Email Social Network Over Time

Cheng Ma, Yulin Shen, Hao Dong, Bo Wang
{cm4102, ys61, hd846, bw1332}@nyu.edu

Project page (on Github):https://github.com/NYU-CS6313-Fall16/Email-Net-Over-Time-20
Video: https://vimeo.com/196754153
Working demo: https://bw1332.github.io/VisualizationProj/

What is the problem you want to solve and who has this problem?

Email is one of the most high frequently-used way to communicate and it is also the most official way to get touch with others through the Internet. How does a group of people use e-mail to enlarge their social network?

It is easy to show who sent or received e-mail at what time or what is the main topic between two people by hundreds of thousands of rows of raw data. However, it is much more difficult for us to detect the close relationship among these attributes.

Our main purpose of this project is to display the volume and content change during specified time intervals and try to figure out interesting trend or pattern within this group of people.

What are the driving analytical questions you want to be able to answer with your visualization?

- **What does the communication frequency between one person and another look like and how does it change during a certain period of time?**
  The communication frequency reflects the relationship among a number of persons. People who contact each other frequently all the time may have a close relationship. This question helps users to know whether one have a good relationship with someone else and how the relationship changes. It will also lead users to exploring the reason why the relationship changes.

- **What are the popular topics between two persons ?**
  The topics in the emails between two persons shows why they contact each other. They may have common interests or just work for the same task. Topics extracted from subjects can show users what two persons talk about and also give users a way to delve interesting stories between the two persons.

- **How many emails are sent or received within this social network during a certain period of time and how does the amount change over time?**
  The amount of emails sent or received within the community reflects the activity of the community. By visualizing this, users can know in which period of time the community is quite active.

- **How many people does one person contact during a certain period of time averagely?**
  The average amount of contacts one person owns reflects how people organize in the community. Collaborated with the question related to amount of emails of the whole community, users can know whether most people in the community are connected or they are divided into several groups and only contact those who within the same group.

What does your data look like? Where does it come from? What real-world phenomena does it capture?

The raw data comes from Enron Email Dataset collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). The raw data is a json file that includes email address and name of both senders and receivers, time, content that extract from Enron Email Dataset. To solve the problems we proposed, we focus on the attributes as follows:

Attribute Name	Attribute Type	Meaning	Values Range	Derived?
Email Address	categorical	Email address of a person	All the email address	no
Name	categorical	Name of a person owning an email	All the names of people	Yes , by reformatting the email address
Volume of email sent	quantitative	The number of emails one person has sent	All emails sent [0, 57]	Yes, by calculating the amount of emails one sent
Volume of email between two persons	quantitative	The number of emails two persons have sent to each other	All email passed between two persons [0, 117]	Yes, by calculating the amount of emails one received
Month	ordinal	Month	1~12	Yes, by reserving the month

		Identification		in which an email was sent only.
Topic	categorical	The high frequency words derived from emails' subjects	All meaningful words in the subjects of emails	Yes, by obtaining the words and their frequency from the contents of emails
Contacts	categorical	People one sent emails to or received emails from	All the people	Yes, by collecting the people who sent emails to or received emails from a certain person

What have others done to solve this or related problems?

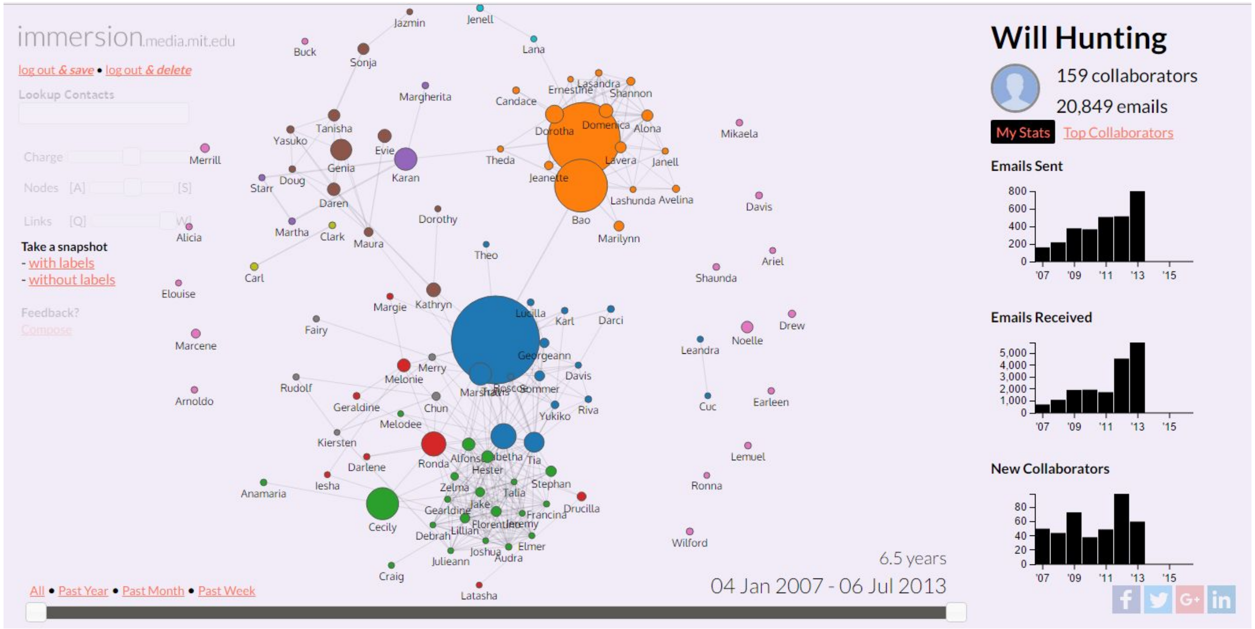
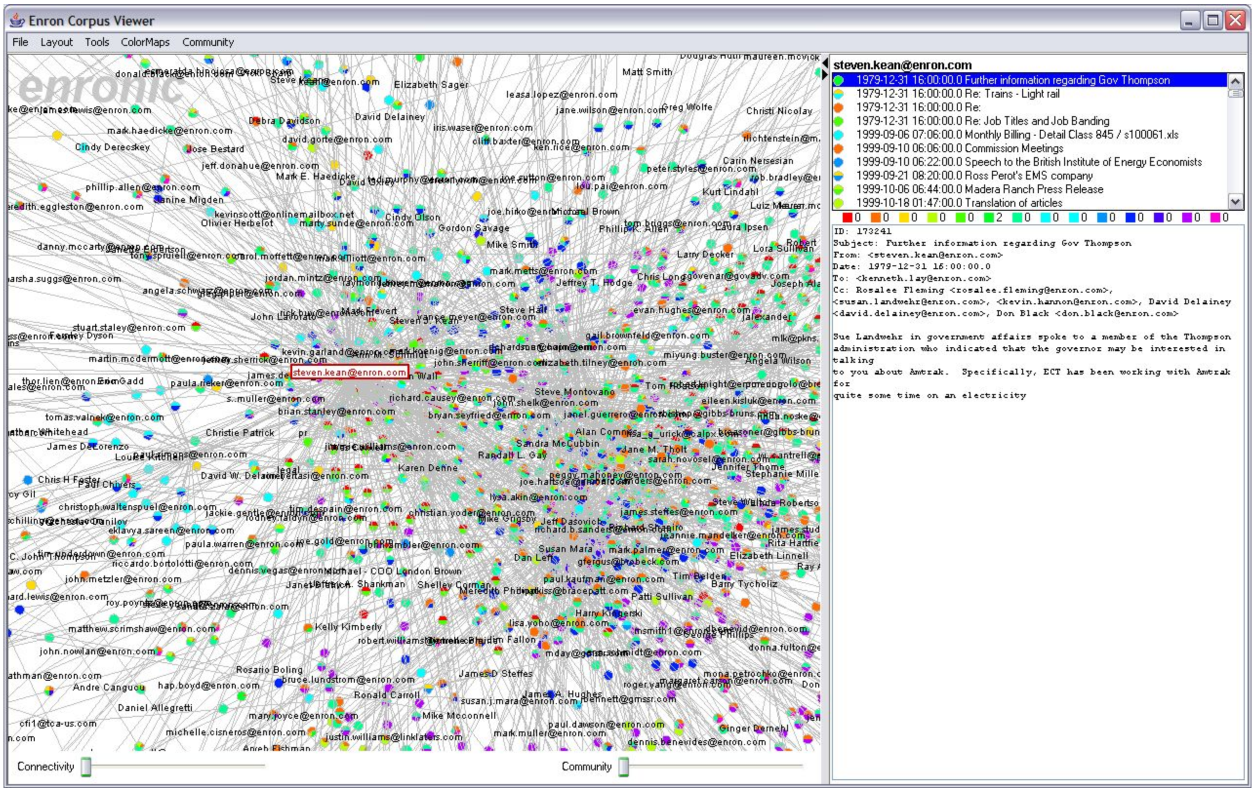


Figure 1

https://immersion.media.mit.edu/demo

As a people-centric project, this project is to dive into a person's history of email and establish a social network based on that history. This project inspires us how to show the social network among people and helps us to design the main chart in our initial mockup.



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 2

http://hci.stanford.edu/~jheer/projects/enron/v1/

This project attempts to use ANLP (Applied Natural Language Processing techniques) to explore Enron email data which is also the dataset we use. A social network visualization is implemented in this project, which inspires us how to show the email traffic among individuals.

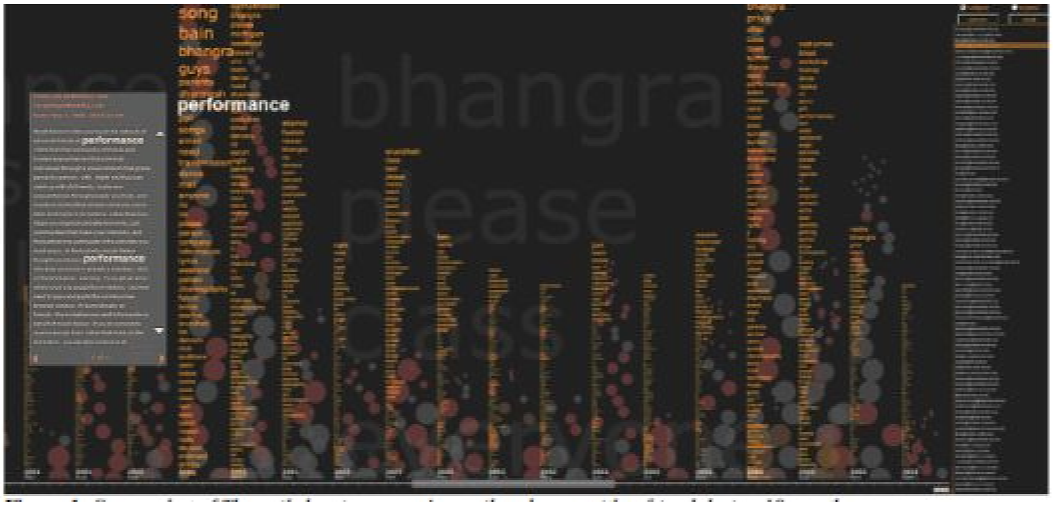


Figure 3

http://www.redlog.net/papers/viegas_themail.pdf

This project portrays relationships using the interaction histories in email archives. It uses keyword in the email contents to show the change of relationships among people. It inspired us that extracting keywords from emails is a good way to show the content of emails change over time.

Design Iterations

Describe how your mockup evolved over time, what kind of visualizations you tried, what worked and did not work, etc. For every iteration/design, please insert an image and describe how to read them. Use this section to showcase your design thinking and decisions that went into building the final visualizations. This is one of the most critical part of the project.

1. Initial Mockup

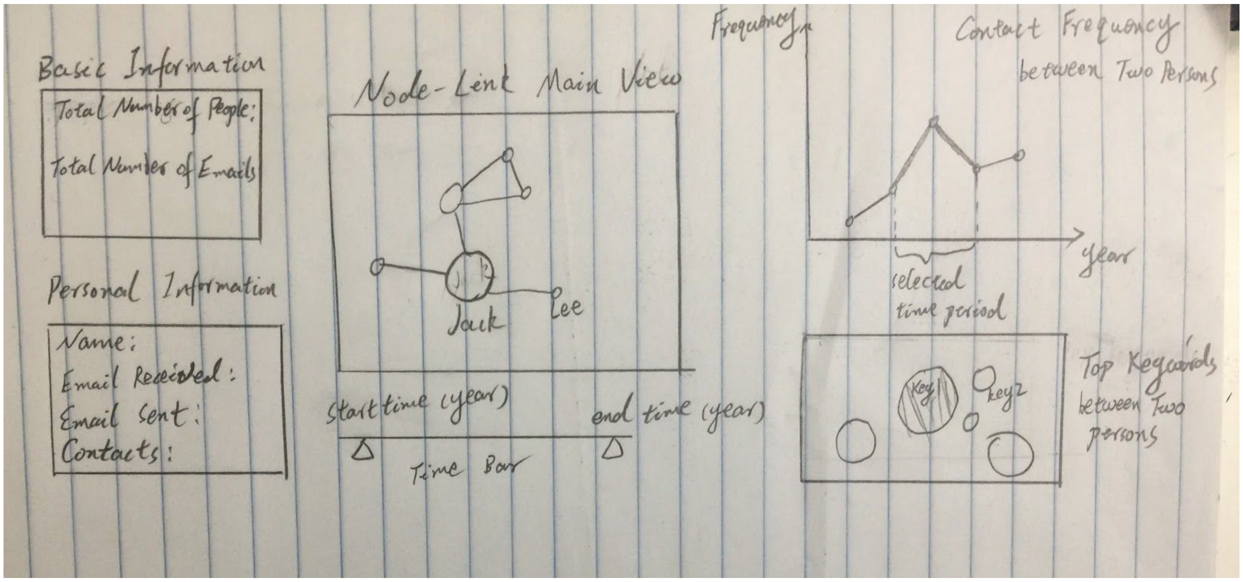


Figure 4

The mockup is designed to work as following:

- The main view represents the social network of email users in the given dataset during a certain period of time, which includes a time bar under the main view that allows the user to change the duration of the time.
- Different charts within this user interface is closely connected.
- When a certain period of time is selected by the user, the amount of email received and sent and the total number of people in the main view will be calculated and updated on the left side of the main view.
- When a person in the main view is selected, the amount of email he received and sent and the total number of people that he or she has communicated during the selected period of time will be shown in the list on the left of the main view.

- When the link between two persons is chosen, the line chart will show how the contact frequency , which is represented by the amount of sent and received mails, between two persons changes over time. The frequency during the period of time selected by user will be highlighted at the same time. The keyword graph which displayed on the right of the main view will be updated simultaneously. Each keyword will be encoded with color hue and the frequency of the keywords is linear with the radius of a circle.

However, this mockup did not work well:

- The link-node chart may contains more than 40 nodes and a great number of links among the nodes. As a result, the chart is quite complicate and it is hard to recognize the link between two nodes.
- When the time period changes, the position of each node is changed. As a result, users have to rely on their memory to find how the chart changes.
- The whole design focus too much on the information between persons. However, we need to show how the volume and contents of emails within the whole community changes.

2. Project Update

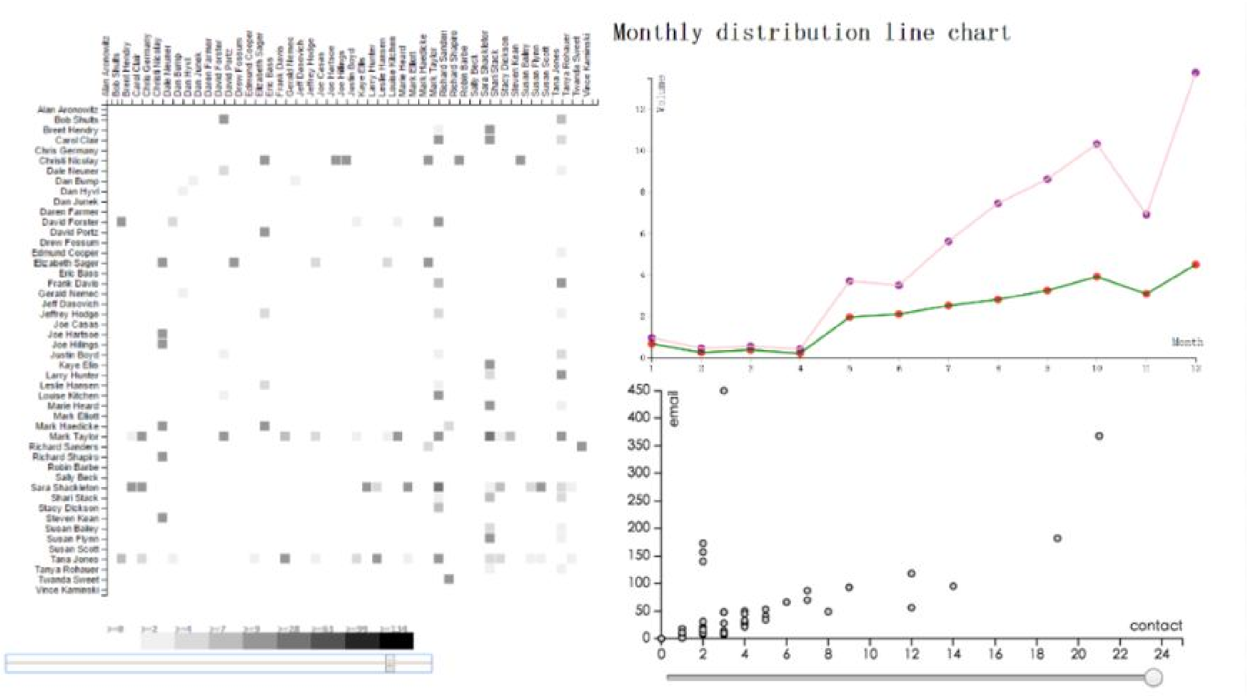


Figure 5

This design works as following:

- The whole page consists of three graphs: a heatmap, a line chart and a scatter plot chart.
- The heatmap shows the contact frequency, represented by the volume of emails sent and received, between two persons within the community in a month.

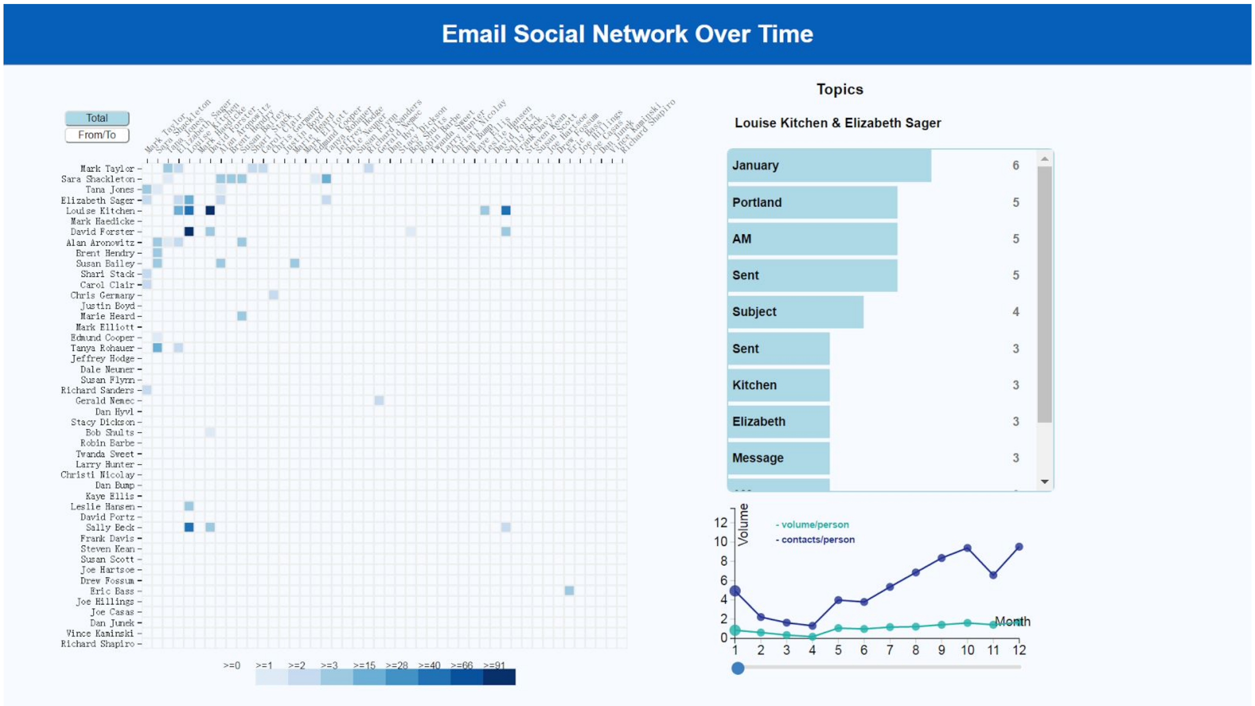
- The line chart shows the average amount of emails a person sent and received and the average amount of contacts one have in each month.
- The scatter plot chart shows the relationship between one’s amount of contacts and amount of emails he or her sent and received in a month. Each circle represents a person. When the mouse is moved on a circle, the name of the person will be shown.

However, this mockup did not work well:

- The three charts are not connected. When the user explores one of them, the other two will not change based on the user’s action.
- The scatterplot chart is redundant. All the information it shows can be extracted from the heatmap
- No graph shows how the contents of emails changes over time

Final Visualization

Insert images of your final visualization, describe how to read it



Email Social Network Over Time

Figure 6

The whole page consists of three parts.:

- The main chart is a heatmap on the left hand side with two buttons on the top left. The buttons are used to switch between two different modes. When click on the one that labeled *Total*, each cell in the heatmap represents the total volume of email sent and received. When click on the one that labeled *From/To* , the x-position represents those who sent the email and y-position represents those who received mail. The darker the cell, the more email sent and received. When mouse over the cell, both x-position and y-position as well as the cell will be highlighted. And when click on the cell, a tooltip will turn up to show the email address of these two people and corresponding volume.

- On the right top is a bar chart that represents the high frequency topics between two people. When click on the cell in the heatmap, the bar chart will simultaneously change to the topics between the selected people. The topics are ranked based on the frequency. When nothing turns up in the bar chart, this means that there is no communication between these two people within this month.
- On the right bottom is a static line chart that used to represent the general change of the whole dataset. The green line displays the volume of email per person and the dark blue line displays the contacts per person. The line chart is attached with a time bar below that enables user to change the dataset through different month. When the month is changed, both the heatmap and the bar chart will change at the same time.

Findings

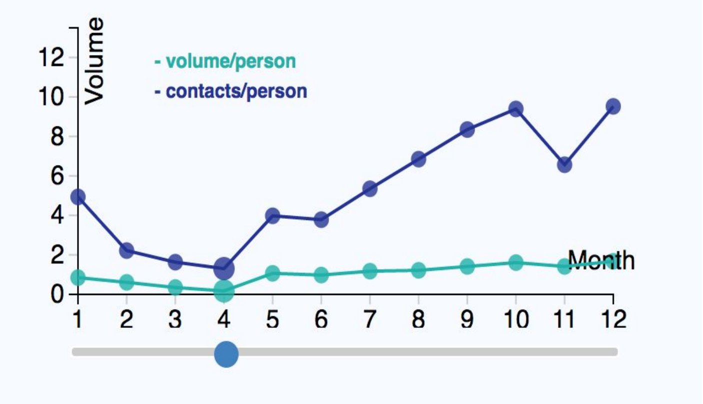


Figure 7

- **What is the trend of the social network?**
According to the Figure 7, the contacts/person line slightly dropped down from January to April and kept increasing from April to December. But volume/person didn’t change much. So there were more people joined the network during the year. But the role they played in their communications might not change much.

- **What does the whole community look like?**
According to the Figure 7, every person in the community contacts less than two other people averagely each month. It can be inferred that peoples are divided into several small groups and seldomly contact those who outside their small group.

- **Who sent and received the largest amount of emails in the whole time? And whose is the smallest?**

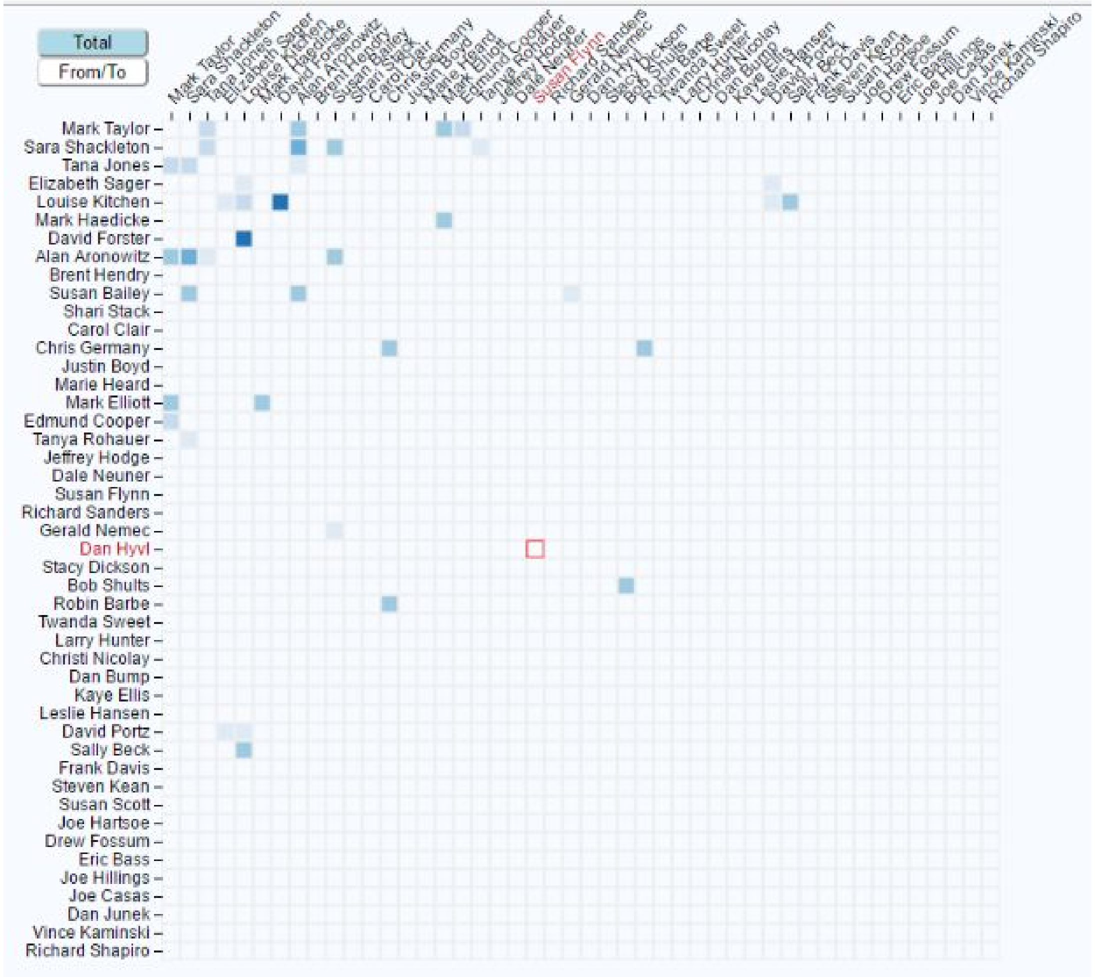
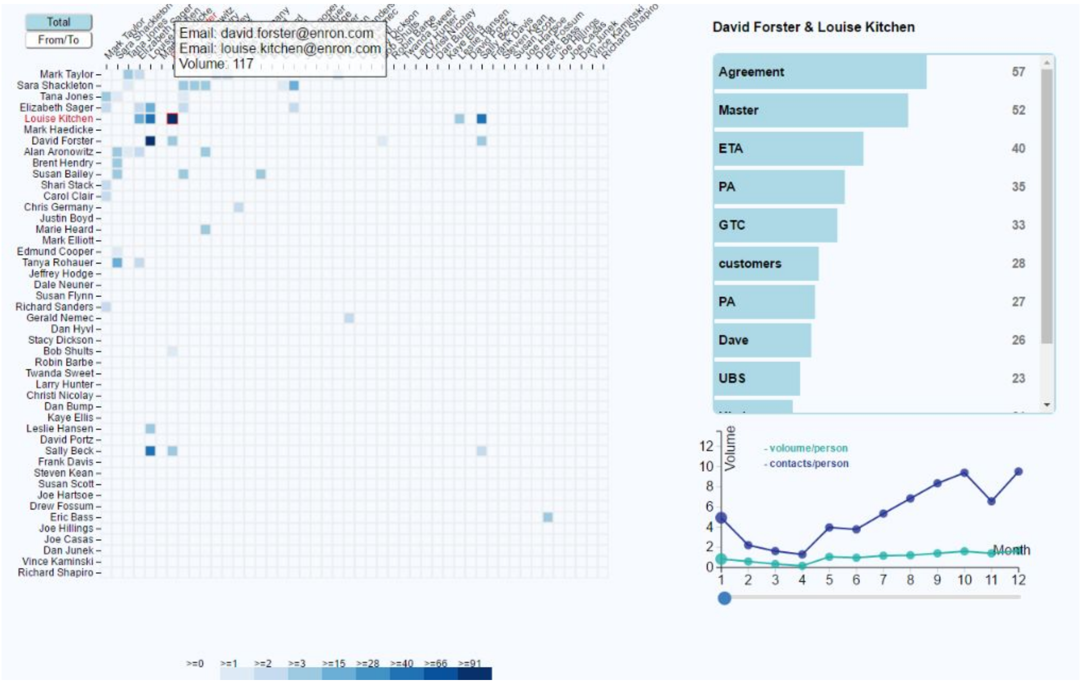


Figure 8

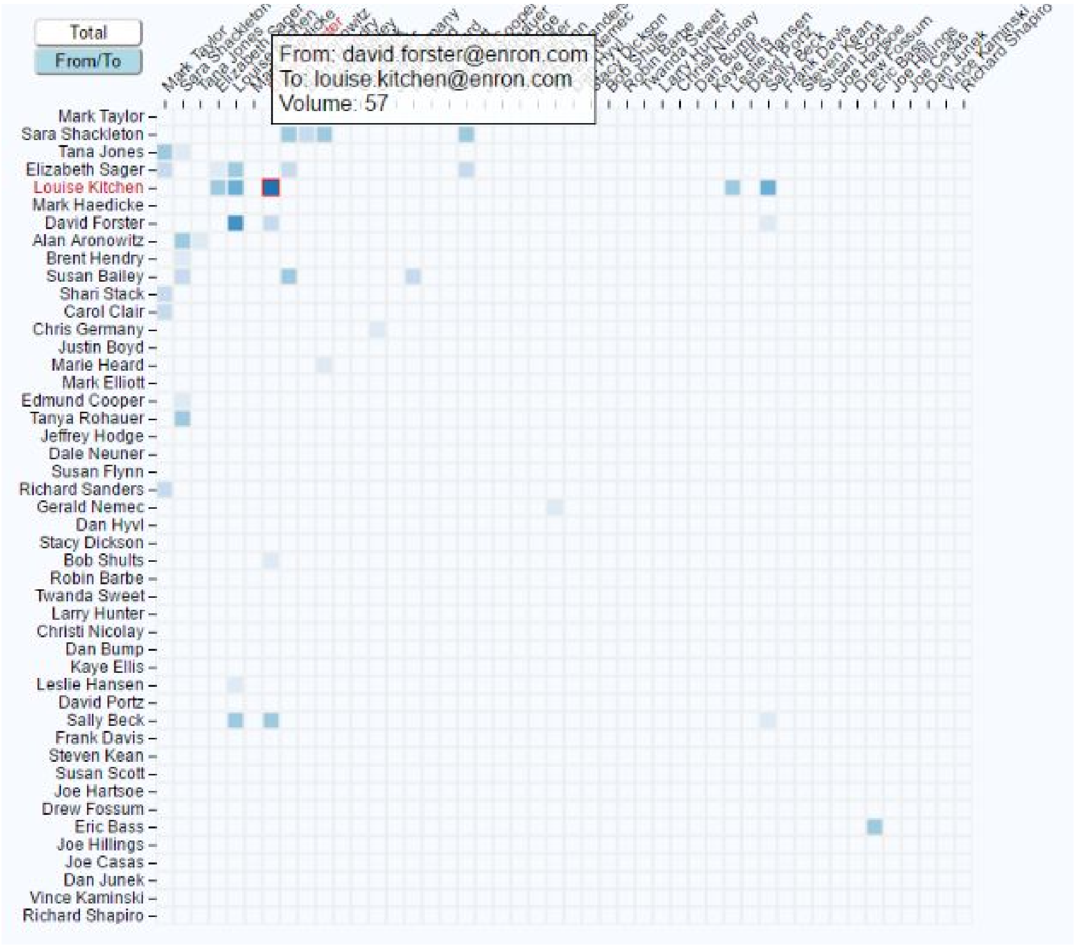
The names in either x Axis or y Axis of the heatmap are sorted according to the amount of emails one received and sent in the whole time. Therefore, Mark Taylor is the one who received and sent the most emails while Richard Shapiro is the one received and sent the fewest emails.

- **Which two persons contact most frequently in January?**



Email: david.forster@enron.com
Email: louise.kitchen@enron.com
Volume: 117

Figure 9



From: david.forster@enron.com
To: louise.kitchen@enron.com
Volume: 57

Figure 10

According to Figure 9 and Figure 10, David Forster and Louise Kitchen contact most frequently in January. They totally sent 117 emails between each other. Among the 117 emails, 57 emails were sent by David Forster while the rest 30 emails were sent by Louise Kitchen.

In fact, more interesting things could be found in our visualization. For example:

- **The progress of the Sara & Brent‘s development**
  From our visualization, the communications between Sara Shackleton and Brent Hendry increased heavily from July to September and drop from October to December. Why did that happen? Probably we could find the reason from the topic bar chart.

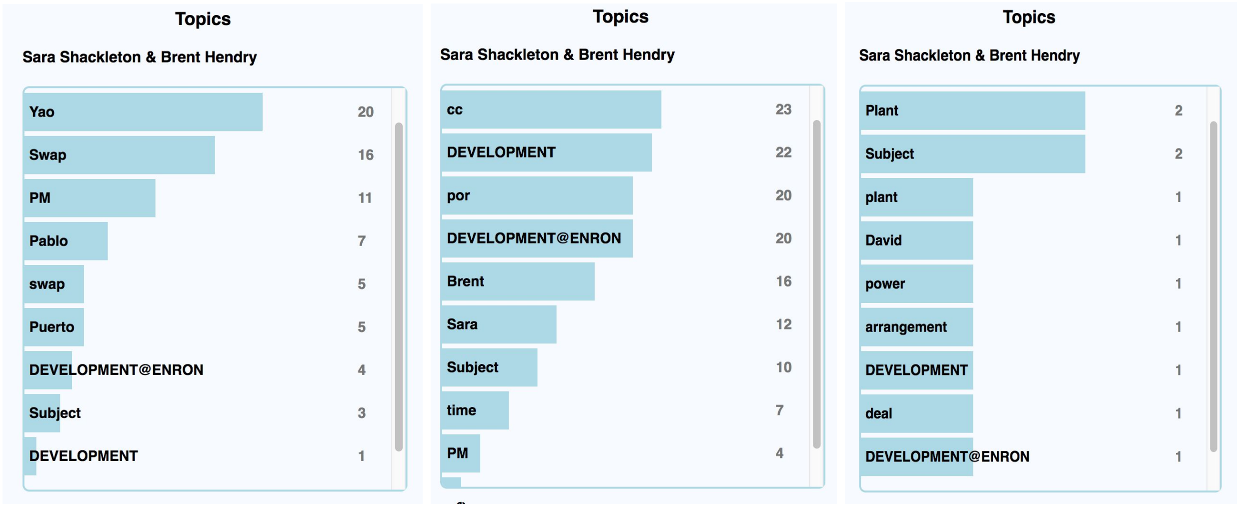


Figure 11

  Figure 11 shows topics between Sara and Brent in July, September and December respectively. We can find the frequency of topic “Development” increased from 5 to 42 from July to September but decreased smoothly to 2 in December. Therefore, probably their development had huge progress in Sep and came to the end in the end of the year.

- **What is the different role they played in the network?**
  If we look deeper into these two people, we could find more things such as the roles they probably played in this network. During the whole year, Sara seemed to have more different contacts in each month while Brent only had a few contacts and they almost the same persons in each month. If we look into the topics, then we can find that Sara had a lot of words about people, time, and other abstract things while Brent had more detail topics like “development”, “credit”, “support”. Thus, Sara is more like a manager while Brent is more like a executor for specific things. Perhaps, Sara is at higher level than Brent.

In fact, if there are more related information to support our analysis or more professional analyzer who understand the meaning of some words, we could find more useful things from our visualization.

Limitations and Future Works

- So far, we only focus on 12 months within a certain year. However, the time identification can also be weekly, daily or hourly. With other time identification, probably we can figure out the relationship between email volume and holidays or working hours.
- The algorithm we used to find the topics of emails is naive. A better algorithm should be used to extract the topic in the future, so we can better understand what they are talking about.
- In the future, we can research the emails of this community within several continuous years and extract the topics from the emails so that more interesting stories among these people can be revealed.