer.

Table 4-10. Scenarios (Tasks) that we started with in Pilot Study I.

Scenario	Tasks
Scenario 1: Search Scenario	*i.* This is the dataset of David Delaney (*Enron* CEO), *ii.* Search for management Committee called ENA in his dataset, *iii.* Find out who sent those emails and why, *iv.* When and where is the meeting, *v.* Who are in the meeting, *vi.* How frequent is the meeting.
Scenario 2: Most Frequent Correspondents Scenario	i. This is the dataset of Andy Zipper (*Enron* vice president), ii. Find out who are his most frequent correspondents in June 2001, iii. Types of their correspondences (private/general) regarding to the size of exchanged messages.
Scenario 3: Email Comparison Scenario	i. This is the dataset of Jeffrey Shankman (*Enron* president), ii. Select the emails of following email addresses: 1) Jeffrey.shankman@enron.com, 2) a..shankman@enron.com, 3) Jeffry.a. shankman@enron.com iii. Find out the type (Sender/Receiver) and duration of each email address, iv. Is there any switching between email addresses and when.
Scenario 4: Time Comparison Scenario	i. This is the datasets of four *Enron* Employees (Albert Meyers, Judy Townsend, Matthew Lenhart and Susan Pereira), ii. Compare the dataset in June-Dec 2000 with June-Dec 2001, iii. What is the difference between these two time periods in terms of gaps and network mess.
Scenario 5: Functionalities Scenario	i. This is the datasets of 11 *Enron* traders, ii. Who are the most active ones in terms of sending and receiving email as *To* and *Cc*, iii. Compare the histogram of Eric Bass and Eric Saibi, iv. How many emails Eric Bass sent and received in Nov 2000.

4.2.1.3 Results and Discussion

Some changes are as follow for the rest of the studies:

- As participants were overloaded with too much information they received in the introduction, we decided to only focus on the visualization part and remove the scenarios related to the system's functionalities (Frequency, Centrality and histograms).
- We reordered the four scenarios from easy to difficult based on the users' comments.
- We added more meaningful questions to the questionnaire (e.g. questions about whether they are able to recognize similar scenarios using *EmailTime*).

Based on the results of Pilot Study I, we decided to focus only on capabilities of the visualization plot and have an hour testing session for the next studies.

4.2.2 Pilot Study II

The purpose of this pilot study was to find out strengths and weaknesses of the visualization and the control panel and whether users are able to recognize the capabilities of *EmailTime* through four scenarios. The focus was on investigating the dataset of individuals.

4.2.2.1 Research Question

We hypothesized that *EmailTime* visualization enables the analyst/user to:

1. Compare different time periods to each other and recognize their differences with respect to the crowded eras, large gaps (no activity), large emails, etc.
2. Find the most frequent correspondents of a person and type of their correspondences (private or general messages based on the size of sent messages).
3. Compare different email addresses of one (or more) person to each other according to the role (sender, receiver or both) of email address, duration and activity level of each email address and discover the switches between email addresses.
4. Find an event using the filters (e.g. a biweekly meeting)**.**

4.2.2.2 Participants

The pilot study experiments for *EmailTime* visualization took place on July 29th, 30th and 31st, 2010 with six subjects (2 females and 4 males, n=6). Participants were SIAT graduate students from SFU. They were between 20 and 29 years of age and none of them were color-deficient. Most of them were a little familiar with the concept of visual analytics and visualization and had the experience of working with software with the purpose of visualization such as Tableau, Inspire, graphs of Excel, etc. All of them were used to working with the computer and had more than one email address to which they logged-in more than once a day or always.

4.2.2.3 Method and Scenarios

Experiments were held in a quiet lab environment. All the participants completed the same tasks. Each participant had 60 minutes to complete the experiment. Data gathering was in the form of Think Aloud Protocol, written answers in questionnaires and our observation notes. There was no written answer in the task form.

At the beginning of each experiment they were asked to sign a Consent Form and fill a Pre-Questionnaire Form (See Appendix 3: User Study Documents) which was about their demography, possible color deficiency, computer skill, familiarity with the visual analysis concept and experience with the visualization tools, etc. Then we introduced the subjects with the *EmailTime* visualization. In the introduction, we explained the plot visualization and part of the control panel using the dataset of *Enron* Trader, "Eric Bass". We let the participants play with the system before starting the tasks and ensure they knew:

- The meaning of the plot, a node, the axes and the size of a black node.
- The difference between the black, blue and green nodes
- Roll over the nodes and find the subject and sender of that node.
- Distinguish a row by highlighting it.
- How to pan and zoom in/out the plot.
- How to change the X and Y axes.
- How to search for an email with a specific subject.

- Limit the visibility to Sent, To, Cc or any combination of those.
- How to work with the Email List and select multiple email addresses.
- How to see the content of an email in the plot.

Next, they were asked to accomplish four scenarios (See Table 4-11). Questions were answered through the whole testing sessions. In each scenario we concentrated on one capability of the system when investigating the datasets of *Enron* individuals namely; Matthew Lenhart (*Enron* employee), Andy Zipper (*Enron* vice president), Jeffery Shankman (*Enron* president) and David Delaney (*Enron* CEO).

Finally, they filled out a Post-Questionnaire Form (in Appendix 3: User Study Documents) about the difficulty level of each task, likes (useful and convenient aspects) and dislikes (troublesome and confusing aspects) of the visualization plot and control panel and whether they are able to recognize similar scenarios.

Table 4-11 contains the scenarios that users accomplished in this pilot study. For User Study, scenarios were the same but we made the tasks of each scenario more specific (See Appendix 3: User Study Documents for the tasks of User Study).

Table 4-11. Scenarios (and tasks) in Pilot Study II.

Scenario	Tasks
Scenario 1: Time Comparison Scenario	i. This network displays all the emails that ***Matthew Lenhart*** (*Enron* employee) had been involved in any sort of way (Sent or Received). ii. Compare the network in ***June-Dec 2000*** with ***June-Dec 2001***. iii. Explain the story behind it regarding (keep in mind that *Enron* Corporation fell in October 2001): • The ***differences*** between the two periods • ***Crowded eras*** • ***Large gap*** (no activity) • ***Large*** emails
Scenario 2: Most Frequent Correspondents Scenario	i. This network displays all the emails that ***Andy Zipper*** (*Enron* vice president) had been involved in any sort of way (Sent or Received). ii. Explain the story behind it regarding: • The ones who mostly ***sent*** emails to him during ***April to June 2001*** especially from ***late May*** to ***June*** (the ***top 5***) • The type of the correspondence for each (***private*** or ***general*** message with respect to the ***size of sent messages***)
Scenario 3: Email Comparison Scenario	i. This network displays all the emails that ***Jeffrey Shankman*** (*Enron* president) had been involved in any sort of way (Sent or Received). ii. Select the emails of following email addresses: 1) Jeffrey.shankman@enron.com, 2) a..shankman@enron.com, 3) Jeffry.a. shankman@enron.com iii. Explain the story behind it based on: • The ***type (Sender/Receiver)*** of each email address • The ***life period (duration)*** of each email address • The ***activity*** level of each email address • Any ***switching*** between email addresses
Scenario 4: Search Scenario	i. This network displays all the emails that ***David Delaney*** (*Enron* CEO) had been involved in any sort of way (Sent or Received). ii. Search for emails about a management committee

	with this subject “***ENA Management Committee***” in the network (search is case-sensitive). iii. Explain the story behind it regarding: • The ones who **sent** those emails • The differences between the role of them (who seems to be the ***leader*** of the meeting and who acts as a ***secretary*** and reminding the meeting) • The ***location*** and ***time*** of the meeting • The ***type of meeting*** (*Once, Weekly, Biweekly, Monthly*)

4.2.2.4 Results and Discussion

Table 4-12 displays the results of “task difficulty” question, the first question in the post-questionnaire. The order of scenarios from difficult to easy is Scenario 2, Scenario 4, Scenario 1, and Scenario 3. Generally, from Table 4-11, the scenarios were not easy as they contain inferential and deductive tasks (In Scenario 2, finding out a row of the black nodes is the most frequent correspondent and the size of the sent black node represents the general and private message. In Scenario 3, discovering which email addresses were switched. In Scenario 4, identifying the leader and secretary).

Table 4-12. Number of participants (percentage) in answering "task difficulty" question (Q1) of post-questionnaire.

Scenario	Easy (%)	Medium (%)	Hard (%)
Scenario 1	66.6	33.3	-
Scenario 2	50.0	33.3	16.6
Scenario 3	83.3	16.6	-
Scenario 4	66.6	16.6	16.6

In questions 2 (See Post-Questionnaire Form in Appendix 3: User Study Documents), for each scenario, we asked participants whether they are able to recognize similar scenario using *EmailTime*. The answers were on a five point scale:

Not at all   Totally

1   2   3   4   5

Table 4-13 presents the number of participants in each bin.

Table 4-13. Before binning, number of participants in answering question 2, ability to recognize similar scenarios using *EmailTime*.

Task	Not at all true	Not true	Neutral	True	Totally true
	1	2	3	4	5
Scenario 1	-	-	-	3	3
Scenario 2	-	-	-	5	1
Scenario 3	-	-	-	2	4
Scenario 4	-	-	1	1	4

We also binned this five point scale into three categories: Not True, Neutral, True (See Table 4-14) as we were interested to see how many participants “agree” and “disagree” and we didn’t want a very sensitive measure of agreement. According to the results (See Table 4-14), almost all the participants mentioned after the testing session and learning how to use the system, they are able to recognize similar scenarios.

Table 4-14. After binning, number of participants (percentage) in answering question 2, ability to recognize similar scenarios using *EmailTime*.

Scenario	Not true (%)	Neutral (%)	True (%)
Scenario 1	-	-	100.0
Scenario 2	-	-	100.0
Scenario 3	-	-	100.0
Scenario 4	-	16.6	83.3

From question 3, 4, 5, and 10 (See Post-Questionnaire Form in Appendix 3: User Study Documents), we list like and dislike aspects of the control panel (See Table 4-15) and the visualization plot (See Table 4-16) based on the participant’s comments. Both could be improved.

Table 4-15. Like and dislike aspects of control panel based on participant's comments

Like	Dislike
Selecting a time periods and changing dates	Prefer a non-case sensitive search
Access to the functionality is good	"Selected item" tab, that displays message content, is hard to read as it displays all the information in the same colour
Filtering based on different things	Difficulty in selecting multiple email addresses using "ctrl + click"
Searching option	Prefer a button for activating search instead of pressing the "Enter" key in the keyboard.
Email List	The arrows on the top of the control panel are too small, hard to click on

Table 4-16. Like and dislike aspects of visualization plot based on participant's comments

Like	Dislike
Size of sent nodes that indicates private or public correspondence	Cannot click on circles inside other circle, overlapping problem
Green is a good colour for Cc, easy to find and clear	Black and blue colour for Sent and Received nodes as To were close to each other. Green is a bit light and make it hard to see
Pan, Zoom in and out	Sometimes hard to zoom in, out or pan. Zoom in and out seems vice versa, in other direction.
The grid layout, the plot	Understanding the plot and concept of sending and receiving is a bit difficult initially
Being able to compare sending and receiving emails for one person	Equal quantum for Time (X) axis, more organized date axis is preferred
Highlighting is useful	Plot sometimes is too crowded and close lines are sometimes confusing
Colour coding	Hard to remember the colour, code it into the system in a way
Tooltip is helpful	Tooltip is a bit slow, make it faster to display
Gaps are obvious in plot	Vertical Axis was not displaying all the email addresses, make it fisheye
-	Displays whose dataset is plotted in a label
-	Prefer to select part of the data in plot and it get zoomed in

4.2.3 User Study

The purpose of this user study was to find out whether users are able to complete the scenarios in specified time. The focuses were on the task completion time and correct answers.

4.2.3.1 Hypothesis

We hypothesized that through the visualization users are able to use the system's capabilities in order to find:

- Changes of activities over time (e.g. switching from one email address to another),
- Role of the owner of email addresses in an event (e.g. secretary or leader in a biweekly meeting in an organization dataset) and,
- Correspondence patterns between email users over time (such as most frequent correspondents and type of their correspondences; general or private messages)

These capabilities are a basis in interpreting the data that is visualized by *EmailTime*, such as:

1. *Time Comparison*; compare different time periods to each other and recognized their differences with respect to the crowded eras, large gaps (no activity), large emails, etc.
2. *Most Frequent Correspondents*; find out the most frequent correspondents of a person and type of their correspondences (private or general message based on the size of sent messages).
3. *Email address Comparison*; compare different email addresses to each other according to the type (sender, receiver or both), duration and activity level of each email address and discovering switches between email addresses.

These capabilities are the results of the plot visualization and the way we display the emails on it.

4.2.3.2 Participants

The second user study experiments for *EmailTime* took place on October 7th, 8th, 12th and 13th 2010 with 13 subjects (5 female and 8 male, n=13). Participants were SIAT graduate students from SFU. They were between 20 and 29 years of age and none of them were color-deficient. Most of them were a little familiar with the concept of visual analytics and visualization and had the experience of working with software with the purpose of visualization such as Tableau, Inspire, graphs of Excel, etc. All of them were used to work with the computer and had more than one email address which they logged-in more than once a day or always.

4.2.3.3 Scenarios and Tasks

Users accomplished 4 scenarios with respect to our hypothesis and system's capabilities (See Tasks in Appendix 3: User Study Documents):

- Scenario 1 is about time comparison,
- Scenario 2 is about finding most frequent correspondents,
- Scenario 3 is about email address comparison,
- Scenario 4 is about finding roles in an event.

4.2.3.4 Method

The testing sessions were similar to Pilot Study II, but the focus was on the task completion time and the number of correct answers. Therefore, the differences are:

- The tasks were more specific. User need to write the answers so the task sheet was changed,
- We timed the participants,
- We didn't ask about like and dislikes of the visualization or control panel so the post questionnaire was changed.

4.2.4 Results and Discussion

4.2.4.1 Analysis of Task Results

Generally, some parts of the scenarios 2, 3 and 4 contain inferential and deductive tasks, which make the tasks difficult (See Tasks in Appendix 3: User Study Documents).

The inferential parts are:

- In Scenario 2 (Most Frequent Correspondents), to infer that a row of the black nodes is the most frequent correspondent and the size of the sent black nodes represents the general and private message.
- In Scenario 3 (Email address Comparison), to discover which email addresses were switched.
- In Scenario 4 (Search), to identify the leader and secretary.

The majority of the participants were able to accomplish the tasks but four users became confused in the deductive task of the scenario 2 and asked the observer (me) to explain and guide them. No users became confused in answering the deductive tasks of scenario 3 and 4. Two participants answered the deductive task in the scenario 4 with uncertainty (but correctly).

Based on the types of questions and results in the previous studies, we estimated that the users would complete the scenarios in between 7 to 10 minutes (See Table 4-18 for tasks' completion time). We expected users to categorize scenario 2, 3 and 4 as medium or hard because they needed to do inference and draw conclusions in those scenarios (See Table 4-20). According to Table 4-20, the order of scenarios from hard to easy would be Scenario 2 (9 minutes), Scenario 3 (9 minutes), Scenario 4 (7.53 minutes) and Scenario 1 (7.61 minutes). This order is not exactly the same as the order in Pilot Study II, which is Scenario 2, Scenario 4, Scenario 1, and Scenario 3. We believe part of this difference is due to small changes in the task's format (as the task became more specific).

For the first scenario (See Tasks in Appendix 3: User Study Documents), the answers were on a five point scale:

Not at all   Totally
1   2   3   4   5

We binned those into three categories: Not true, Neutral, True (See Table 4-17) as we were interested to see how many participants "agree" and "disagree". We also didn't want our measure of agreement be very sensitive.

Each cell in Table 4-17 specifies the number of participants in percentage. In this scenario, we asked them to compare two time periods. No participant had a wrong answer in this scenario. The average completion time was 7.61 minutes (See Table 4-18 for more details).

Table 4-17. Number of participants (percentage) in answering the first scenario.

Question	First Period			Second Period		
	Not true	Neutral	True	Not true	Neutral	True
Seems Normal	-	-	100.0	100.0	-	-
Has Large Gaps	100.0	-	-	-	-	100.0
Has Crowded Eras	100.0	-	-	-	-	100.0
Has Large Emails	100.0	-	-	-	-	100.0

In the second scenario (See Tasks in Appendix 3: User Study Documents), users need to infer that the crowded black lines are the most frequent correspondents and the size of the black sent message node is determined by the recipient count and represents the type (general or private) of the message. No participant had a wrong answer in this scenario but four participants asked questions, became confused and began with a wrong approach. The average completion time was 9 minutes (See Table 4-18 for more details). In order to find the most frequent correspondents of Andy Zipper, the wrong approach was users selected Andy Zipper from the email address list which displays a line of Andy Zipper’s emails. The correct approach is users should filter out the received emails and keep only the sent emails in order to find out who frequently sent emails to Andy.

In the third scenario (See Tasks in Appendix 3: User Study Documents), the inferential part was discovering which email addresses were switched. Every participant answered that correctly. The average completion time was 9 minutes (See Table 4-18 for more details).

In the fourth scenario (See Tasks in Appendix 3: User Study Documents), we asked the participants to infer the identity of the leader and secretary. Although some participants became confused at first or doubted their conclusion, every participant answered correctly. For the fourth scenario the average completion time was 7.53 minutes (See Table 4-18 for more details).

Generally, a few users took a different approach in accomplishing some tasks. For example, in the third scenario two participants looked at each email address one by one instead of examining all email addresses simultaneously. In addition, one participant found the time through the message content instead of the visualization plot. Also in the fourth scenario, two participants found the attendees of a meeting through the content of a message instead of the plot visualization. More over, a few participants began with a wrong approach to answer the inferential questions and then realized what to do. For example, in scenario 2, to find out who sent Andy Zipper the message, four participants first filtered for “Andy Zipper” then realized could not find the information and decided to filter in the sent nodes (See Table 4-19).

We identified three groups of users based on their approach in completing the scenarios:

- Users of the first group were quick and confident in accomplishing the tasks,
- Users of the second group were curious and interested in using the tool to analyze more of the plot,
- Users of the third group were unconfident about their answers and what they did.

Table 4-18. Completion time (in minute) for each user in each scenario.

Users	Scenario Completion Time (min)				General Comment on the User's Performance by the Observer (me)
	#1	#2	#3	#4	
User #1	5	7	7	8	Very quick and confident
User #2	7	7	11	6	Curious and analyzed a lot – VA background
User #3	14	8	10	5	Play a lot and see
User #4	8	13	12	8	Very slow and not confidence about what to do
User #5	6	12	9	9	-
User #6	8	12	7	10	Curious and interested in the tool to play and find more
User #7	6	5	6	7	Very quick and confident
User #8	7	15	8	7	Good – art background
User #9	13	8	9	7	Curious and wanted to see more in content – VA background
User #10	7	8	11	8	Act with no confidence
User #11	7	9	9	8	Good
User #12	4	6	8	7	Very quick and confident
User #13	7	7	10	8	-
Ave.	7.61	9	9	7.53	-
STDEV	2.84	3.02	1.77	1.26	-

Table 4-19 summarizes the problems during each scenario.

Table 4-19. Summary of the problems in each scenario.

Scenario	Summary of the Problem in each Scenario in the Study
#1	Visualization disappeared (crashed) in the beginning of two testing sessions (while users were playing with the system), as users were too fast in zooming in and out, so we reset the visualization.
#2	Some users got confused in interpreting of the sent emails with “Large” and “Medium” number of recipients, Some of the participants got confused and began with the wrong approach and I guided four of them in accomplishing the task. (they did not get the complete concept of the plot visualization in the introduction), A few of them could not infer that the size of the sent nodes represents the number of participants which indicates the type of correspondence (general or private message), Two participants did not filter out the received nodes to find the senders (which makes it easier to answer the tasks).
#3	Multi selection in the email list was difficult, as users need to hold the “ctrl” button and scroll down. I explained what I meant by “switching” between email addresses (Stop using one email address and start using another one).
#4	I mentioned to three users to search for the complete keywords (e.g. “ENA Management Committee”).

4.2.4.2 Results on Post Questionnaire Analysis

For question 1 (task difficulty – See Post-Questionnaire Form), we summarized the results in Table 4-20. As mentioned before, scenarios contained some analytical and inferential judgements from the visualization. It appears that users completed easier tasks faster based on the completion time in Table 4-18 and the task difficulty in Table 4-20.

Table 4-20. Number of participants (percentage) in answering task difficulty level question (Q1).

Task	Easy (%)	Medium (%)	Hard (%)
Scenario 1	61.54	23.07	15.39
Scenario 2	23.07	46.16	30.77
Scenario 3	53.84	38.46	7.70
Scenario 4	53.84	46.16	-

For question 2 of Post-Questionnaire Form, ability to recognize similar scenarios using *EmailTime,* the answers were on a five point scale:

Not at all				Totally
1	**2**	**3**	**4**	**5**

Table 4-21 presents the number of participants in each bin.

Table 4-21. Before binning, number of participants in answering question 2, ability to recognize similar scenarios using *EmailTime*.

Task	Not at all true	Not true	Neutral	True	Totally true
	1	2	3	4	5
Scenario 1	-	1	-	2	10
Scenario 2	-	-	4	7	2
Scenario 3	-	-	-	5	8
Scenario 4	-	-	2	6	5

We also binned those into three categories: Not True, Neutral, True (See Table 4-22) as we were interested to see how many participants “agree” and “disagree” and we didn’t want to have a very sensitive measure of agreement. From Table 4-22, in scenario 2, about 30% of users were neutral and in scenario

4, 15% were neutral. But generally most of them mentioned that they are able to recognize similar scenarios.

Table 4-22. After binning, number of participants (percentage) in answering question 2, ability to recognize similar scenarios using *EmailTime*.

Task	Not true (%)	Neutral (%)	True (%)
Scenario 1	7.70	-	92.30
Scenario 2	-	30.77	69.23
Scenario 3	-	-	100.0
Scenario 4	-	15.39	84.61

For questions 6, 7, 8 and 9 the answers were on a five point scale:

1	**2**	**3**	**4**	**5**
not at all true		somewhat true		very true

Table 4-23 presents the number of participants in each bin. As we were interested to see how many participants “agree” and “disagree” in questions 6, 7, 8 and 9 and we didn’t want to have a very sensitive measure of agreement, we binned those into three categories: not true (disagree), somewhat true (neither agree nor disagree), true (agree) (See Table 4-24). Each cell specifies the number of participants in percentage.

Table 4-23. Before binning, number of participants in answering Q6, 7, 8 and 9.

Question	Not at all true	Not true	Somewhat true	True	Very true
	1	2	3	4	5
Q6 (feel competent)	-	-	4	7	2
Q7 (easy to learn to use)	-	1	5	3	4
Q8 (need introduction)	-	1	1	2	9
Q9 (need to concentrate)	-	4	3	2	4

Table 4-24. After binning, number of participants (percentage) in answering Q3, 4, 5 and 6.

Question	Not true (%)	Somewhat true (%)	True (%)
Q3 (feel competent)	-	30.77	69.23
Q4 (easy to learn to use)	7.70	38.46	53.84
Q5 (need introduction)	7.70	7.70	84.61
Q6 (need to concentrate)	30.77	23.07	46.16

According to Table 4-24, more than 80% of the users mentioned the introduction was necessary to understand the concept of the visualization and learn how to work with the system. About half of the users agreed that they needed to concentrate while working with the system and completing the tasks and it was easy to learn to use.

5: CONCLUSION AND FUTURE WORK

We have introduced *EmailTime*, a new style of visualizing email traffic that helps an analyst see patterns of correspondence over a significant period of time in email archives. This tool enables the user to analyze and visualize hundreds of stored emails over time and discern patterns of correspondence. Interaction is an important aspect of the tool as it provides zooming, panning, filtering, highlighting and evaluating measurements, such as centrality, frequency and histogram.

It helps analysts make more sense out of a collection of emails. It can be used for the datasets of an individual or an organization to visualize the correspondence and to analyse their email behaviour.

After considering the limitations of this work, we can improve it from different perspectives such as system implementation, system testing, usability study with different hypotheses, etc. Therefore as our next steps, we are considering:

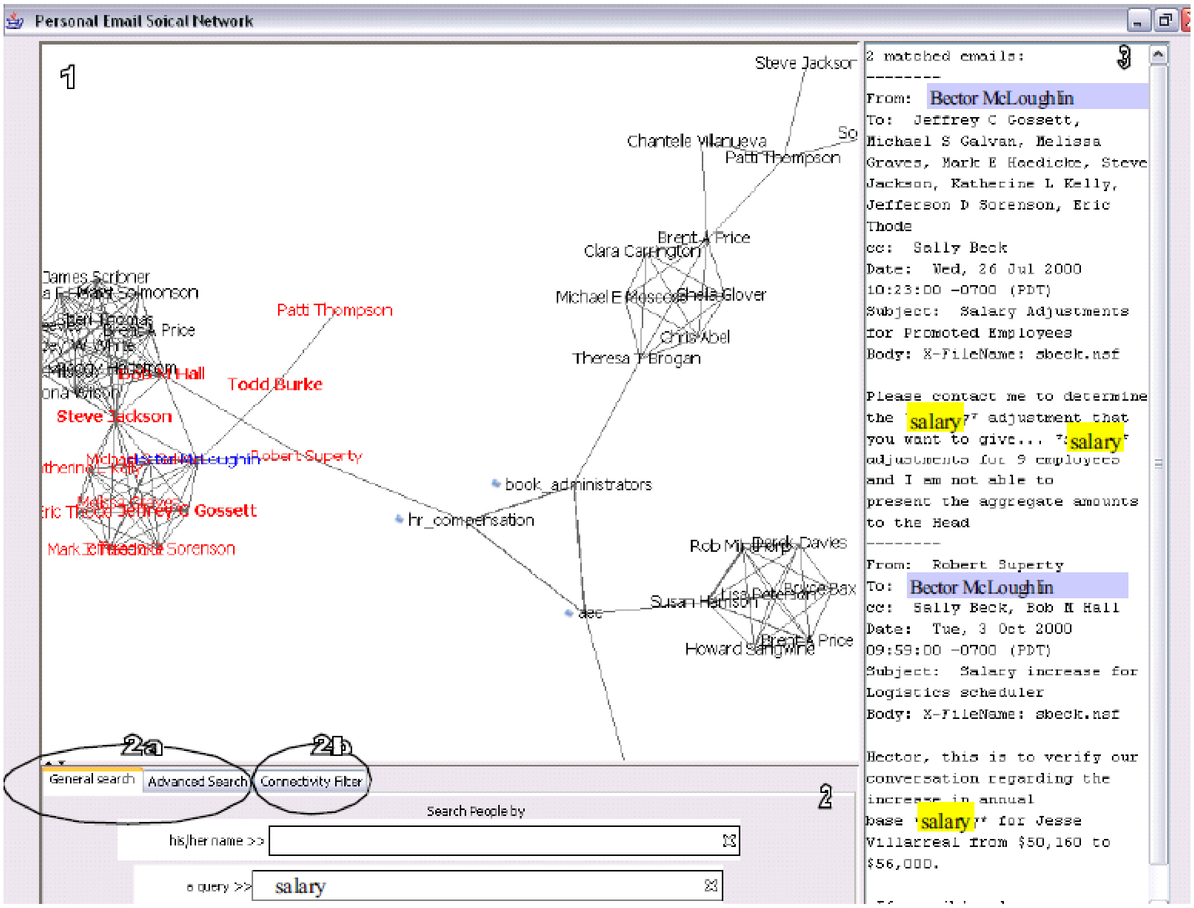
- From system implementation perspective:
    - Improve system's interaction by incorporating the results of user studies,
    - Add and/or improve the system's functionalities (Centrality, Frequency, histogram),
    - Display *Bcc* email would be even more interesting,

  - Make it available as a web application and encourage users to visualize the content of their own mailbox.

- From system testing perspective:
  - Make a comparison with another email visualization system,
  - Import a different kind of dataset to *EmailTime* (e.g. chat messages).
- From usability study perspective:
  - Usability study on the system functionalities (frequency, centrality, histogram, etc),
  - Usability study on the coding of information in the plot (e.g. colour coding or shape coding could better specify the node type).

6: APPENDICES

Appendix 1: Figures of Related Works



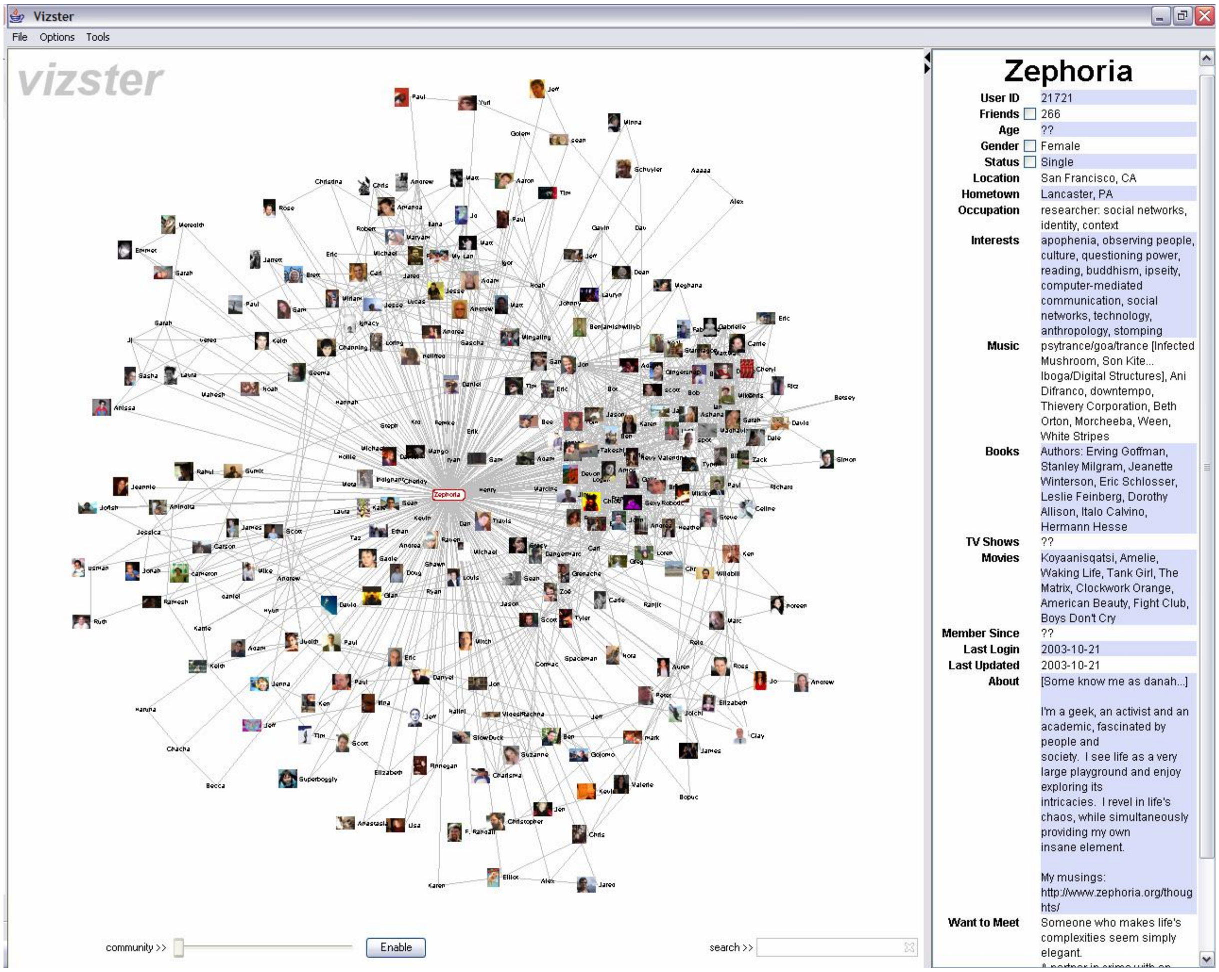
2 matched emails:

From: Bector McLoughlin
To: Jeffrey C Gossett, Michael S Galvan, Melissa Graves, Mark E Haedicke, Steve Jackson, Katherine L Kelly, Jefferson D Sorenson, Eric Thode
cc: Sally Beck
Date: Wed, 26 Jul 2000 10:23:00 -0700 (PDT)
Subject: Salary Adjustments for Promoted Employees
Body: X-FileName: sbeck.nsf

Please contact me to determine the 'salary'' adjustment that you want to give... ''salary'' adjustments for 9 employees and I am not able to present the aggregate amounts to the Head
----
From: Robert Superty
To: Bector McLoughlin
cc: Sally Beck, Bob M Hall
Date: Tue, 3 Oct 2000 09:59:00 -0700 (PDT)
Subject: Salary increase for Logistics scheduler
Body: X-FileName: sbeck.nsf

Hector, this is to verify our conversation regarding the increase in annual base 'salary'' for Jesse Villarreal from $50,160 to $56,000.

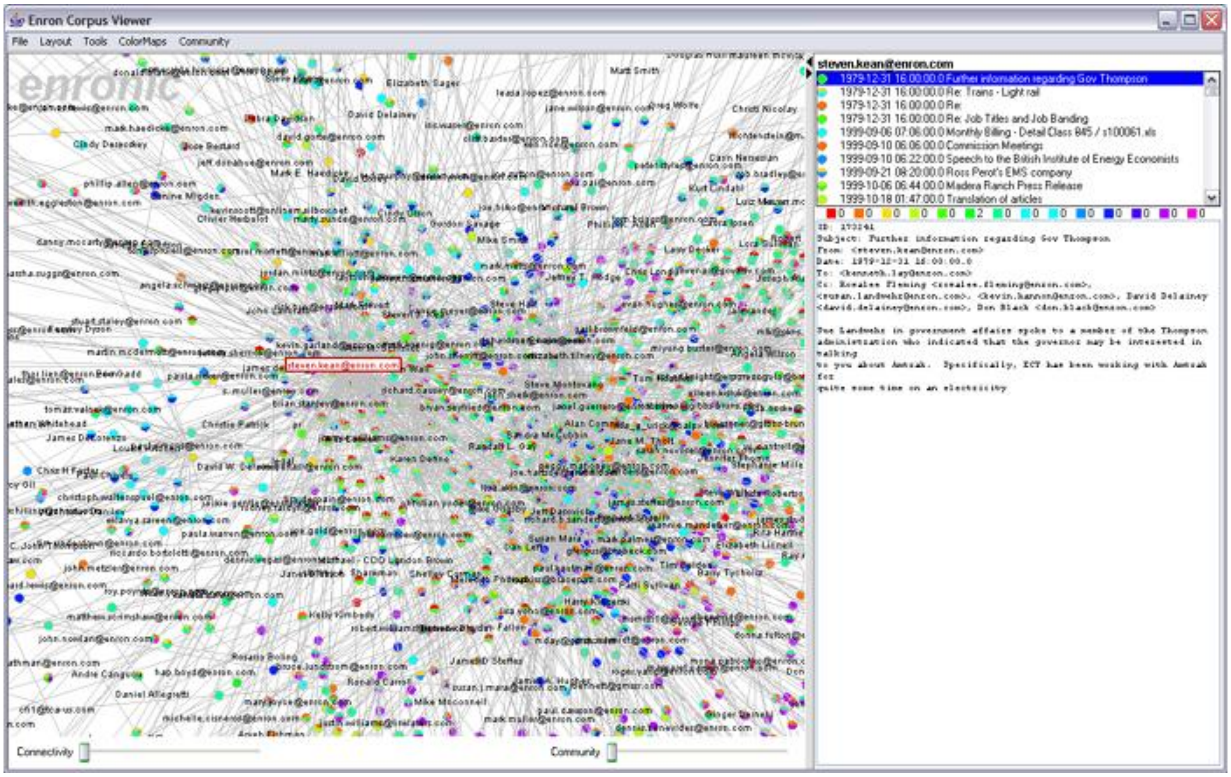
Figure 6-1. Screen shot of *VizPEAM* [1]. 1) view panel; 2) control panel with 2a) search filter with search for 'salary' and 2b) connectivity filter; 3) result panel shown two matched emails for the selected correspondent 'BectorMcLouglin'.



Zephoria

User ID	21721
Friends □	266
Age	??
Gender □	Female
Status □	Single
Location	San Francisco, CA
Hometown	Lancaster, PA
Occupation	researcher: social networks, identity, context
Interests	apophenia, observing people, culture, questioning power, reading, buddhism, ipseity, computer-mediated communication, social networks, technology, anthropology, stomping
Music	psytrance/goa/trance [Infected Mushroom, Son Kite... Iboga/Digital Structures], Ani Difranco, downtempo, Thievery Corporation, Beth Orton, Morcheeba, Ween, White Stripes
Books	Authors: Erving Goffman, Stanley Milgram, Jeanette Winterson, Eric Schlosser, Leslie Feinberg, Dorothy Allison, Italo Calvino, Hermann Hesse
TV Shows	??
Movies	Koyaanisqatsi, Amelie, Waking Life, Tank Girl, The Matrix, Clockwork Orange, American Beauty, Fight Club, Boys Don't Cry
Member Since	??
Last Login	2003-10-21
Last Updated	2003-10-21
About	[Some know me as danah...] I'm a geek, an activist and an academic, fascinated by people and society. I see life as a very large playground and enjoy exploring its intricacies. I revel in life's chaos, while simultaneously providing my own insane element. My musings: http://www.zephoria.org/thoughts/
Want to Meet	Someone who makes life's complexities seem simply elegant. A partner in crime with an

Figure 6-2. Screen shot of the *Vizster* visualization system [2]. The left side presents a network display with controls for community analysis and keyword search. The right side consists of a panel displaying a selected member's profile information. Words in the profile panel that occur in more than one profile will highlight on mouse-over; clicking these words will initiate searches for those terms. The checkboxes in the profile panel will initiate an "X-ray" view of that particular profile dimension.



steven.kean@enron.com

```
1979-12-31 16:00:00.0 Further information regarding Gov Thompson
1979-12-31 16:00:00.0 Re: Trains - Light rail
1979-12-31 16:00:00.0 Re:
1979-12-31 16:00:00.0 Re: Job Titles and Job Banding
1999-09-06 07:06:00.0 Monthly Billing - Detail Class 845 / s100061.xls
1999-09-10 06:06:00.0 Commission Meetings
1999-09-10 06:22:00.0 Speech to the British Institute of Energy Economists
1999-09-21 08:20:00.0 Ross Perot's EMS company
1999-10-06 06:44:00.0 Madera Ranch Press Release
1999-10-18 01:47:00.0 Translation of articles
```

🟥 0 🟧 0 🟨 0 🟩 0 🟩 0 🟩 2 🟩 0 🟦 0 🟦 0 🟦 0 🟦 0 🟪 0 🟪 0 🟪 0

```
ID: 173241
Subject: Further information regarding Gov Thompson
From: <steven.kean@enron.com>
Date: 1979-12-31 16:00:00.0
To: <kenneth.lay@enron.com>
Cc: Rosalee Fleming <rosalee.fleming@enron.com>,
<susan.landwehr@enron.com>, <kevin.hannon@enron.com>, David Delainey
<david.delainey@enron.com>, Don Black <don.black@enron.com>

Sue Landwehr in government affairs spoke to a member of the Thompson
administration who indicated that the governor may be interested in
talking
to you about Amtrak.  Specifically, ECT has been working with Amtrak
for
quite some time on an electricity
```

Figure 6-3. Screen shot of *Enronic* [3]. Visualization of Enron social network is on the left. Colour legend for category labels and the Message Viewer are on the right.

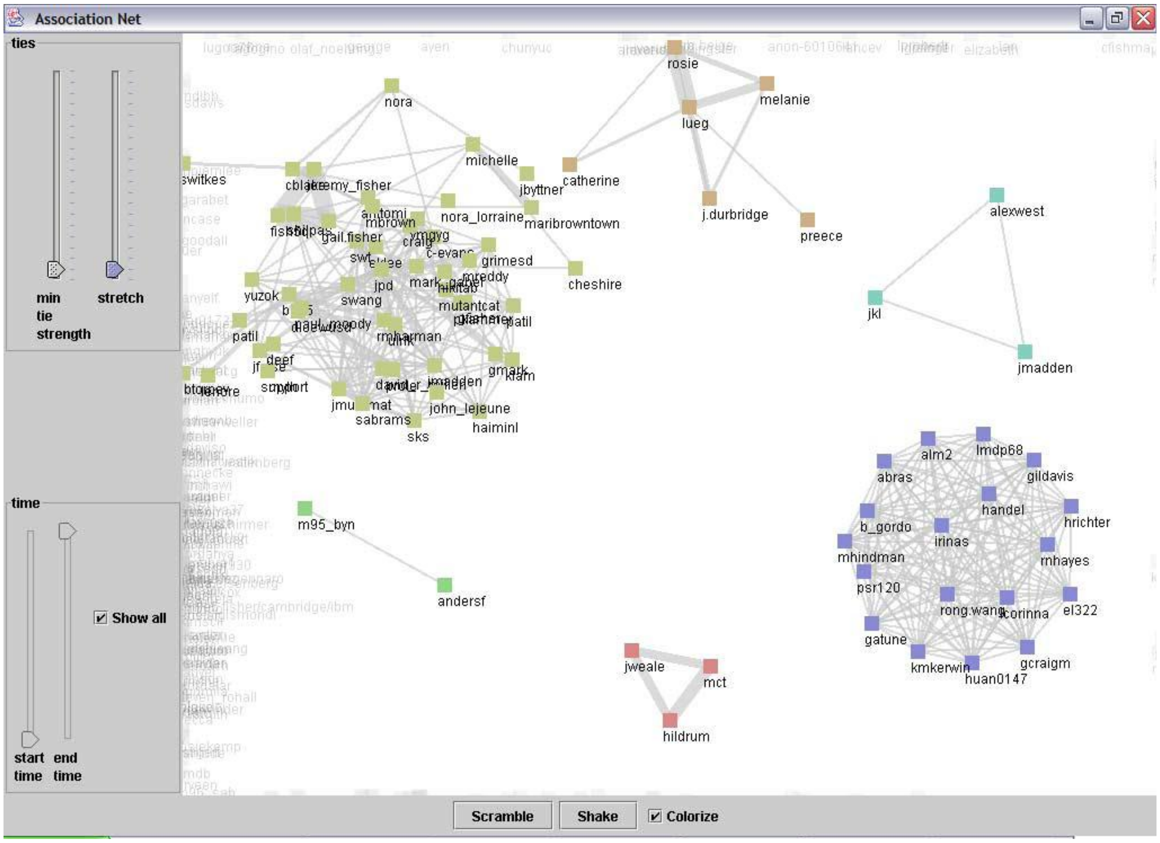
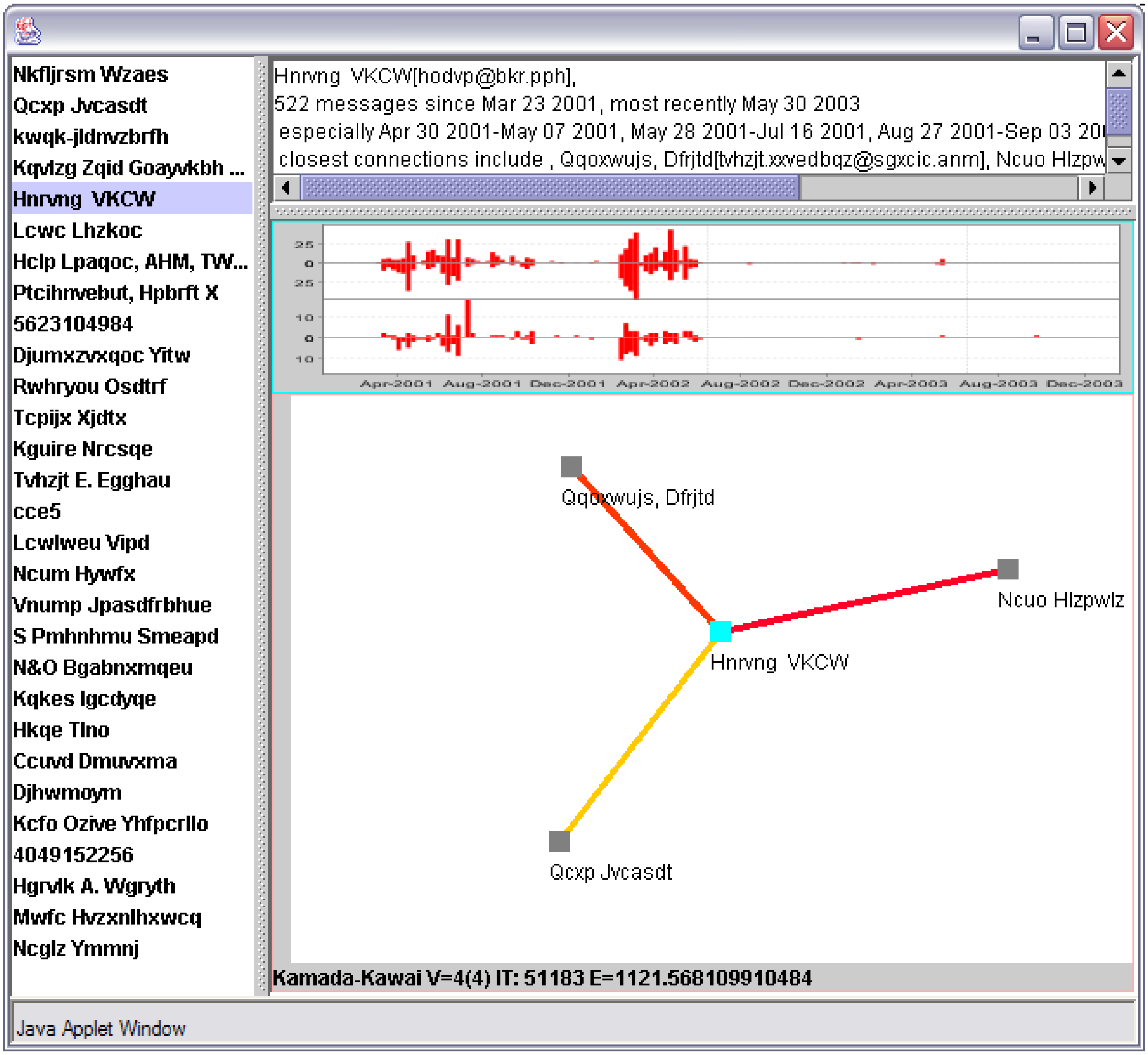


Figure 6-4. The *Soylent* [4] network view.



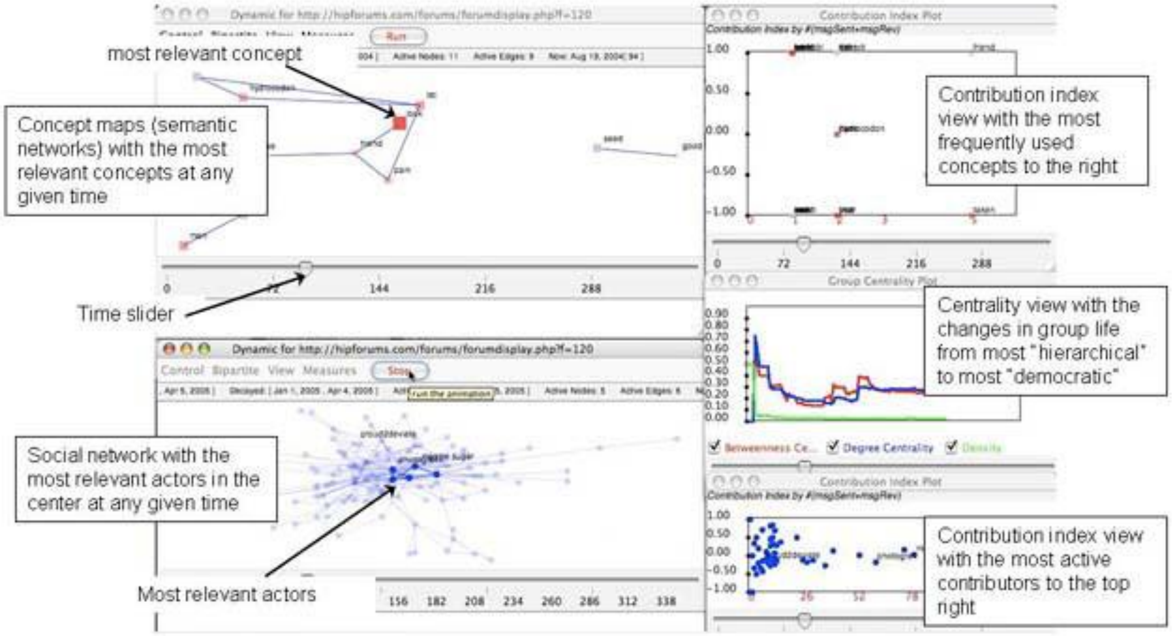
Nkfljrsm Wzaes
Qcxp Jvcasdt
kwqk-jldnvzbrfh
Kqvlzg Zqid Goayvkbh ...
Hnrvng VKCW
Lcwc Lhzkoc
Hclp Lpaqoc, AHM, TW...
Ptcihnvebut, Hpbrft X
5623104984
Djumxzvxqoc Yitw
Rwhryou Osdtrf
Tcpijx Xjdtx
Kguire Nrcsqe
Tvhzjt E. Egghau
cce5
Lcwlweu Vipd
Ncum Hywfx
Vnump Jpasdfrbhue
S Pmhnhmu Smeapd
N&O Bgabnxmqeu
Kqkes Igcdyqe
Hkqe Tlno
Ccuvd Dmuvxma
Djhwmoym
Kcfo Ozive Yhfpcrllo
4049152256
Hgrvlk A. Wgryth
Mwfc Hvzxnlhxwcq
Ncglz Ymmnj

Hnrvng VKCW[hodvp@bkr.pph],
522 messages since Mar 23 2001, most recently May 30 2003
especially Apr 30 2001-May 07 2001, May 28 2001-Jul 16 2001, Aug 27 2001-Sep 03 20
closest connections include , Qqoxwujs, Dfrjtd[tvhzjt.xxvedbqz@sgxcic.anm], Ncuo Hlzpw

Kamada-Kawai V=4(4) IT: 51183 E=1121.568109910484

Java Applet Window

Figure 6-5. ***TellMeAbout*** **demo [4].The left side has a list of people; The top panel tells that the user exchanged 522 messages with Hnrnvng. The second panel shows the frequency of, in order: Outgoing messages TO Hnrvng, Incoming messages FROM Hnrvng, Outgoing attachments TO Hnrvng, Incoming attachments FROM Hnrvng. Indexed by date.**


most relevant concept

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

Time slider

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

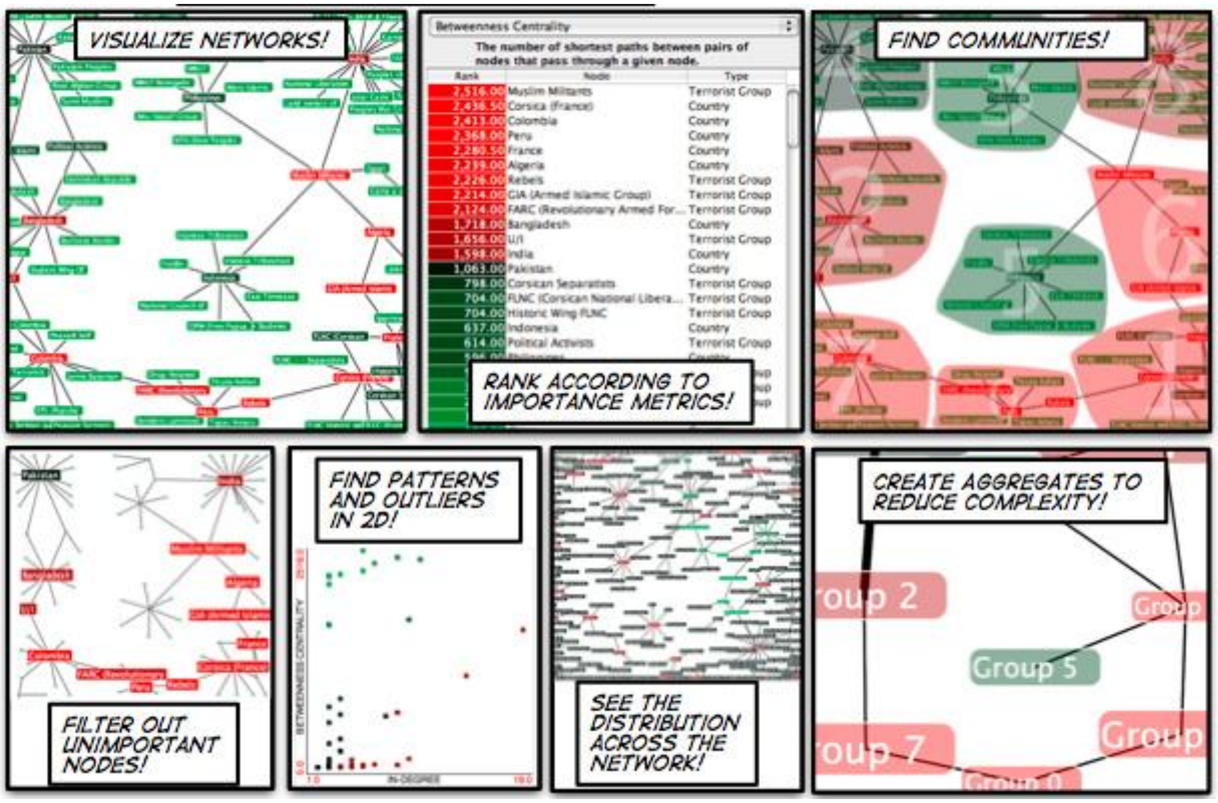
Most relevant actors

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

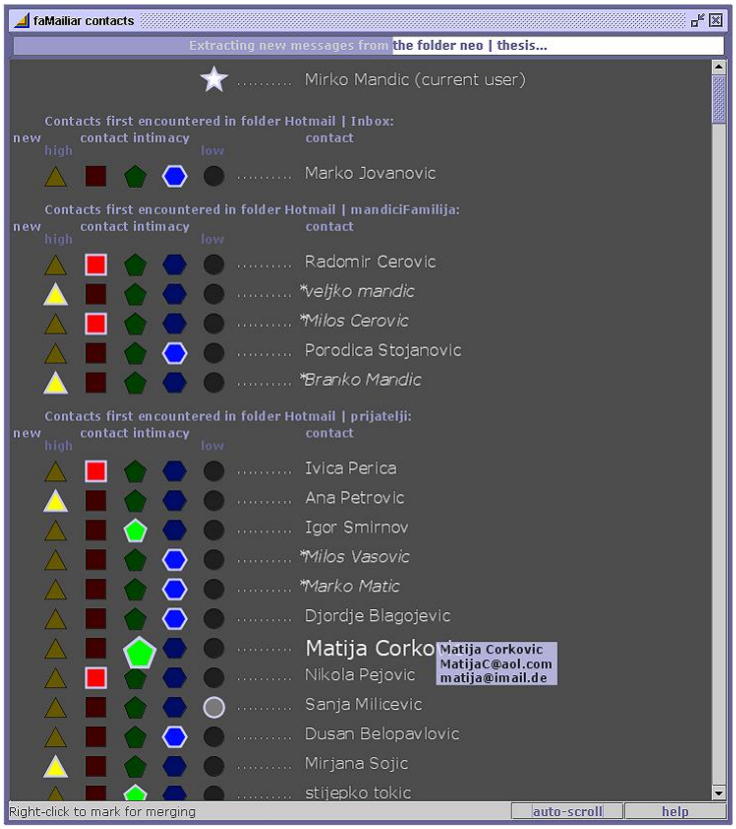
Figure 6-6. Five main views of *TeCFlow* [5].


VISUALIZE NETWORKS!
Betweenness Centrality
RANK ACCORDING TO IMPORTANCE METRICS!
FIND COMMUNITIES!
FILTER OUT UNIMPORTANT NODES!
FIND PATTERNS AND OUTLIERS IN 2D!
BETWEENNESS CENTRALITY
IN-DEGREE
SEE THE DISTRIBUTION ACROSS THE NETWORK!
CREATE AGGREGATES TO REDUCE COMPLEXITY!
Group 5

The number of shortest paths between pairs of nodes that pass through a given node.

Rank	Node	Type
2,516.00	Muslim Militants	Terrorist Group
2,436.50	Corsica (France)	Country
2,413.00	Colombia	Country
2,368.00	Peru	Country
2,280.50	France	Country
2,239.00	Algeria	Country
2,226.00	Rebels	Terrorist Group
2,214.00	GIA (Armed Islamic Group)	Terrorist Group
2,124.00	FARC (Revolutionary Armed For...	Terrorist Group
1,718.00	Bangladesh	Country
1,656.00	U/I	Terrorist Group
1,598.00	India	Country
1,063.00	Pakistan	Country
798.00	Corsican Separatists	Terrorist Group
704.00	FLNC (Corsican National Libera...	Terrorist Group
704.00	Historic Wing FLNC	Terrorist Group
637.00	Indonesia	Country
614.00	Political Activists	Terrorist Group

Figure 6-7. ***SocialAction*** **views [8]. Used with permission University of Maryland Human-Computer Interaction Lab, http://www.cs.umd.edu/hcil/socialaction**



Extracting new messages from the folder neo | thesis...

.... Mirko Mandic (current user)

Contacts first encountered in folder Hotmail | Inbox:

new contact intimacy contact

high low

.... Marko Jovanovic

Contacts first encountered in folder Hotmail | mandiciFamilija:

new contact intimacy contact

high low

.... Radomir Cerovic

.... *veljko mandic

.... *Milos Cerovic

.... Porodica Stojanovic

.... *Branko Mandic

Contacts first encountered in folder Hotmail | prijatelji:

new contact intimacy contact

high low

.... Ivica Perica

.... Ana Petrovic

.... Igor Smirnov

.... *Milos Vasovic

.... *Marko Matic

.... Djordje Blagojevic

.... Matija Corko

Matija Corkovic
MatijaC@aol.com
matija@imail.de

.... Nikola Pejovic

.... Sanja Milicevic

.... Dusan Belopavlovic

.... Mirjana Sojic

.... stijepko tokic

Right-click to mark for merging

Figure 6-8. Snapshot of *famailiar* [9, 10]. Contact window.

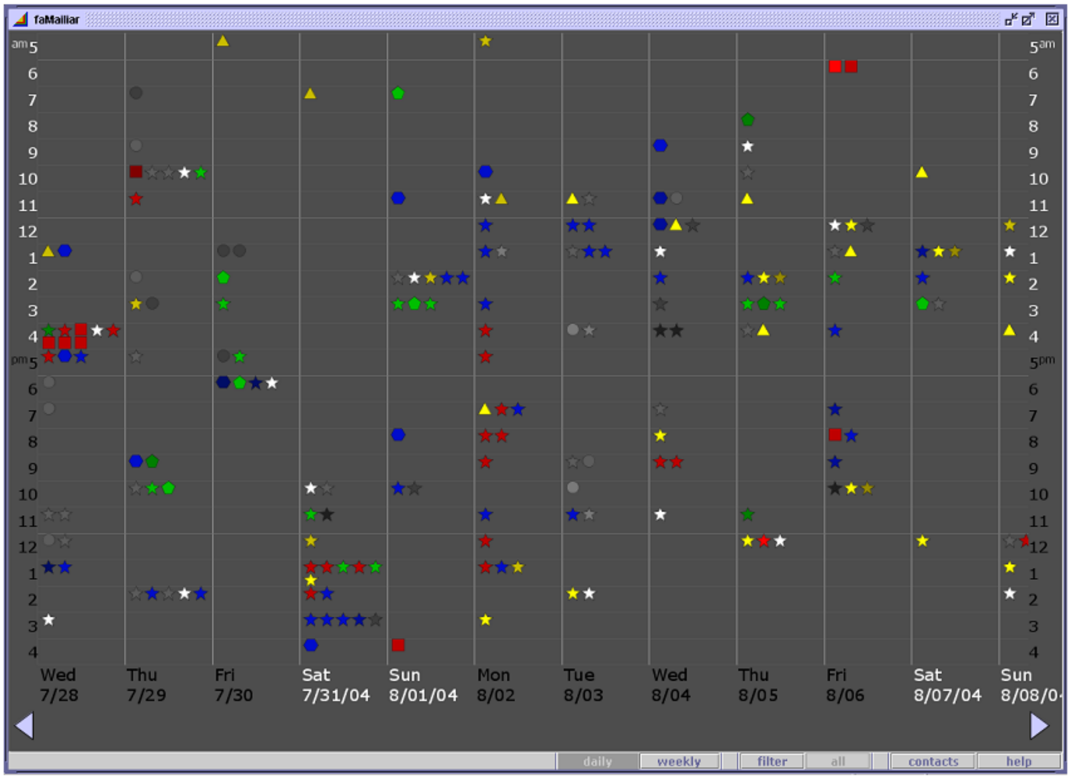
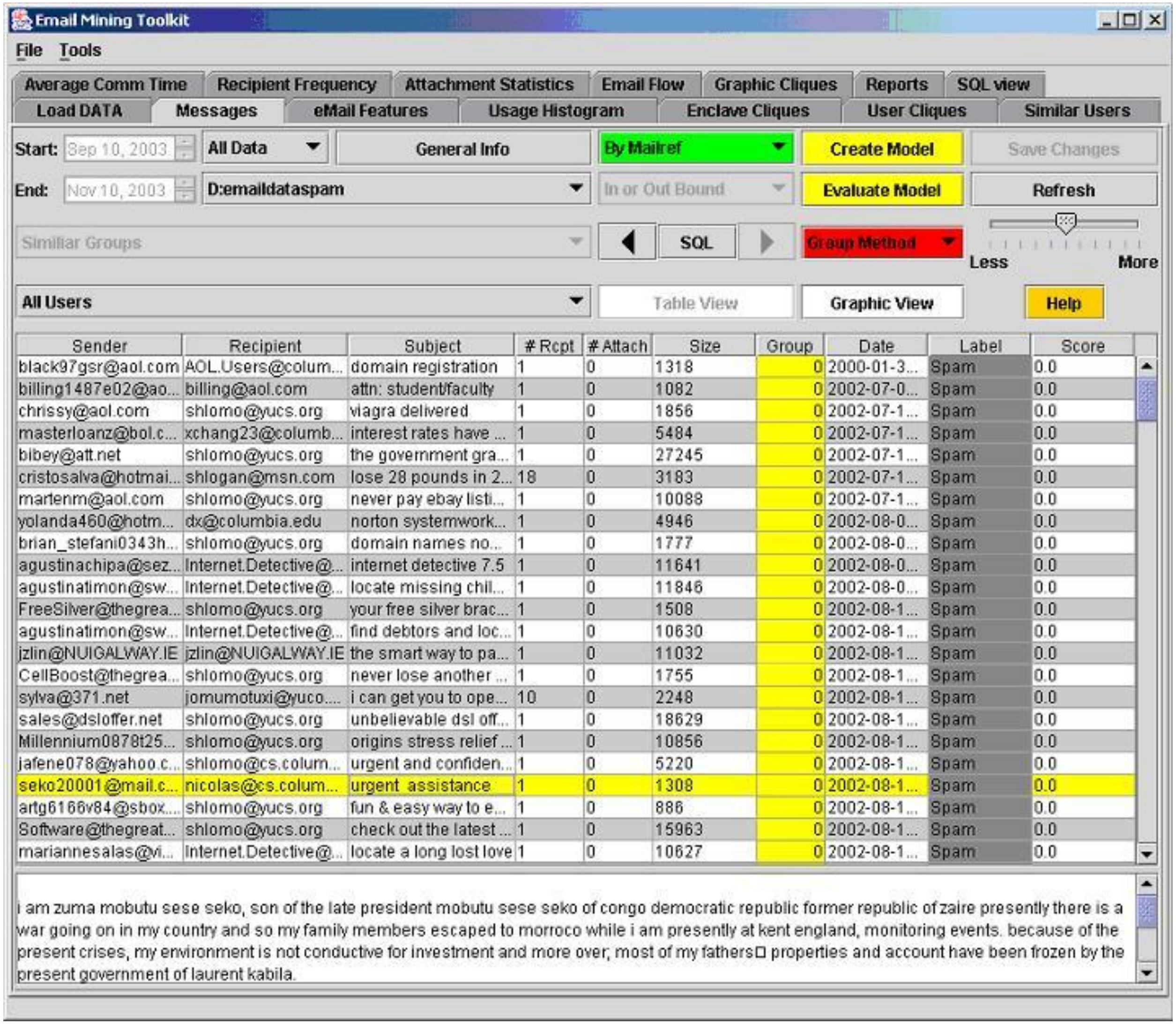


Figure 6-9. Snapshot of *famailiar* [9, 10]. Daily view of a user's email.

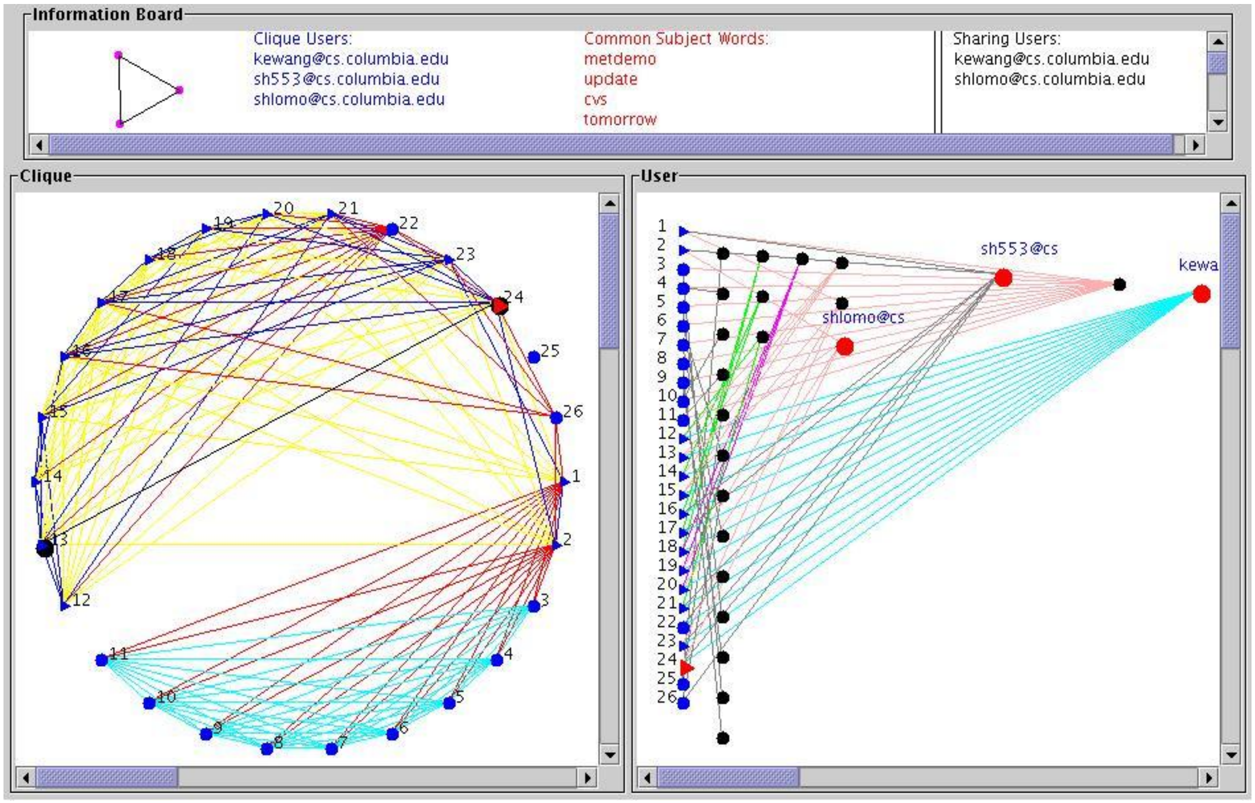


Start: Sep 10, 2003 All Data By Mailref
End: Nov 10, 2003 D:emaildataspam In or Out Bound
Similiar Groups Group Method
All Users

Sender	Recipient	Subject	# Rcpt	# Attach	Size	Group	Date	Label	Score
black97gsr@aol.com	AOL.Users@colum...	domain registration	1	0	1318	0	2000-01-3...	Spam	0.0
billing1487e02@ao...	billing@aol.com	attn: student/faculty	1	0	1082	0	2002-07-0...	Spam	0.0
chrissy@aol.com	shlomo@yucs.org	viagra delivered	1	0	1856	0	2002-07-1...	Spam	0.0
masterloanz@bol.c...	xchang23@columb...	interest rates have ...	1	0	5484	0	2002-07-1...	Spam	0.0
bibey@att.net	shlomo@yucs.org	the government gra...	1	0	27245	0	2002-07-1...	Spam	0.0
cristosalva@hotmai...	shlogan@msn.com	lose 28 pounds in 2...	18	0	3183	0	2002-07-1...	Spam	0.0
martenm@aol.com	shlomo@yucs.org	never pay ebay listi...	1	0	10088	0	2002-07-1...	Spam	0.0
yolanda460@hotm...	dx@columbia.edu	norton systemwork...	1	0	4946	0	2002-08-0...	Spam	0.0
brian_stefani0343h...	shlomo@yucs.org	domain names no...	1	0	1777	0	2002-08-0...	Spam	0.0
agustinachipa@sez...	Internet.Detective@...	internet detective 7.5	1	0	11641	0	2002-08-0...	Spam	0.0
agustinatimon@sw...	Internet.Detective@...	locate missing chil...	1	0	11846	0	2002-08-0...	Spam	0.0
FreeSilver@thegrea...	shlomo@yucs.org	your free silver brac...	1	0	1508	0	2002-08-1...	Spam	0.0
agustinatimon@sw...	Internet.Detective@...	find debtors and loc...	1	0	10630	0	2002-08-1...	Spam	0.0
jzlin@NUIGALWAY.IE	jzlin@NUIGALWAY.IE	the smart way to pa...	1	0	11032	0	2002-08-1...	Spam	0.0
CellBoost@thegrea...	shlomo@yucs.org	never lose another ...	1	0	1755	0	2002-08-1...	Spam	0.0
sylva@371.net	jomumotuxi@yuco....	i can get you to ope...	10	0	2248	0	2002-08-1...	Spam	0.0
sales@dsloffer.net	shlomo@yucs.org	unbelievable dsl off...	1	0	18629	0	2002-08-1...	Spam	0.0
Millennium0878t25...	shlomo@yucs.org	origins stress relief ...	1	0	10856	0	2002-08-1...	Spam	0.0
jafene078@yahoo.c...	shlomo@cs.colum...	urgent and confiden...	1	0	5220	0	2002-08-1...	Spam	0.0
seko20001@mail.c...	nicolas@cs.colum...	urgent assistance	1	0	1308	0	2002-08-1...	Spam	0.0
artg6166v84@sbox....	shlomo@yucs.org	fun & easy way to e...	1	0	886	0	2002-08-1...	Spam	0.0
Software@thegreat...	shlomo@yucs.org	check out the latest ...	1	0	15963	0	2002-08-1...	Spam	0.0
mariannesalas@vi...	Internet.Detective@...	locate a long lost love	1	0	10627	0	2002-08-1...	Spam	0.0

i am zuma mobutu sese seko, son of the late president mobutu sese seko of congo democratic republic former republic of zaire presently there is a war going on in my country and so my family members escaped to morroco while i am presently at kent england, monitoring events. because of the present crises, my environment is not conductive for investment and more over, most of my fathers□ properties and account have been frozen by the present government of laurent kabila.

Figure 6-10. Snapshot of *Email Mining Toolkit (EMT)* [11, 12, 13, 14]. Message window.



Information Board

Clique Users:
kewang@cs.columbia.edu
sh553@cs.columbia.edu
shlomo@cs.columbia.edu

Common Subject Words:
metdemo
update
cvs
tomorrow

Sharing Users:
kewang@cs.columbia.edu
shlomo@cs.columbia.edu

A. Clique Panel.

B. User Panel

Figure 6-11. Snapshot of *Email Mining Toolkit (EMT)* [11, 12, 13, 14]. *a.* Clique Panel. Nodes (Cliques) are small polygons and the number of their edges is the number of members of the clique (e.g. a triangle is a 3-clique). Edges are the common members. b. User Panel. Blue nodes in the left most column are the (indexed) cliques and black nodes (users) are one distinct email addresses placed in different columns depending upon the number of cliques they belong to.

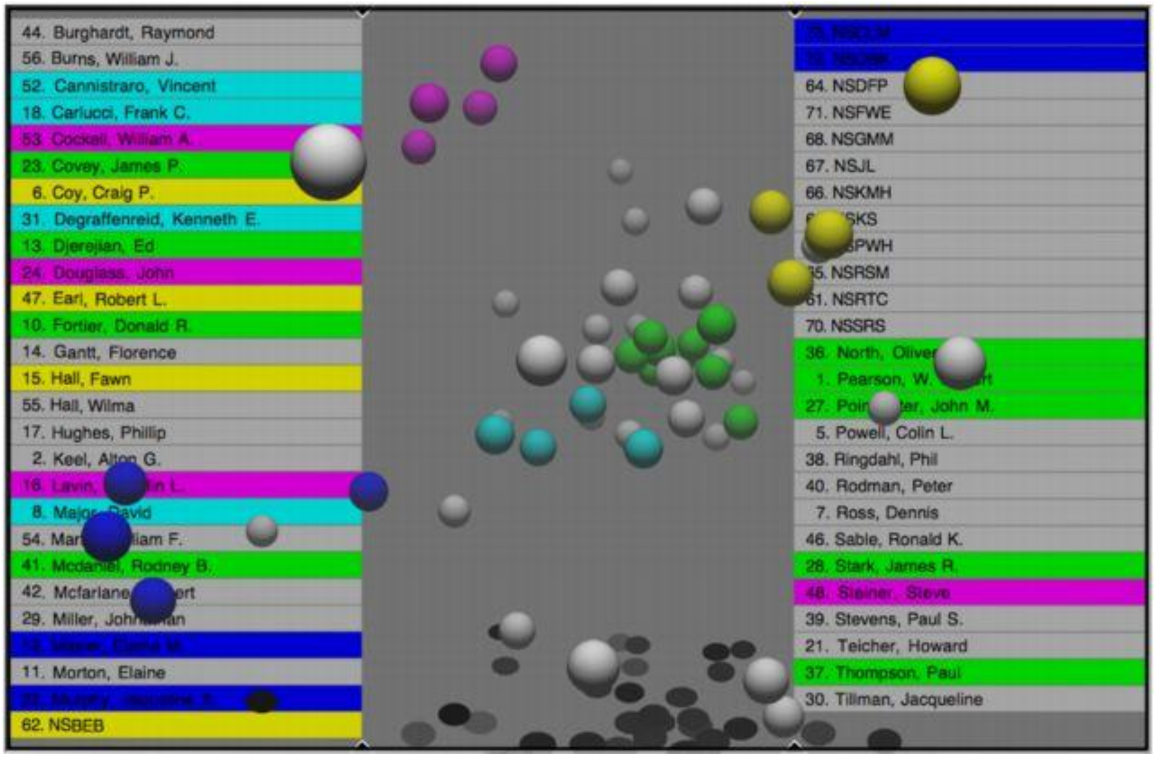


Figure 6-12. *eArchivarius* [16].

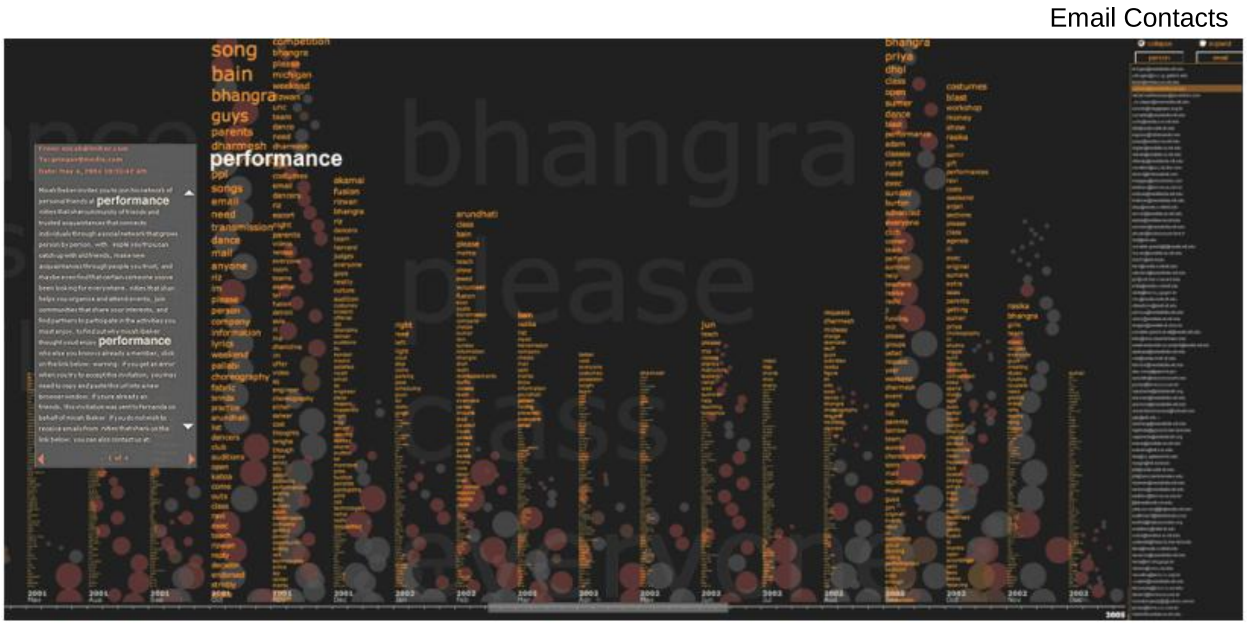


Figure 6-13. Snapshot of *Themail* [17]. A user's email exchange with a friend during 18 months. It shows multiple layers of information; monthly (yellow) and yearly (white) words. The more frequent and distinctive a word is, the bigger it is.

v

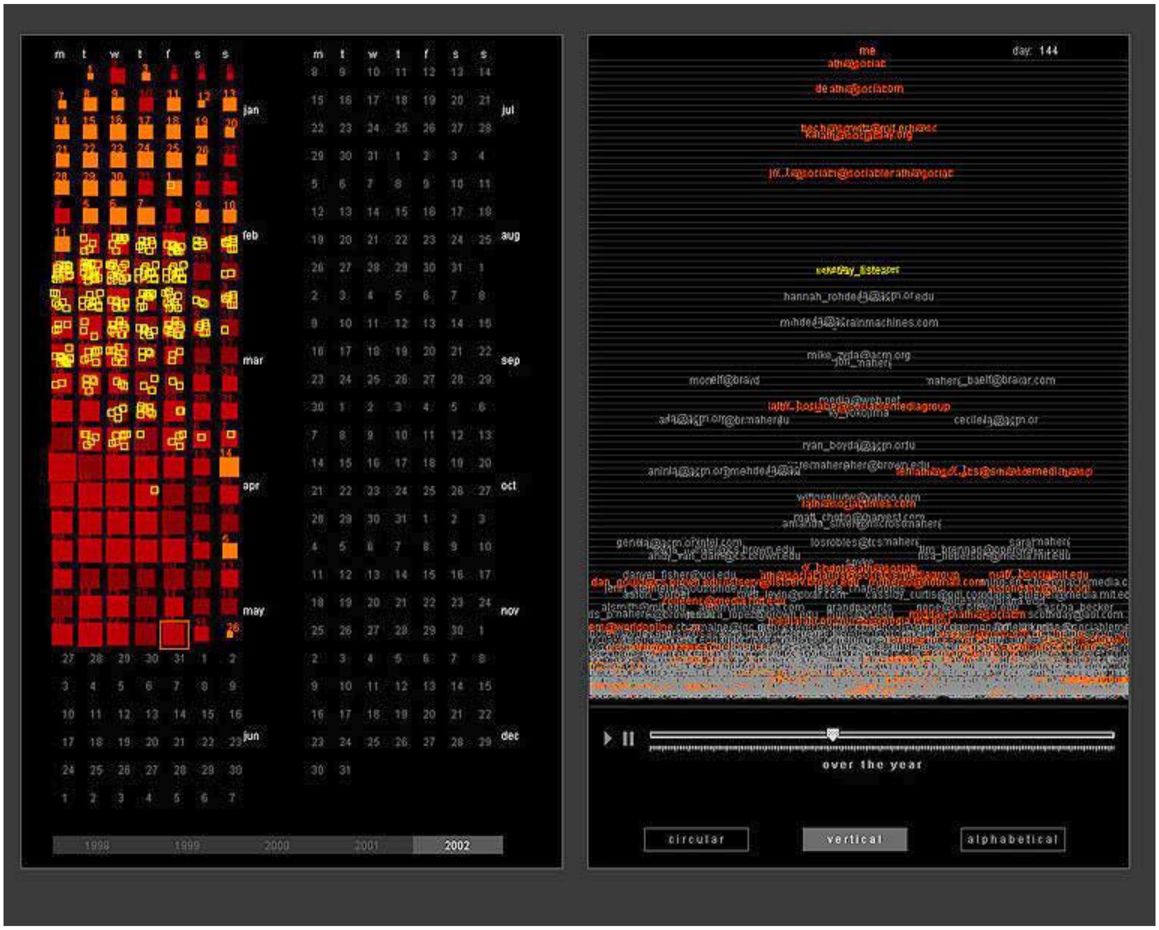


Figure 6-14. *PostHistory* [18, 19] interface with calendar panel on the left and contacts panel on the right. Names on the right panel move higher to reflect more intense email exchanges with ego. As time progresses and the intensity of exchange changes, names either slide back down or stay stationary.

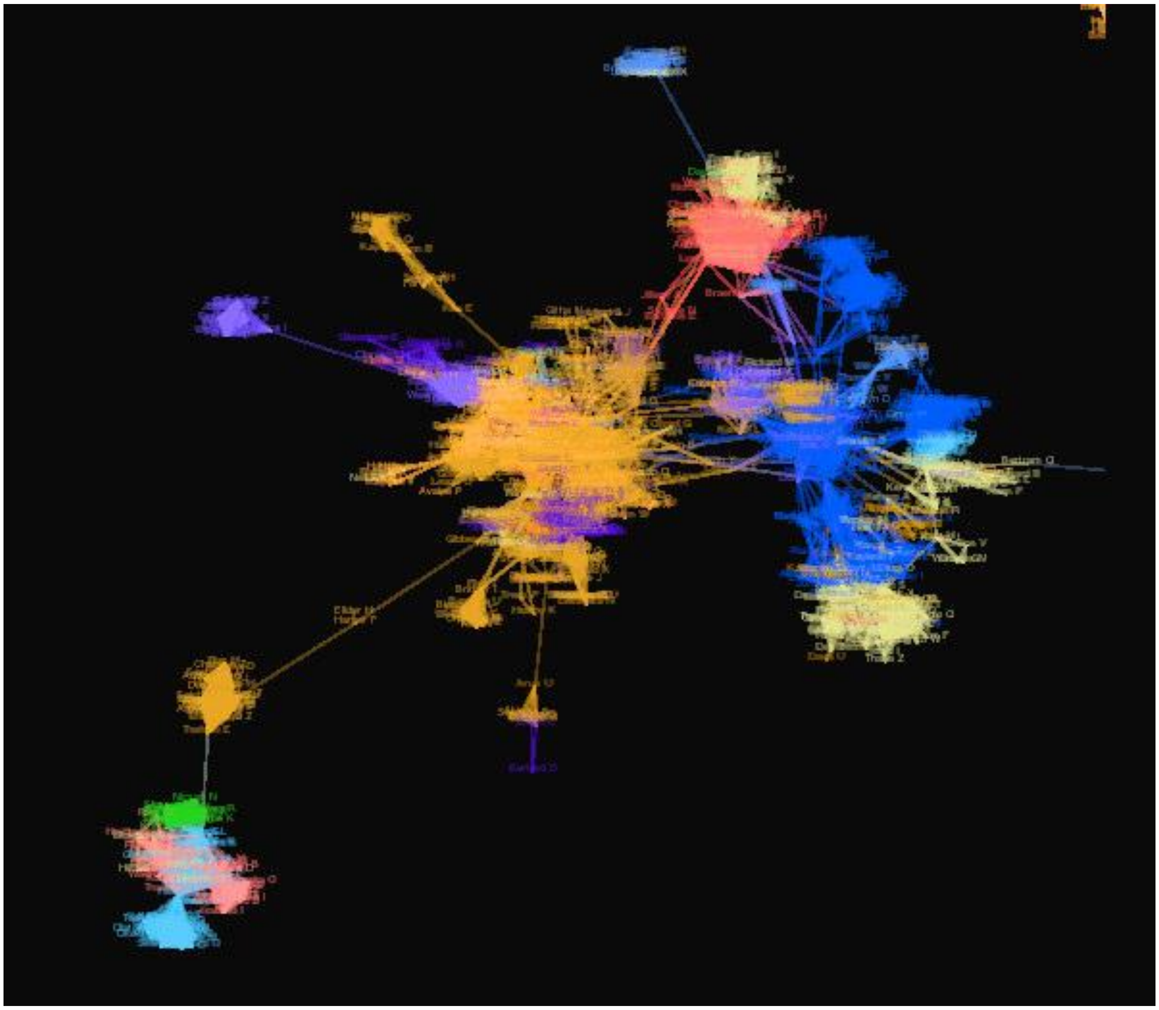
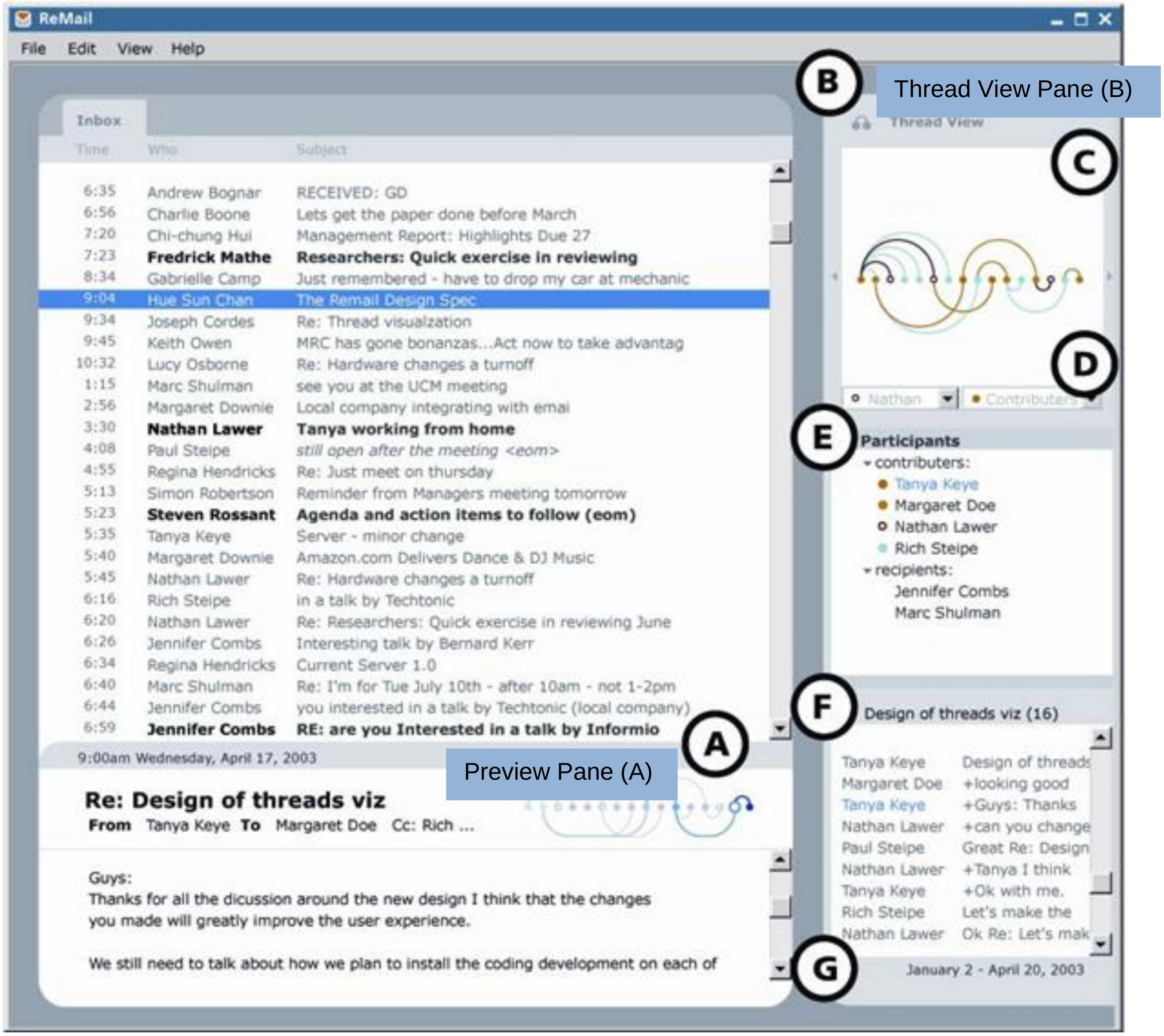


Figure 6-15. A complex cluster of contacts in *SNF* [18, 19]. The colors indicate that the cluster includes people from different contexts of ego's social life: family, school friends and work colleagues.



Time	Who	Subject
6:35	Andrew Bognar	RECEIVED: GD
6:56	Charlie Boone	Lets get the paper done before March
7:20	Chi-chung Hui	Management Report: Highlights Due 27
7:23	**Fredrick Mathe**	**Researchers: Quick exercise in reviewing**
8:34	Gabrielle Camp	Just remembered - have to drop my car at mechanic
9:04	Hue Sun Chan	The Remail Design Spec
9:34	Joseph Cordes	Re: Thread visualzation
9:45	Keith Owen	MRC has gone bonanzas...Act now to take advantag
10:32	Lucy Osborne	Re: Hardware changes a turnoff
1:15	Marc Shulman	see you at the UCM meeting
2:56	Margaret Downie	Local company integrating with emai
3:30	**Nathan Lawer**	**Tanya working from home**
4:08	Paul Steipe	*still open after the meeting <eom>*
4:55	Regina Hendricks	Re: Just meet on thursday
5:13	Simon Robertson	Reminder from Managers meeting tomorrow
5:23	**Steven Rossant**	**Agenda and action items to follow (eom)**
5:35	Tanya Keye	Server - minor change
5:40	Margaret Downie	Amazon.com Delivers Dance & DJ Music
5:45	Nathan Lawer	Re: Hardware changes a turnoff
6:16	Rich Steipe	in a talk by Techtonic
6:20	Nathan Lawer	Re: Researchers: Quick exercise in reviewing June
6:26	Jennifer Combs	Interesting talk by Bernard Kerr
6:34	Regina Hendricks	Current Server 1.0
6:40	Marc Shulman	Re: I'm for Tue July 10th - after 10am - not 1-2pm
6:44	Jennifer Combs	you interested in a talk by Techtonic (local company)
6:59	**Jennifer Combs**	**RE: are you Interested in a talk by Informio**

9:00am Wednesday, April 17, 2003

Re: Design of threads viz

From Tanya Keye **To** Margaret Doe Cc: Rich ...

Guys:
Thanks for all the dicussion around the new design I think that the changes you made will greatly improve the user experience.

We still need to talk about how we plan to install the coding development on each of

Participants

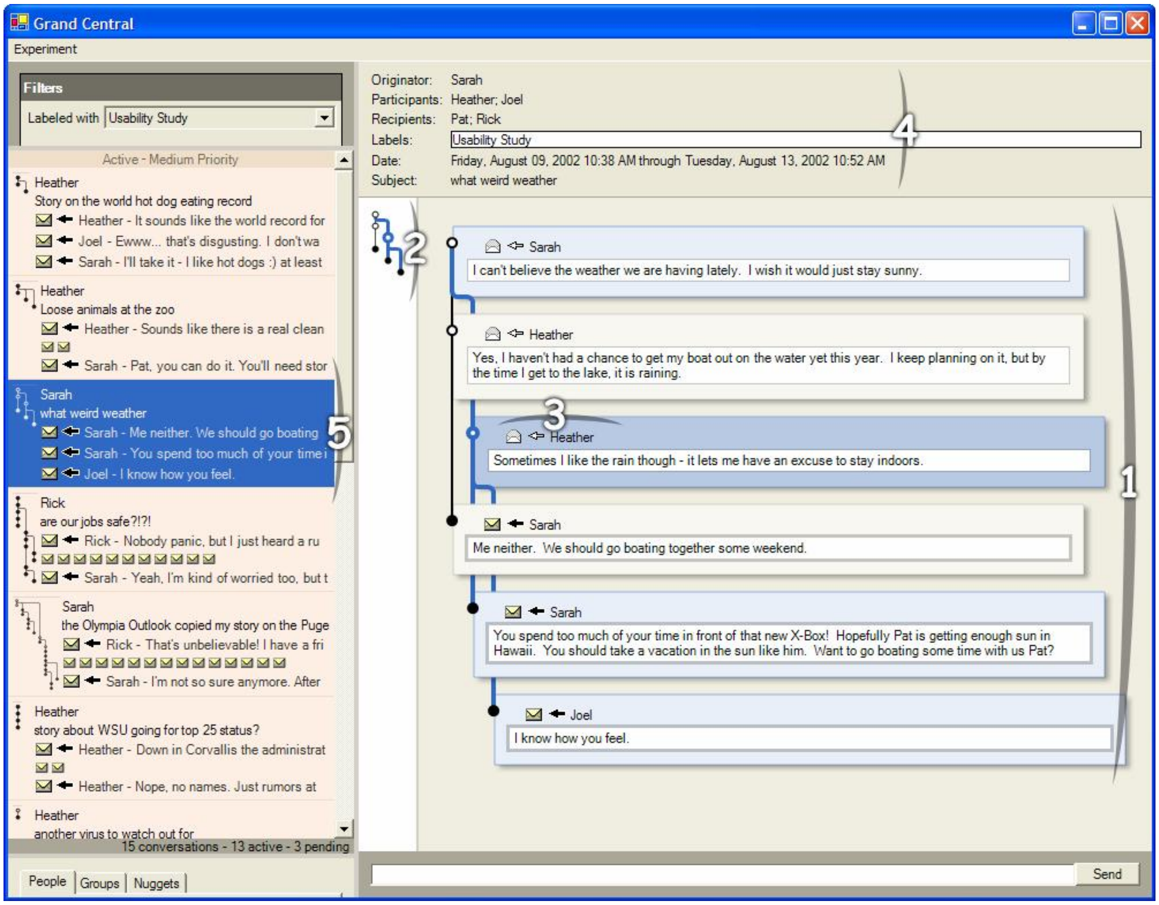
- contributers:
  - ● Tanya Keye
  - ● Margaret Doe
  - ○ Nathan Lawer
  - ● Rich Steipe
- recipients:
  - Jennifer Combs
  - Marc Shulman

Design of threads viz (16)

Tanya Keye	Design of threads
Margaret Doe	+looking good
Tanya Keye	+Guys: Thanks
Nathan Lawer	+can you change
Paul Steipe	Great Re: Design
Nathan Lawer	+Tanya I think
Tanya Keye	+Ok with me.
Rich Steipe	Let's make the
Nathan Lawer	Ok Re: Let's mak

January 2 - April 20, 2003

Figure 6-16. Screenshot of *Reinvented Email (ReMail)* [21]. Preview Pane (A) and Thread View Pane (B) display *Thread Arc* [22]. IBM Research Remail project, http://www.research.ibm.com/remail/



Labeled with: Usability Study

Active - Medium Priority

Heather
Story on the world hot dog eating record
Heather - It sounds like the world record for
Joel - Ewww... that's disgusting. I don't wa
Sarah - I'll take it - I like hot dogs :) at least

Heather
Loose animals at the zoo
Heather - Sounds like there is a real clean
Sarah - Pat, you can do it. You'll need stor

Sarah
what weird weather
Sarah - Me neither. We should go boating
Sarah - You spend too much of your time i
Joel - I know how you feel.

Rick
are our jobs safe?!?!
Rick - Nobody panic, but I just heard a ru
Sarah - Yeah, I'm kind of worried too, but t

Sarah
the Olympia Outlook copied my story on the Puge
Rick - That's unbelievable! I have a fri
Sarah - I'm not so sure anymore. After

Heather
story about WSU going for top 25 status?
Heather - Down in Corvallis the administrat
Heather - Nope, no names. Just rumors at

Heather
another virus to watch out for

15 conversations - 13 active - 3 pending

Originator: Sarah
Participants: Heather; Joel
Recipients: Pat; Rick
Labels: Usability Study
Date: Friday, August 09, 2002 10:38 AM through Tuesday, August 13, 2002 10:52 AM
Subject: what weird weather

Sarah
I can't believe the weather we are having lately. I wish it would just stay sunny.

Heather
Yes, I haven't had a chance to get my boat out on the water yet this year. I keep planning on it, but by the time I get to the lake, it is raining.

Heather
Sometimes I like the rain though - it lets me have an excuse to stay indoors.

Sarah
Me neither. We should go boating together some weekend.

Sarah
You spend too much of your time in front of that new X-Box! Hopefully Pat is getting enough sun in Hawaii. You should take a vacation in the sun like him. Want to go boating some time with us Pat?

Joel
I know how you feel.

Figure 6-17. The conversation visualization incorporated into a conversation-based email Client [23].

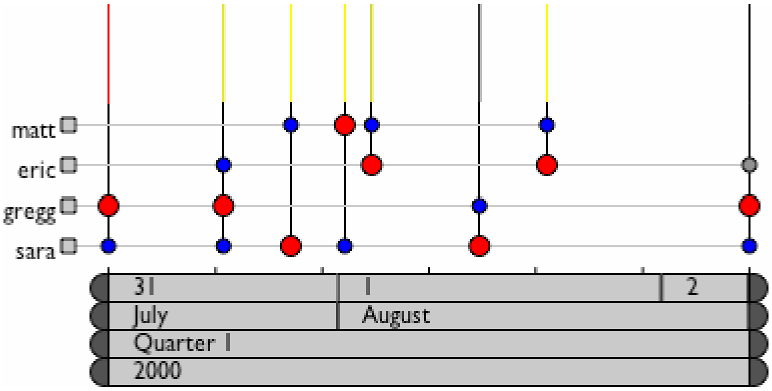


Figure 6-18. Visualization of a discussion [24].

Appendix 2: Tables and Figures of Case Study

Tables

There were seven organizational positions in the public *Enron* dataset namely CEO, President, Vice President, Manager, Director, Trader and Employee. For Jan 2000 to Dec 2001, the statistics of these seven categories in the *Enron* datasets are shown in the following tables.

Table 6-1. Mean and Standard deviation of numbers of sent and received emails.

Organizational Position	Count	No. Emails (sent + received)		No. Sent		No. Received (To + Cc)	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
CEO	4	3083.5	725.3	1433.7	1595.6	1649.7	1184.3
Director	12	666.92	348.4	186.2	174.9	480.6	274.7
Employee	35	3161.2	4593.8	2129.4	4176.5	1446.7	1878.6
Manager	14	4030.7	7802.1	2026.4	4014.5	2004.3	3968.6
President	4	2748.2	1310.6	1185.0	1167.9	1563.2	1148.3
Trader	11	1242.6	1868.8	664.3	1448.1	578.2	641.2
Vice President	21	2520.2	2734.9	1163.7	1222.6	1356.5	1906.7
Total	101	2623.7	4251.3	1458.9	3023.7	1308.5	2080.5

Table 6-2. Mean and Standard deviation of numbers of received emails as determined by *To*: and *Cc*: fields and numbers of email addresses.

Organizational Position	Count	No. To		No. Cc		No. Email Adr	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
CEO	4	1608.7	1180.6	41.0	20.4	8.0	5.2
Director	12	470.9	272.9	9.7	19.4	3.5	1.5
Employee	35	1340.7	1706.1	105.9	217.2	3.7	2.1
Manager	14	1824.7	3529.8	179.5	444.3	3.8	2.4
President	4	1485.2	1091.1	78.0	70.2	4.2	2.1
Trader	11	561.4	629.5	16.8	25.5	3.2	1.5
Vice President	21	1250.9	1716.9	105.5	193.8	4.2	1.9
Total	101	1217.3	1872.1	91.2	228.5	3.9	2.3

Table 6-3. Mean and Standard deviation of Contribution Index, To-CI and Cc-CI.

Organizational Position	Count	CI		To-CI		Cc-CI	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
CEO	4	.1681	.9409	.1588	.9492	.5068	.5887
Director	12	.5008	.3787	.4932	.3815	.8233	.3006
Employee	35	.0188	.4526	.0004	.4510	.8828	.1599
Manager	14	.0159	.5320	.0329	.5270	.8637	.1548
President	4	.2078	.5947	.1866	.6015	.8061	.1526
Trader	11	.1331	.5655	.1212	.5716	.7423	.4976
Vice President	21	.0348	.4662	.0115	.4640	.8168	.2681
Total	101	.1004	.5099	.0831	.5112	.8261	.2796

Table 6-4. Mean and Standard deviation of numbers of sent emails with Single, Small, Medium and Large number of recipients.

Organizational Position	Count	Single		Small		Medium		Large	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
CEO	4	1131.5	1267.2	275.7	426.4	20.5	31.1	6.0	2.7
Director	12	174.2	173.1	10.4	10.9	1.3	2.2	.3	1.1
Employee	30	740.2	1152.2	164.6	331.2	15.3	31.7	3.1	10.9
Manager	12	719.3	725.9	92.8	82.7	10.6	13.6	.5	1.1
President	4	1067.2	1075.5	108.2	99.5	7.5	5.8	2.0	3.3
Trader	11	555.8	1164.1	78.9	204.1	29.3	84.3	.2	.9
Vice President	20	864.2	924.5	151.5	222	10.3	17.8	2.2	5.2
Total	93	700.2	978.7	124.8	245.2	13.4	35.6	1.9	6.9

Table 6-5. Number of emails have been sent to multiple recipients (Small, Medium and Large number of recipients).

Organizational Position	Recipient Count		
	Small	Medium	Large
CEO	1103	82	24
President	433	30	8
Vice President	3031	207	44
Director	125	16	4
Manager	1114	128	6
Trader	868	323	3
Employee	4938	461	97

Table 6-6. Descriptive Statistics for the normalized number of sent and received emails.

	Organizational Position	Mean	Std. Deviation	Number
Normalized number of sent emails as determined by From: field	CEO	41.5944	47.04706	4
	Director	24.9573	18.93799	12
	Employee	52.5670	33.88464	35
	Manager	50.7989	26.60297	14
	President	39.6076	29.73977	4
	Trader	43.3409	28.27891	11
	Vice President	48.2587	23.31482	21
	Total	46.1931	29.65013	101
Normalized number of received emails as determined by To: field	CEO	57.0629	46.75877	4
	Director	73.4858	18.94149	12
	Employee	49.6595	21.42394	35
	Manager	46.9289	25.28366	14
	President	57.5423	29.29326	4
	Trader	54.2030	26.62694	11
	Vice President	48.7076	21.82922	21
	Total	53.0142	24.51737	101
Normalized number of received emails as determined by Cc: field	CEO	1.3428	.61089	4
	Director	1.5569	2.40760	12
	Employee	2.3210	2.61096	35
	Manager	2.2722	2.38607	14
	President	2.8501	2.15506	4
	Trader	2.4561	4.20014	11
	Vice President	3.0337	2.71052	21
	Total	2.3686	2.70294	101
Normalized number of received emails as determined by both To: and Cc: fields	CEO	58.4056	47.04706	4
	Director	75.0427	18.93799	12
	Employee	51.9805	22.14503	35
	Manager	49.2011	26.60297	14
	President	60.3924	29.73977	4
	Trader	56.6591	28.27891	11
	Vice President	51.7413	23.31482	21
	Total	55.3827	25.29683	101

Table 6-7. Descriptive Statistics for the normalized number of sent emails with the Small, Medium and Large number of recipients.

	Organizational Position	Mean	Std. Deviation	Number
Normalized number of sent emails with Small number of recipients	CEO	15.8171	10.71973	4
	Director	7.6762	7.46363	12
	Employee	17.4947	9.89317	30
	Manager	13.5589	6.62454	12
	President	9.2513	5.40231	4
	Trader	8.6582	6.58517	11
	Vice President	12.8756	8.81862	20
	Total	13.2547	9.03912	93
Normalized number of sent emails with Medium number of recipients	CEO	.9507	.94000	4
	Director	.7467	1.24107	12
	Employee	1.4315	3.25716	30
	Manager	2.2380	4.43789	12
	President	.9057	.86744	4
	Trader	4.2196	9.64879	11
	Vice President	.9543	1.65772	20
	Total	1.6311	4.21073	93
Normalized number of sent emails with Large number of recipients	CEO	8.5227	11.06461	4
	Director	.0623	.21583	12
	Employee	.6013	1.93382	30
	Manager	.0402	.09793	12
	President	.2800	.49676	4
	Trader	.2755	.91367	11
	Vice President	.5863	1.77644	20
	Total	.7445	2.95497	93

Figures

We used SPSS [31] to create the figures.

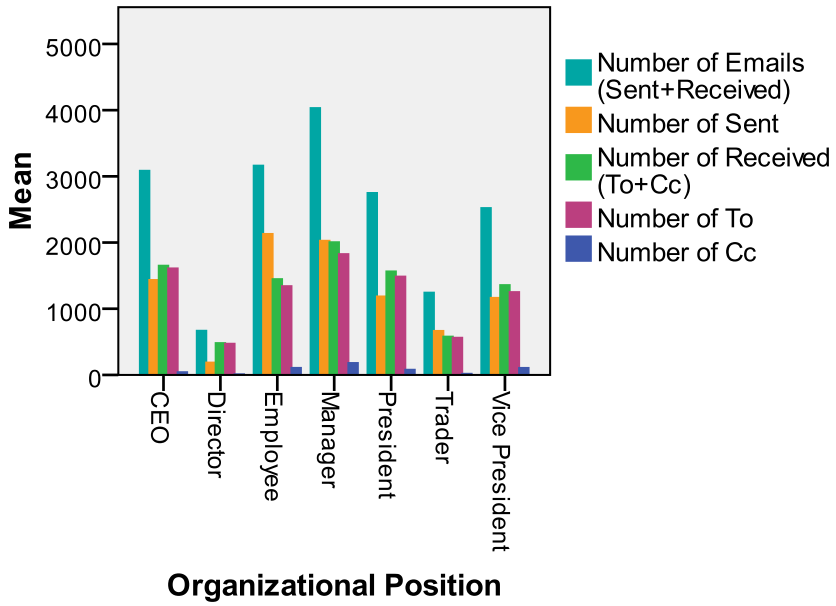
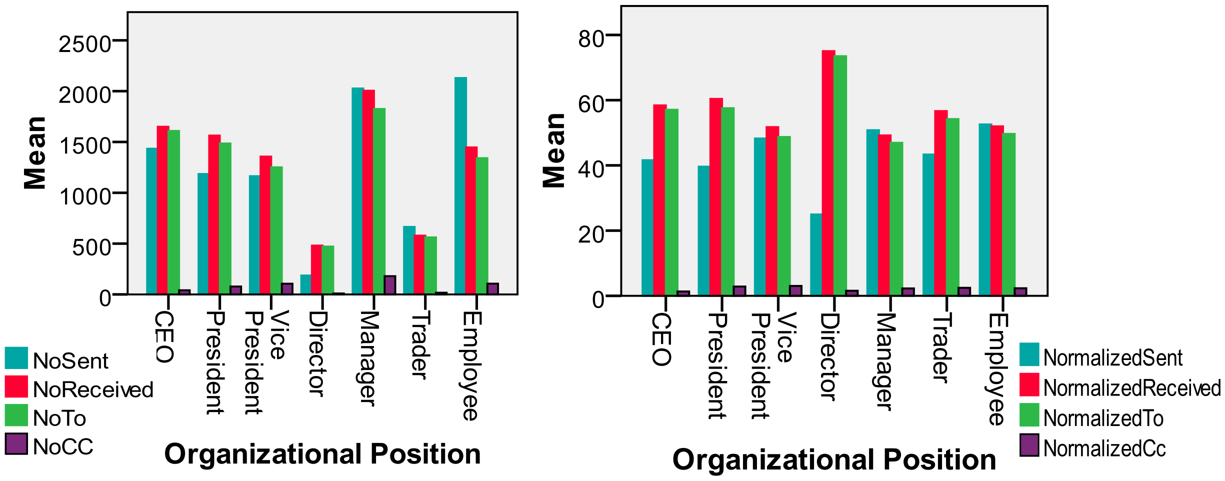


Figure 6-19. **Average number of exchanged, sent emails (as determined by *From*: field) and received emails (as determined by *To:* and *Cc*: fields) for *Enron* organizational positions from Jan 2000 to Dec 2001.**



A. Actual numbers

B. Normalized numbers

Figure 6-20. **Average number of sent emails (as determined by *From*: field) and received emails (as determined by *To:* and *Cc*: fields) for *Enron* organizational positions from Jan 2000 to Dec 2001. A) Actual numbers. B) Normalized numbers.**

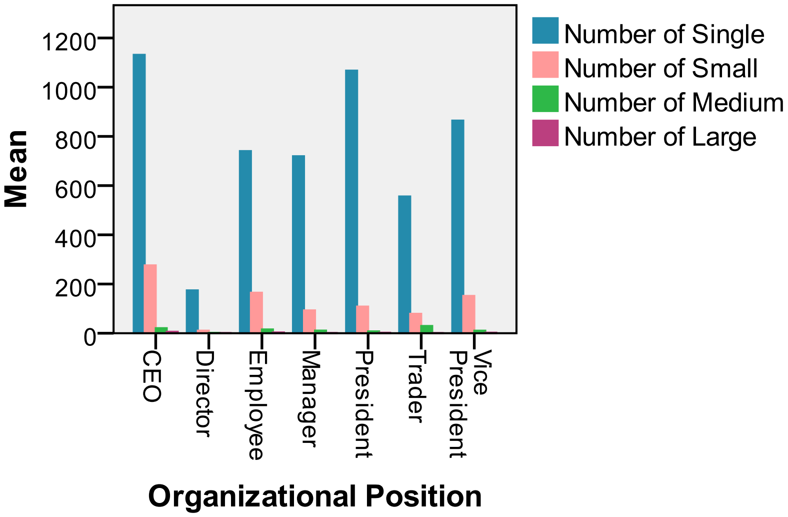
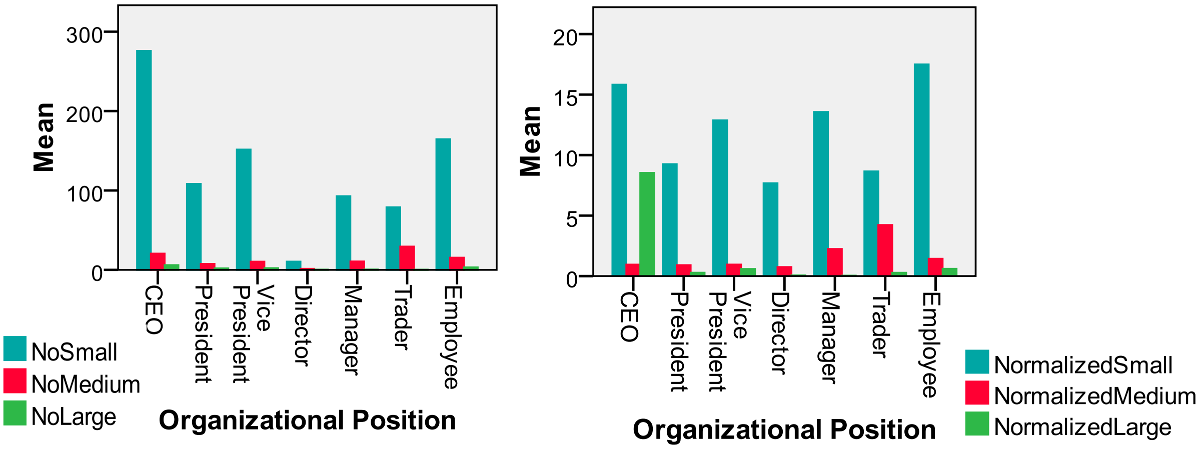


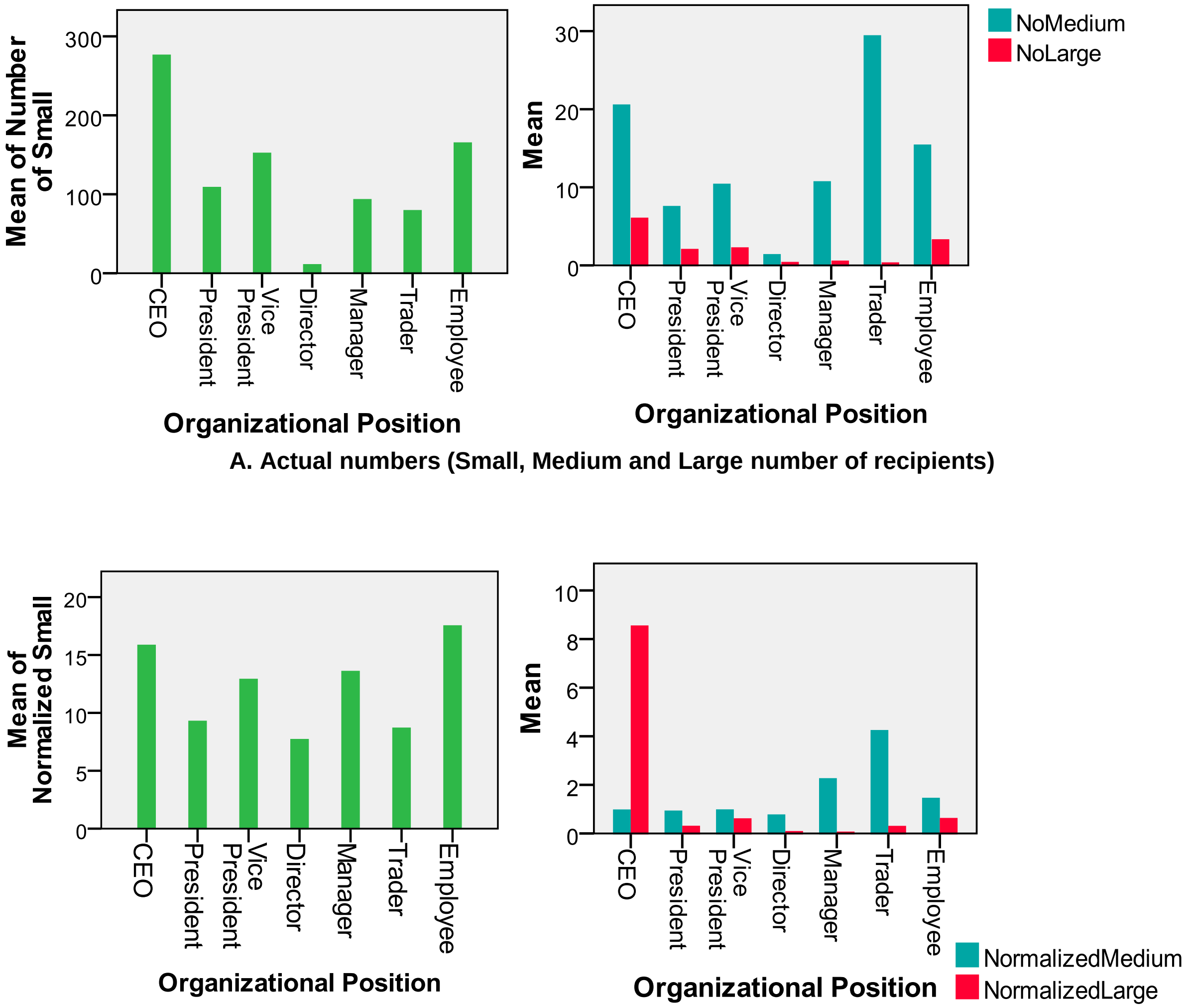
Figure 6-21. **Counts of sent email with the Single, Small, Medium and Large number of recipients.**



A. Actual numbers

B. Normalized numbers

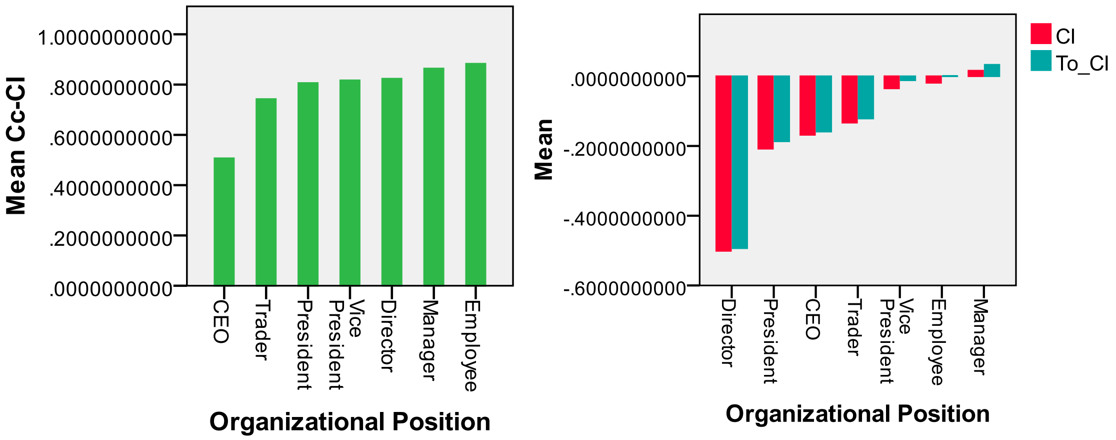
Figure 6-22. **Counts of sent email with the Small, Medium and Large number of recipients. A) Actual numbers. B) Normalized numbers.**



A. Actual numbers (Small, Medium and Large number of recipients)

B. Normalized numbers (Small, Medium and Large number of recipients)

Figure 6-23. Results of recipient count of sent emails. A) Average number of Small, Medium and Large recipient count for sent emails. B) Average of normalized number for Small, Medium and Large recipient count for sent emails.



A. Average of CI

B. Average of CI and To-CI

Figure 6-24. Contribution Index (CI) for organizational positions. A) Average of CI and To-CI. B) Average of Cc-CI.

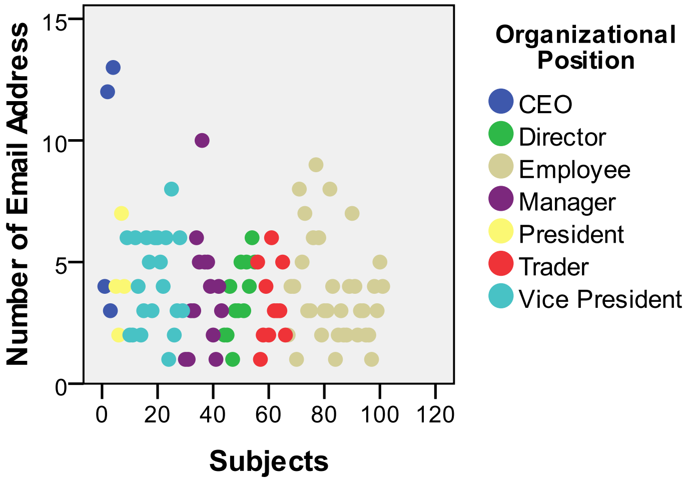
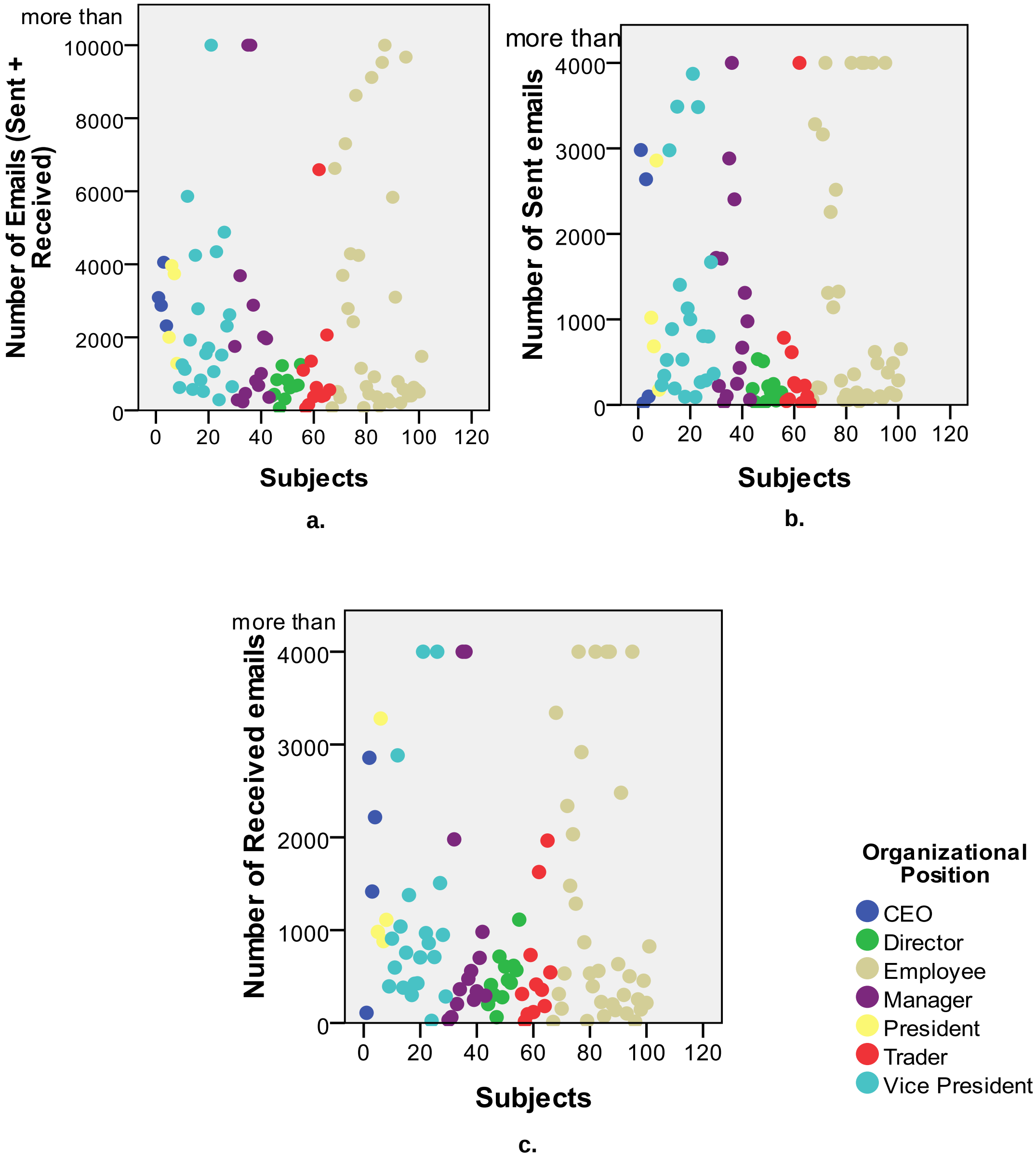


Figure 6-25. **Average number of email addresses for *Enron* workers from Jan 2000 to Dec 2001. Colour specifies the *Enron* organizational positions.**



a.

b.

c.

Figure 6-26. a. Number of emails. b. Number of sent emails. c. Number of received emails for *Enron* workers from Jan 2000 to Dec 2001. Colour specifies the *Enron* organizational positions.

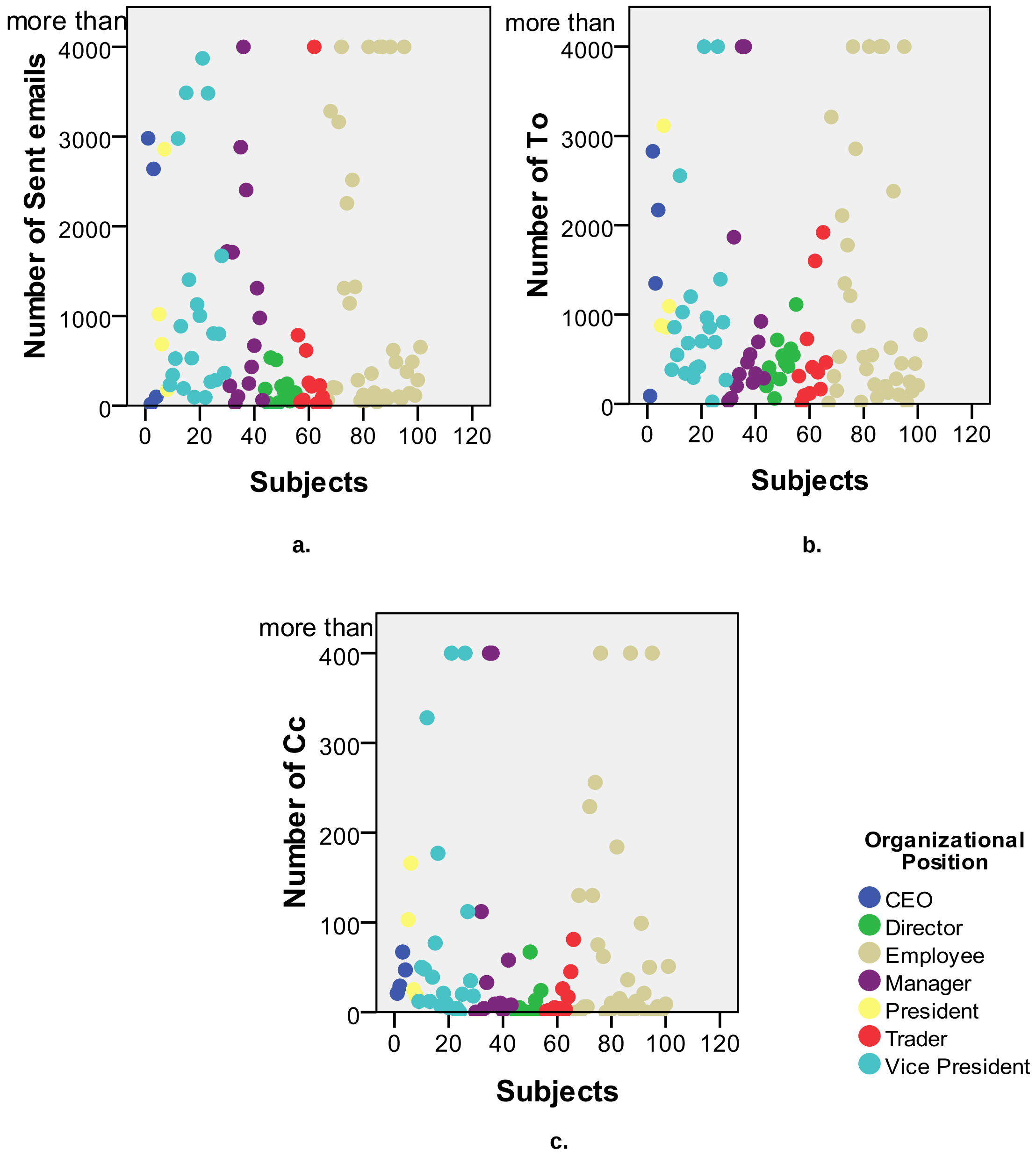


Figure 6-27. a. Number of sent emails as determined by *From*: field. b. Number of received emails as determined by *To:* Field. c. Number of received emails as determined by *Cc:* field for *Enron* worker from Jan 2000 to Dec 2001. Colour specifies the *Enron* organizational positions.

Appendix 3: User Study Documents

Consent Form

School of Interactive Arts
+ Technology

STREET ADDRESS
250 -13450 102 Avenue
Surrey, BC V3T 0A3
CANADA

MAILING ADDRESS
Simon Fraser University
Surrey
250 -13450 102 Avenue
Surrey, BC V3T 0A3
CANADA

phone: +1 778.782.7474
fax: +1 778.782.7478

Participant Consent Form
Feb 10, 2010

Dear Participant,

Hello, I am interested in your thoughts and experiences while investigating the *EmailTime* a new email visualization tool. This is a study that is conducted by Minoo Erfani Joorabchi, graduate student at Simon Fraser University. The purpose of this study is investigating and testing a new email visualization tool for displaying the content of email boxes over the course of time for analysis.

The University and those conducting this study subscribe to the ethical conduct of research and to the protection at all times of the interests, comfort, and safety of participants. This form and the information it contains are given to you for your own protection and to ensure your full understanding of the procedures, risks, and benefits described below.

Procedure: During the session, you will be asked to work with email visualization tool for analyzing the email dataset while I watch you. Once you complete, I will give you a questionnaire to fill about what you liked and did not like about those email visualizations. It would be great if you could try to remember any problems you see with the visualizations and the things you like as you work with those. You can do that through writing down your thoughts using paper and pen or just remember them. I hope you will enjoy your testing session. Please note that you do not have to participate if you do not want to. Also, at any point during the session, you can change your mind and stop if you do not wish to continue, no questions asked.
Risks to the participant, third parties or society: no risks involved
Benefits of study to the development of new knowledge: This study will result in investigating the new email visualization tool for analyzing the email dataset.
Freedom to Withdraw: Please note that you do not have to participate if you do not want to. Also, at any point during the session, you can change your mind and stop if you do not wish to continue, no questions asked.
Statement of Confidentiality: The name of the participants will be confidential and the data and the results will be stored in locked rooms and/or on password-secured storage devices.

Your signature on this form will signify that you have received a document, which describes the procedures, possible risks, and benefits of this research study, that you have received an adequate opportunity to consider the information in the document, and that you voluntarily agree to allow the minor named below to participate in the study.

I certify that I understand the procedures:

Name of participant: ____

and I know that I have the right to withdraw from the study at any time.

Signature of Participant____

Date____

and that any complaints about the study may be brought to:

Dr.Hal Weinberg, Director
Office of Research Ethics
Simon Fraser University
hal_weinberg@sfu.ca
778-782-6593
I may obtain copies of the results of this study, upon its completion by contacting the researchers named above by sending an email to Minoo Erfani Joorabchi (mea18@sfu.ca). Also our location is in School of Interactive Art and Technology, SFU, Surrey, BC.

Sincerely,
Minoo Erfani Joorabchi

Title: EmailTime Visualization Testing
Investigator Name: Minoo Erfani Joorabchi (mea18@sfu.ca)
Investigator Department: SFU School of Interactive Art and Technology

Pre-Questionnaire Form

Complete Name: --

Age: 20-24 25-29 30-34 35-40

Gender: Female Male

Are you colour-deficient?
If yes, please explain --------------------

How often do you work with the computer?
More than once a day Once a day Once a week Once a month Almost Never

Do you have the experience of working with visualization tools (e.g. Tableau, Inspire, graphs of Excel, etc)? ------- If yes, please name those--------------------

Are you familiar with the concept of visual analysis (a little)?

How many email accounts do you actively use?

How often do you log-in your email accounts?
Always More than once a day Once a day Once a week

Do you know what From, To, Cc, and Bcc mean in an email?

Which email viewers do you use most frequently? (e.g. Yahoo)

Have you tried search functions in the email viewers?

Signature-----------------------------

Tasks (for User Study)

The dataset that we are using for this experiment is a public real large dataset which belongs to Enron Corporation. Enron Corporation fell in October 2001. Think aloud is encouraged.

1. Time Comparison Scenario:

iv. This network displays all the emails that ***Matthew Lenhart*** had been involved in any sort of way (Sent or Received).

v. Compare the network for the period ***June-Dec 2000*** with the period ***June-Dec 2001***.

vi. In each of the cells in the following table, provide a sorting from 1 through 5 that indicates the strength of the properties for each time period (Enron Corporation fell in October 2001):

Period	**June-Dec 2000**	**June-Dec 2001**
Seems Normal	Not at all Totally 1 2 3 4 5	Not at all Totally 1 2 3 4 5
Has Large Gaps	Not at all Totally 1 2 3 4 5	Not at all Totally 1 2 3 4 5
Has Crowded Area	Not at all Totally 1 2 3 4 5	Not at all Totally 1 2 3 4 5
Has Large Emails	Not at all Totally 1 2 3 4 5	Not at all Totally 1 2 3 4 5

- Do you have any further observations?

2. Top Frequent Correspondents Scenario:

i. This network displays all the emails that ***Andy Zipper*** had been involved in any sort of way (Sent or Received).

ii. In the following table write the name of people that Andy Zipper has corresponded with :

- The ones who mostly *sent* emails to him during ***April to June 2001*** especially from ***late May*** to ***June***
- The types of their correspondence for each person (***private*** or ***general*** message)

Email Address of Correspondent	**Type of Correspondence**
1.	
2.	
3.	
4.	

- Do you have any further observations?

3. Email Comparison Scenario:

i. This network displays all the emails that ***Jeffrey Shankman*** had been involved in any sort of way (Sent or Received).

ii. Select three of his email addresses:

1) Jeffrey.shankman@enron.com, 2) a..shankman@enron.com, 3) Jeffry.a. shankman@enron.com

iii. Fill the following table regarding to:

- The ***role (Sender/Receiver)*** of each email address
- The ***activity*** level of each email address using the following scale:

Low				High
1	2	3	4	5

- The ***life period (duration)*** of each email address

Email Address of Jeffrey	Role	Activity	Duration
Jeffrey.shankman@enron.com			
a..shankman@enron.com			
Jeffry.a. shankman@enron.com			

- Any ***switching*** between email addresses Yes No

If yes > From:

To:

- Do you have any further observations?

4. Search Scenario:

i. This network displays all the emails that ***David Delaney*** (Enron **CEO**) had been involved in any sort of way (Sent or Received).

ii. Search for emails about a management committee with this subject "***ENA Management Committee***" in the network (search is case-sensitive).

iii. Fill the following table regarding:

- The ones who sent those emails
- The differences between the role of them (who seems to be the ***leader*** of the meeting and who acts as a ***secretary*** and reminding the meeting)

Sender	Role (Leader/ Secretary)

- The ***location***, ***time*** and ***type*** of the meeting

Location:

Time:

Type (draw circle): *Once* *Daily* *Biweekly*

- List three of the people who participate in the meeting

1.

2.

3.

- Do you have any further observations?

Post-Questionnaire Form

1. Set the task difficulty level in the following table:

Task	Easy	Medium	Hard	Fail or Gave up
Scenario 1				
Scenario 2				
Scenario 3				
Scenario 4				

2. As you may realized each scenario represents a functionality of *EmailTime*. Do you think you are able to recognize similar scenarios using *EmailTime*.

Scenario 1: *comparing time periods.*

Not at all Totally

1 2 3 4 5

Scenario 2: *finding the top frequent Correspondents of a person.*

Not at all Totally

1 2 3 4 5

Scenario 3: *comparing different email address of one person.*

Not at all Totally

1 2 3 4 5

Scenario 4: *finding out a scenario by searching.*

Not at all Totally

1 2 3 4 5

3. In the visualization plot, what are most liked (useful and convenient aspects) and disliked (troublesome or confusing aspects)? Why?

Like -------------------------

Dislike-----------------------

4. In the control panel what are most liked (useful and convenient aspects) and disliked (troublesome or confusing aspects)? Why?

Like -------------------------

Dislike-----------------------

5. Were used colours distinguishable? Yes No

Did you like the assignment of colours? Yes No, if No please explain ----------------

6. After working with *EmailTime* Visualization, I felt pretty competent now.

1	**2**	**3**	**4**	**5**
not at all		somewhat		very
true		true		true

7. I found the *EmailTime* Visualization environment easy to learn to use after introduction part.

1	**2**	**3**	**4**	**5**
not at all		somewhat		very
true		true		true

8. I found the *EmailTime* Visualization environment intuitive to learn before the introduction part.

1	**2**	**3**	**4**	**5**
not at all		somewhat		very
true		true		true

9. I found that I had to really concentrate to learn how to use *EmailTime* Visualization.

1	**2**	**3**	**4**	**5**
not at all		somewhat		very
true		true		true

10. Any more comments?

Signature -------------------------------------

REFERENCE LIST

1. X. Yu, P. Liu, M. Alkandari and M. Perez-Quinones, “Visualizing a personal social network of email archives for re-finding”. *In Proceedings of the 2nd Invitational Workshop on Personal Information Management at SIGIR*, 2006.

2. J. Heer and D. Boyd, “Vizster: visualizing online social networks”. *Proceedings of Symposium on Information Visualization,* Minneapolis, MN: IEEE Press, pp. 33-40, 2005.

3. J. Heer, “Exploring Enron, a sketch of visual data mining of email”. In *Email Archive Visualization Workshop*, HCIL University of Maryland*,* 2005.

4. D. Fisher and P. Dourish, “Social and temporal structures in everyday collaboration”, *Proceedings of the ACM Conference on Human Factors in Computing Systems CHI*, pp. 551-558, 2004

5. P. Gloor, S. Niepel and Y. Li, “Identifying potential suspects by temporal link analysis”. MIT CCS working paper, 2006.

6. A. Chapanond, M. S. Krishnamoorthy and B. Yener, “Graph theoretic and spectral analysis of *Enron* email data”, *Computational & Mathematical Organization Theory*, v.11 n.3, pp. 265-281, 2005.

7. J. Diesne and K. M. Carley, “Exploration of communication networks from the *Enron* email corpus”, *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining*. Newport Beach, pp. 3-14, 2005.

8. A. Perer and B. Shneiderman, “Balancing systematic and flexible exploration of social networks”. In *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, n. 5, pp. 693-700, 2006.

9. M. Mandic and A. Kerne, “FaMailiar - intimacy-based email visualization”. In *IEEE Symposium on Information Visualization*, pp. 14-14, 2004.

10. M. Mandic and A. Kerne, “Using intimacy, chronology and zooming to visualize rhythms in email experience”. In *CHI ‘05,* pp. 1617-1620, 2005.

11. W. J. Li, S. Hershkop and S. J. Stolfo, “Email archive analysis through graphical visualization”. In *Proceedings of ACM CCS VizSEC/DMSEC ‘04,* pp. 128-132, 2004.

12. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern and C. W. Hu, “Behavior profiling of email”. In *Proceedings of ISI’03*, pp.74-90, 2003.

13. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern and C. W. Hu, “A behavior-based approach to securing email systems”. In *Proceedings of Springer Verlag*, 2003.

14. J. Stolfo, C. W. Hu, W. J. Li, S. Hershkop, K. Wang and O. Nimeskern, “Combining behavior models to secure email systems”, *CU Tech Report*, 2003.

15. A. Perer, B. Shneiderman and D. Oard, “Using rhythms of relationships to understand email archives”. *Journal of the American Society for Information Science and Technology*, pp. 1936-1948, 2006.

16. A. Leuski, D. W. Oard and R. Bhagat, “eArchivarius: accessing collections of electronic mail”. In *Proceedings of SIGIR '03*, pp. 468-469, 2003.

17. F. Viegas, B. Fernando, S. Golder and J. Donath, “Visualizing email content: portraying relationships from conversational histories”. In *Proceedings CHI*, pp. 979-988, 2006.

18. F. Viegas and J. Donath, “Social network visualization: can we go beyond the graph?” In *Proceedings of CSCW’04*, vol. 4, pp. 6-10, 2004.

19. F. Viegas, D. Boyd, D. H. Nguyen, J. Potter and J. Donath, “Digital artifacts for remembering and storytelling: posthistory and social network fragments”. In *Proceedings HICSS-37*, pp. 10, 2004.

20. F. Viegas and M. Smith, “Newsgroup crowds and authorlines, visualizing the activity of individuals in conversational cyberspaces”. In *HICSS-37*, pp. 10, 2004.

21. S. L. Rohall, D. Gruen, P. Moody and S. Kellerman, “Email visualizations to aid communications”. In *Proceedings of IEEE Symposium on Information Visualization*, pp. 12-15, 2001.

22. B. Kerr, “Thread arcs: an email thread visualization”. In *IEEE Symposium on Information Visualization*, pp. 211-218, 2003.

23. G. Venolia and C. Neustaedter, “Understanding sequence and reply relationships within email conversations: a mixed-model visualization”. In *Proceedings of CHI*, vol. 5, pp. 361-368, 2003.

24. A. Perer and B. Shneiderman, “Beyond threads, identifying discussions in email archives”. *IEEE Symposium on Information Visualization*, 2005.

25. J. Heer, S. K. Card, and J. A. Landay. “Prefuse: a toolkit for interactive information visualization”. In *CHI 05*, pp. 421–430. ACM Press, 2005.

26. J. Donath, “Visualizing Email Archives” (*draft)* 2004.

27. W. W. Cohen, CALD, CMU. October 2004, from http://www-2.cs.cmu.edu/~enron/

28. B. Klimt and Y. Yang, “Introducing the Enron corpus”, *First Conference on Email and Anti-Spam* (CEAS), Mountain View, 2004

29. http://alumni.media.mit.edu/~fviegas/projects/mountain/

30. Authorised Traders List5_11_01.pdf (http://docs.google.com/viewer?a=v&q=cache:ksNeNA7bX08J:ferc.aspensys.com/FercData/Miscellaneous%2520cd's/Box005/Response%2520to%2520Request%252015/RAC/Compliance/Authorized%2520Trader%2520Lists/Authorised%2520Traders%2520List5_11_01.pdf+enron+employee+list+Authorised+Traders+List5_11_01.pdf&hl=en&gl=ca&pid=bl&srcid=ADGEESh6hNiaUZ1rnfeACh8Je2t_9yoPAv1gHZIHx-Gj6u4zubP5SJalbbe80FqbUymqVXyoJjMc-0iQtow2OlwOSZivHF0q-U8Q8p95pSPVUIuH-GkSqHkluT7q46lGcxJSPVQue115&sig=AHIEtbSPt7pfGxONUsZsZ_0yHoAmfOoE7w)

31. SPSS, http://www.spss.com/

32. Histogram code, http://webfoot.com/code/

33. M. Erfani Joorabchi, J. D. Yim, C. D. Shaw, “EmailTime: Visual Analysis and Statistics for Temporal Email”, *Visualization and Data Analysis,* pp. 7868-24, 2011*.*

34. M. Erfani Joorabchi, J. D. Yim, M. Erfani Joorabchi, C. D. Shaw, “Enron Case Study: Analysis of Email Behavior using *EmailTime”, IEEE VisWeek VAST Poster,* 2010*.*

35. M. Erfani Joorabchi, J. D. Yim , C. D. Shaw, “EmailTime: Visual Analytics and Statistics of Email*”, IEEE VisWeek VAST Poster,* 2010.

36. M. Erfani Joorabchi, J. D. Yim, C. D. Shaw, “EmailTime: Visualization of the Temporal Email,” *Grace Hopper Celebration of Women in Computing,* pp. 231-238, 2010.