Visualization personal email behavior

This project will help in analysing personal email archive and reveal the behavior more easily with visualization.

Li Jin-lj963, Kejia Wang-kw1776, Hao Zhang-hz1109, AnQi Liu-aql215

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

Almost everyone uses emails, whether frequently or infrequently. These emails contain a wealth of information about our tendencies: when we send to them, how often we write to them, and who we communicate with. Large amount of people have the demand to detect these behaviors when using emails. However, this kind of information is not well represented and is difficult to view from the raw data. Our goal is to provide clients with a novel visualization to help them understand their email behaviors.

What questions do you want to be able to answer with your visualization?

→ When did I send emails during last two months/weeks(a certain time period)? On what date and at what time in a day did I send? What is the frequency?

The client may be interested in at what time did they send through a certain period, and how often did they send. Answering these questions will illustrate the whole timeline of sending behavior and the client will have a better visual perception.

→ How does the recipients changes during a period? Who are the ones that the user communicate with the most during a certain time period?

Recipients may vary due to different time period since purpose and relationship may change as fact. By visualising this, a big picture of the recipients of the client will be clear.

(→ What are the keywords (of subject) (with certain recipient A)and how did they vary during a period?

“What do I talk about with my various recipients?” may be appealing to the client. The keywords in the subjects could be one way showing the topic of conversation and give the client a view on the diversity.)

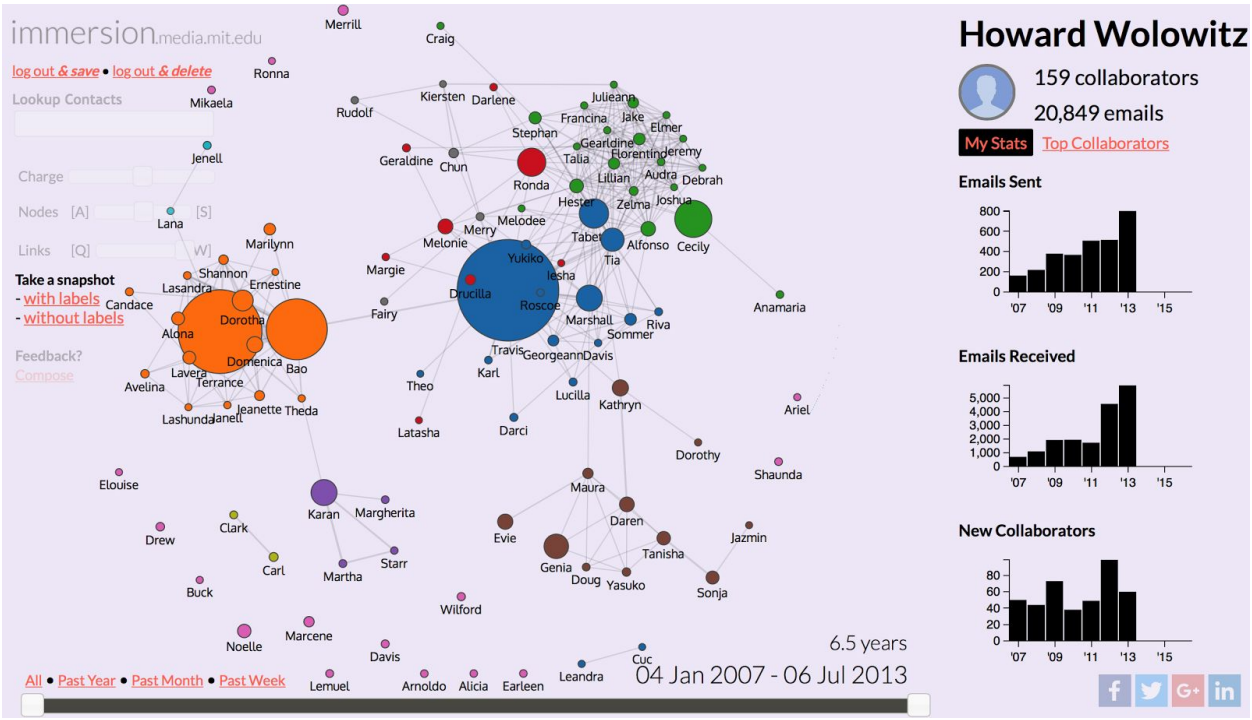
What is your data about? Where does it come from? What attributes are you going to use? What is their meaning? What are their attribute types (data abstraction)? Do you plan to generate derived attributes? If yes, which and why?

The data is about the emails a person sent during a period of time. It comes from Enron Email Dataset. The raw data is a directory structure with each person, and each user's directory structure mirrors the folder structure of their emails with the actual emails in eml or mime format. To solve the problems we proposed, we focus on the attributes as follows:

Attribute Name	Attribute Type	Description	Value Range / Categories	Is Derived
email_id	Categorical	Unique id for each email	Id of all sent emails	N
sent_date	Ordinal	Date and time an email was sent	N/A	N
to_address	Categorical	Email address an email was sent to, including cc and bcc	All email addresses the user sent to	N
subject	Unstructured Text	Subject of an email	N/A	N
content	Unstructured Text	Content of an email	N/A	N
to_person	Categorical	Name of the person an email was sent to	All receivers the user sent to	Y
keyword	Categorical	Keyword of an subject in an email	All keywords derived from the subject	Y
cont_length	Quantitative	Length of the content of an email	N/A	Y

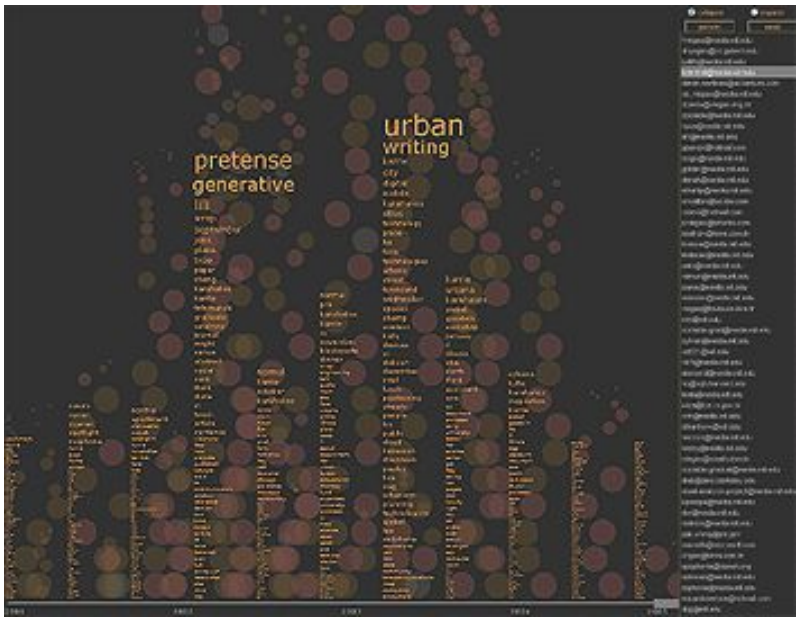
What have others done to solve this or related problems?

Immersion is a people-centric view of the email life. It presents users with a number of different perspectives of their email data(use only the From, To, Cc and Timestamp fields of the mails). It presents users wanting to be more strategic with their professional interactions, with a map to plan more effectively who they connect with.



[Themail](#) is a visualization of the contents of a person's email archive. The application is designed to be used by the owner of the email archive and it addresses two main questions:

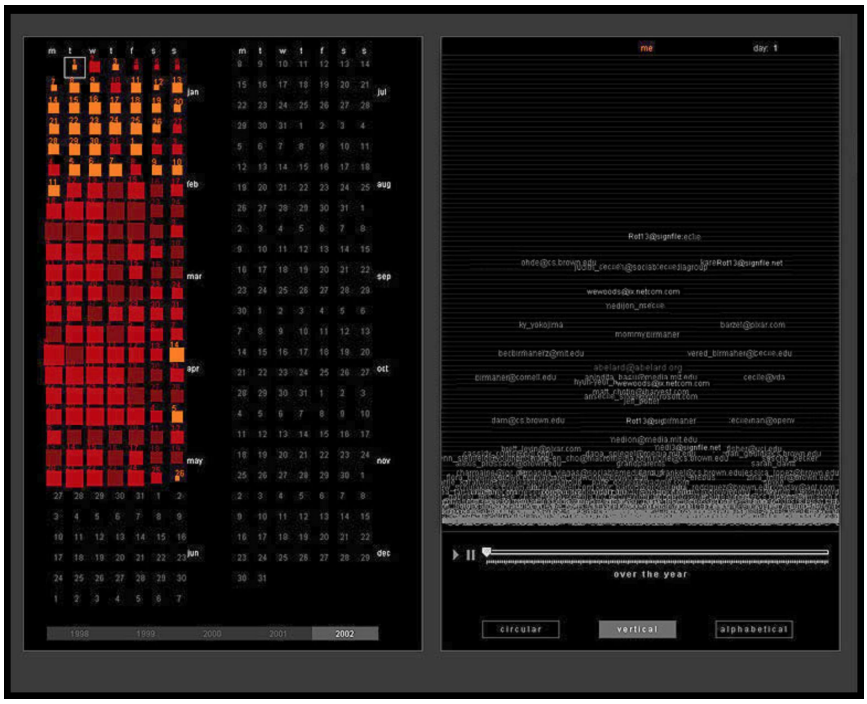
1) What kinds of things do I talk about with my various email contacts?
2) How does my email conversation with person A differs from my conversation with other people?



As in the figure, each column of words refers to emails exchanged in previous months with the selected person. The more salient and unique a word is in your conversation with a specific

person, the bigger that word appears in the visualization. Each circle represents an email message you have exchanged with the selected person.

PostHistory depicts quantitative aspects of a user's email activity on a daily basis. It differentiates between headers and interpret what they mean in terms of social network constructs as well as in terms of formal social structure. It represents for reflection and insightful monitoring of fundamental patterns of interactivity. The visualization aims at impressing on the user a sense of daily accumulation, of growth and scale – dimensions not normally conveyed on current email applications.



What solution do you propose?

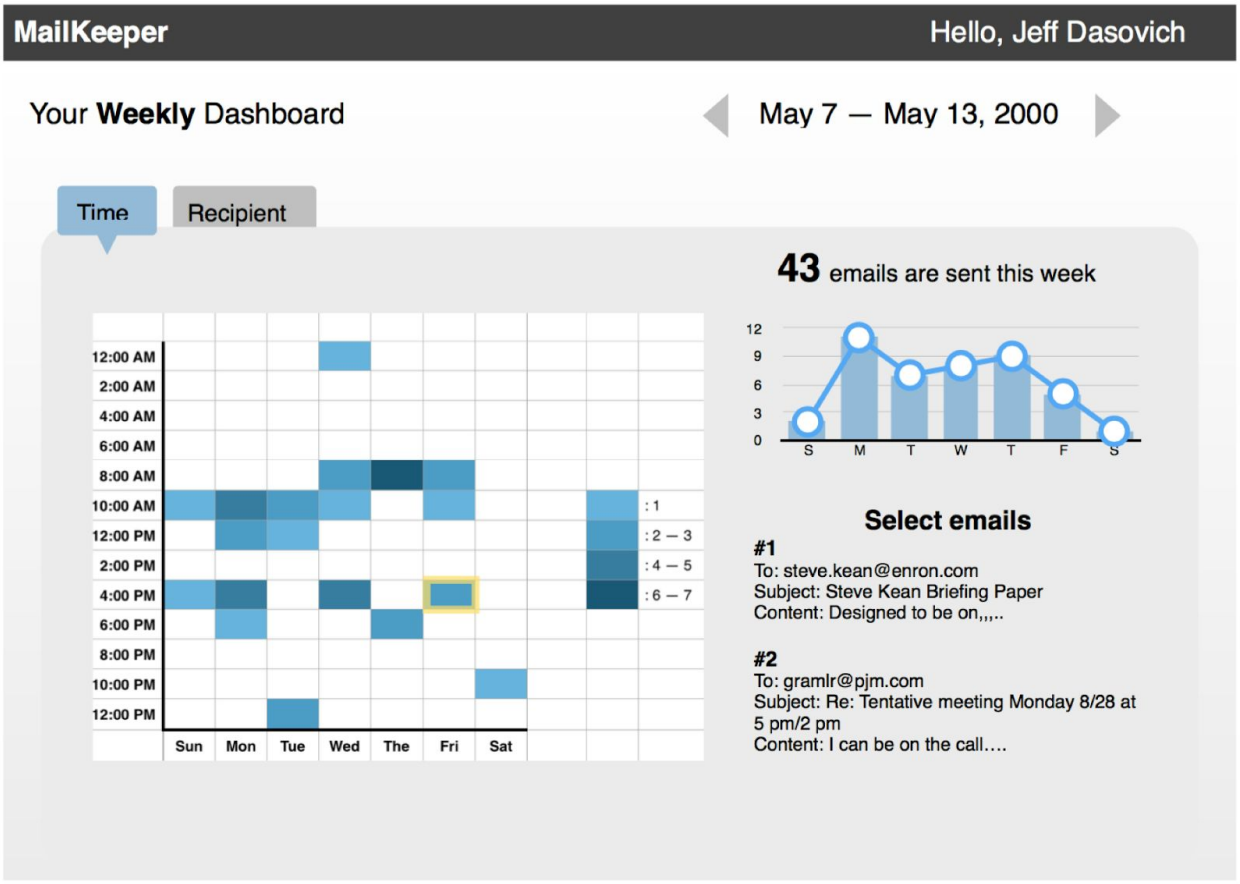


Figure 1

Figure 1 shows the weekly time trend of emails sent by the client. There are two charts shown above. A selector allows client to choose from “weekly”, “two-weekly”, “monthly”, etc. There are two tabs, respectively are two sections, time and recipient.

Below that is a heatmap, which is the main chart showing the frequency of emails sent. In the heatmap, the x-axis is the day in a week currently and the y-axis is the time period in a day. If the user change “weekly” to “monthly”, the x-axis will change to the month of every year and the y-axis will change to the day of this month. The color intensity indicates the amount of the emails sent during that period of time by the client. When user hover the mouse to a specific rectangular area inside the heatmap, the number will show in tooltip and the corresponding area will be highlighted.

At the right side, there is a column separated into top and down two parts. The top part is a bar chart, shows the total amount of emails the certain period of time. The down part is for the content.

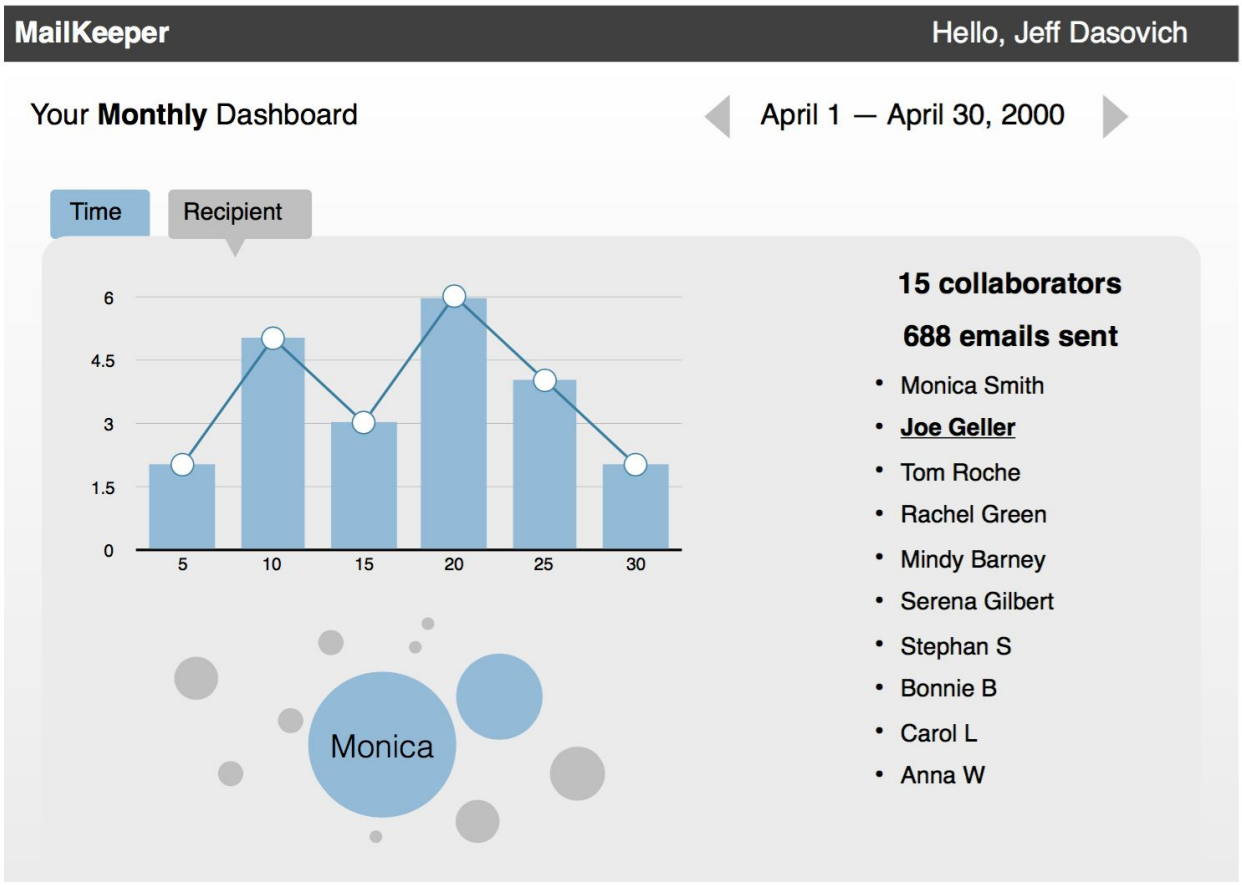


Figure 2

Figure 2 shows the recipients the user have sent email to in a yearly view. The left part consist of two charts. The bar chart shows how the amount of recipients changes throughout the year. Below the bar chart is a bubble set represent the recipients. Each bubble represents a single recipient and the size of the bubble represents the amount of emails the user sent to this recipient. Different color indicates different categories that the recipient belongs to.
The right side is a list shows the name of all recipients sorted by the most frequency. When user hovers mouse to the bubble, the corresponding recipient's name will show as tooltip and the name in the list will be highlighted. When user clicks on a bubble or a name entry in the list, the bubble and the corresponding recipient's name in the list will be highlighted. At the same time, the list in the right side will change to another view shown as Figure 3.

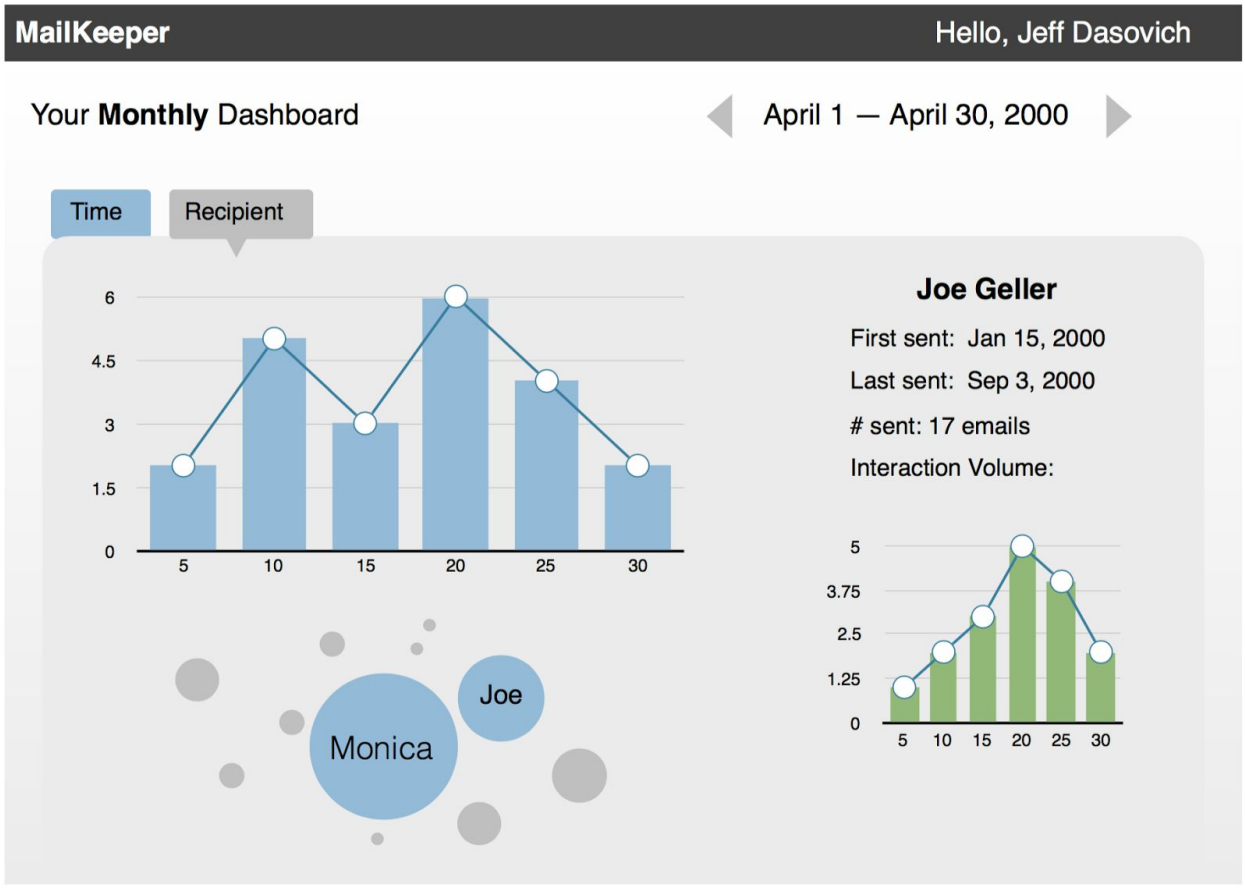


Figure 3

Figure 3 shows the detail information of the chosen recipient. In the right part, the list of recipients is changed to show the name, first sent email, last sent email, total number of emails sent and the interaction volume. The chosen recipient is highlighted and the smaller bar chart shown below indicates the time trend of the number of emails the user sent to this recipient. Figure 2 and 3 together show how the recipients changes during a period, who is the one that the user communicate with the most and the trend of this relation during a period.

How does the solution help you answer the questions stated above?

1. When did I send emails during last two months/weeks(a certain time period)? On what date and at what time in a day did I send? What is the frequency?

On the heatmap, we can see how many emails user sent at a certain day or a certain hour. We can compare color zone and blank zone to know what time the user like sending email most. The user will know when he sends email and how much he sent emails clearly. Like the github activity record, he will know which certain period he like sending the email most. Maybe he does not sent a single email for a long time. He will get it from the blank zone.
Use a bar chart to show the amount of the sending emails change over time.

2. How does the recipients change during a period? Who are the ones that the user communicates with the most during a certain time period?

Use a heatmap to show how many recipients change during a certain period. We can clearly compare the amount of recipients between different time period.
Use a bar chart to show the amount of receipts change over time. It is easy to determine which one does the user communicate most during a certain time.

3. The content part is for potential use for questions about key words.

How do you plan to verify whether you have met your goals with this project?

To verify whether the goal is reached, we need to confirm two stages:
1. The project should answer all questions listed above when the email dataset is imported or being accessed, which meets the basic demand.
2. The final visualization should be tally with the behavior of the client.

How is your project team going to work on the project? Who is going to do what?

There are 3 major phase to complete the project: data handling, web development, document enhancing. We choose to use Github as our repository of source code of the project. Tasks related to web development will be divided to small ones later and assign to group members. Weekly group meetings would be set up to communicate with the progress and sync up with each other.

	Task	Assignee
1	Extract data of one user's sent emails from Enron Dataset	Li Jin
2	Select the attributes needed from the raw data	Whole group
3	Define the standard format of json file and convert the original directory structure and emails into a single json file	Hao Zhang
4	Combine the individual mock-up of each member, discuss and design the UI of the web application: what is the content of the web page, how many charts to show, how the UI changes with interaction	Whole group
5	Web development and implementation of widgets of the visualization (details will be determined later)	Kejia Wang, Li Jin, Hao Zhang, An Qi Liu
6	Web application testing: including functionality, usability, and interface testing	An Qi Liu

7	Final report completion	Whole group