

EmailTime: Visual Analytics of Emails

Minoo Erfani Joorabchi, Ji-Dong Yim and Christopher D. Shaw

Simon Fraser University

ABSTRACT

Although the discovery and analysis of communication patterns in large and complex email datasets are difficult tasks, they can be a valuable source of information. This paper presents *EmailTime*'s capabilities through several examples. *EmailTime* is 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. We suggest that integrating both statistics and visualizations in order to display information about the email datasets may simplify its evaluation.

KEYWORDS: Email, Enron, *EmailTime*, Email Correspondents, Visual Analysis.

Index Terms:

1 EMAILTIME OVERVIEW

Our system, the *EmailTime* [1], visualizes the communication activities found in a collection of the emails over a course of time. The original email dataset is from the Enron Email Corpus [2]. To apply the archive to our visualization, we prepared the dataset from either one's or multiple users' mailboxes, and filter off the emails out of the time period we are interested in.

EmailTime is 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 put time as a primary variable of interest, and plot emails along a time line. *EmailTime* helps email dataset explorers interpret archived messages by providing zooming, panning, filtering and highlighting etc. To support analysis, it also measures and visualizes histograms, graph centrality and frequency on the communication graph that can be induced from the email collection.

In the visualization, the horizontal dimension depicts time and each email message is displayed as multiple visual aliases painted in three different colors, which are a black circle for the Sender (or email address in *From* field), blues for *To* recipients and greens for *CCs*. Sent emails have different sizes to represent the number of recipients (*To* + *CC*) of the email message.

Figure 1 displays the plot of a small network. As mentioned above a message draws multiple circles in three different colors. The message #1 is sent by Aaron to Beth, the message #2 is sent by

Chris to Aaron, and copied to Beth, the message #3 is sent by Beth to Aaron and copied to Chris and David. Therefore the black circle gets larger as the numbers of recipients increased.

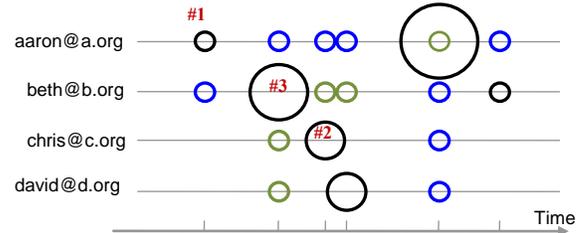


Figure 1. Plotting email messages. A message draws multiple circles in three different colors; black for Sent, blue for Received as *To*, and green for Received as *Cc*. The size of a Sent node represents the number of recipients. (e.g. the message #3 is sent by Chris to Aaron, and copied to Beth.)

2 EMAILTIME CAPABILITIES

EmailTime visualization enables the analysts to infer underlying implications from observation and thus helps them move from observation to investigation. Using the system the analysts/users can:

- Compare the activity of one or more email addresses in the particular time periods (Figure 2).
- Temporal gaps in email addresses' activities are obvious. Therefore if the analysts/users know some events they can easily relate those to them (e.g. holidays, trips..., Figure 2-B)
- All the top frequent correspondents (user clique) who have sent/received relatively a large amount of emails to/from the focused one are easily distinguished – any bunch of emails that are rendered in a row and can be highlighted by plot (Figure 3).
- Types of correspondence (general or private messages) which means that to how many people that email address sent messages (realized by size of messages, Figure 3).
- Compare the email addresses of one individual (Can select one or more email address from the email list and see the related emails in the plot, Figure 4).
- Some changes such as switching from one email address to another one can be recognized (Figure 4).
- Role of each email address (sender, receiver, both) and the behaviour pattern of each email address (e.g. start and end time, more or less activity) by plot and histogram (Figure 4).

2.1 Examples

Figure 2 displays the datasets of four randomly selected Enron Employees in two different time period, June-Dec 2000 and 2001. These two time periods are completely different. The first one (Figure 2-A) is more uniform than the second one whereas the second one (Figure 2-B) has the gap period and crowded part at the end of 2001 which can be the result of Enron's fall. In the first period only these four employees sent the emails but in the second period many others sent emails to the network as well.

SIAT, Simon Fraser University, 250-13450 102 Avenue,
Surrey, Canada

Send correspondence to M.E.J.

M.E.J.: E-mail: meal8@sfu.ca

J.D.Y.: E-mail: jdyim@sfu.ca

C.D.S.: E-mail: shaw@sfu.ca

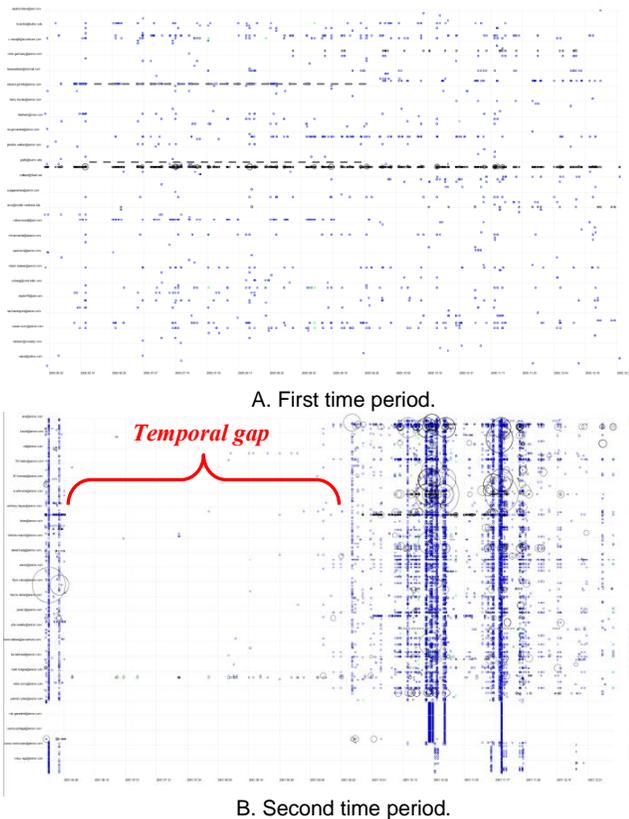


Figure 2. The datasets of four randomly selected Enron Employees; Albert Meyers, Judy Townsend, Matthew Lenhart and Susan Pereira are plotted in the visualization in two different time periods. A) June-Dec 2000. B) June-Dec 2001.

Since our visualization displays a large number of emails and draws multiple instances for a message, the very basic and important filter in the system is the *Node Type Selector* applied to the three types of email node - *Sent*, Received as *To*, and Received as *Cc*. In **Error! Reference source not found.**3, where *To* and *Cc* nodes are filtered out, we can clearly see different patterns of sending activities. For example in Figure 3, Sender 2 (Kay

Chapman) frequently spread emails to many people, while the Sender 5 (David Delaney) usually talked to a small group of people. Also, we can recognize three of active email senders. In similar ways, *To* and *Cc* emails make interesting patterns each.

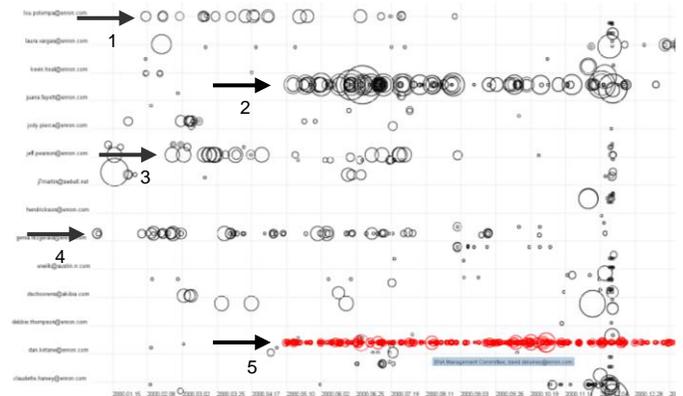


Figure 3. A subset of *Sent* emails (or *From* nodes) within the example dataset in year 2000. The highlighted nodes at the bottom (in red) are the messages that the actor david.delaney@enron.com sent.

The *Email List Tab* in the control panel contains the list of all email addresses found in the dataset and filters the plot visualization to only display the selected ones. In Figure 4 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 in March 2001. Jeffrey Shankman played different role during each time period. He was a sender in his first and last activities and was a receiver in his second activities. He didn't actively use other email addresses.

REFERENCES

- [1] Mino Erfani Joorabchi, Ji-Dong Yim, Christopher D. Shaw, "EmailTime: Visualization of the Temporal Email," Grace Hopper Celebration 2010.
- [2] W. W. Cohen, CALD, CMU. October 2004, from <http://www-2.cs.cmu.edu/~enron/>

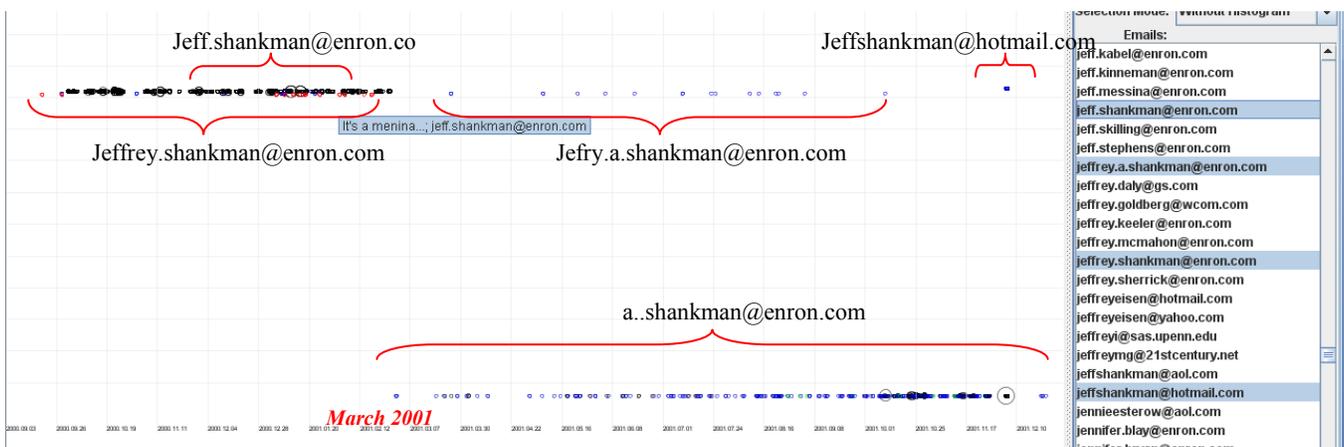


Figure 4. Emails sent or received by the five email addresses of Jeffrey Shankman, the President of Enron Corporation.