Visualization Personal Email Behavior

FINAL PROJECT DESCRIPTION

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

Project page (on Github):
https://github.com/NYU-CS6313-SPRING2016/Group-4-Personal-Browsing
Video: https://vimeo.com/167371397
Working demo: http://nyu-cs6313-spring2016.github.io/Group-4-Personal-Browsing/

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 did we send, how often is that happened, who we communicate with, and what are we talking about. 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-sending behaviors.

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

1 When did I send emails during certain week? On what day and at what time in a day did I send? What is the frequency like?
→Client may be interested in at what time did they send through certain week and how often did it happen. Answering these questions will illustrate the whole timeline of sending behavior and client will have a better visual perception.

2 Who did I sent most to in a certain week? How did they vary with weeks? How often and when did I send emails to specific contacts, and how did that change over time?
→Recipients may vary due to different weeks since purpose and relationship may change as fact. Also, how did certain contact change through the week is informative. By visualising these, a big picture of the recipients of the client will be clear.

3 What are the most frequently used keywords during certain week? How did they vary with weeks?
→"What do I talk about with my various recipients?" may be appealing to the client. The keywords abstracted from contents could be one way showing the topic of conversation and

give the client a view on the diversity.

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

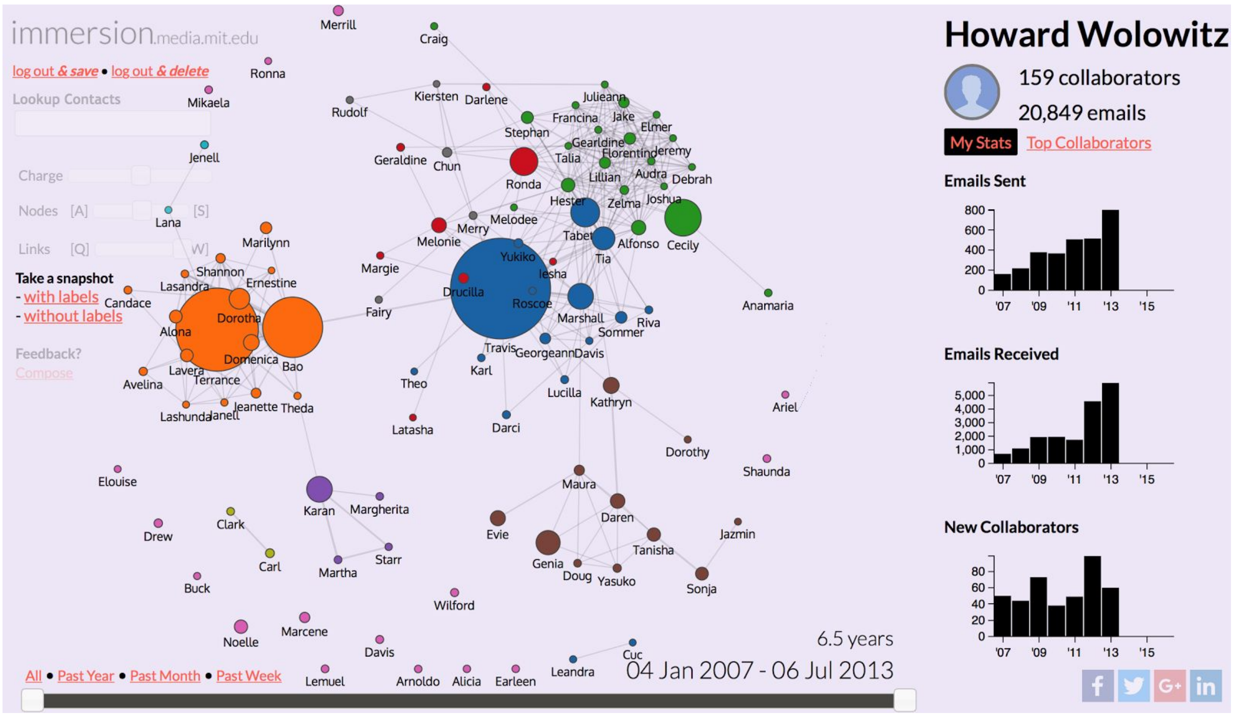
The data is about the emails a person sent during a period of time. It comes from Enron Email Dataset. The raw data is a directory structure with each person, and each user's directory structure mirrors the folder structure of their emails with the actual emails in eml or mime format. To solve the problems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

Attribute Name	Attribute Type	Description	Value Range / Categories	Is Derived
sent_time	Ordinal	Date and time an email was sent	N/A	N
recipient address	Categorical	Email address sent to, including cc and bcc	All email addresses the client sent to	N
content	Unstructured Text	Content of an email	All content client sent	N
week_number	Categorical	Weeks identification	N/A	Y
start_date	Ordinal	Start date of the week	All start date of week	Y
end_date	Ordinal	End date of the week	All end date of week	Y
year	Ordinal	Which year is the week in	All years	Y
days	Set of Attributes	A set of attributes that describe a period of time	N/A	Y
day_num	Ordinal	Number each day of the week	0~6, 0 for Sunday and etc.	Y
daytime	Ordinal	Time of the day	0~23	Y
day	Categorical	Day of the week	Su~Sa, Su for Sunday etc.	Y
num_recipient	Quantitative	Number of recipients sent to in the week	N/A	Y
date	Ordinal	Date of email sent	All dates that client	Y

			has sent emails on	
total	Quantitative	Total number of emails in the week	N/A	Y
sent_detail	Array	An array with hashmap related to daytime attribute	N/A	Y
recipients	Set of Attributes	A set of attribute that describe some properties of recipients	N/A	Y
addr	Categorical	Address of recipients that client sent to	All sent addresses	Y
num	Quantitative	Number of emails sent to the recipient	N/A	Y
keywords	Set of Attributes	A set of attributes that describe some properties of keywords	N/A	Y
word	Categorical	Word abstracted from contents of the week	All keywords	Y
freq	Quantitative	Frequency of the word	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.
Although we are not dealing with the network between recipients of the client, it gives us thought to visualize how recipients vary from week to week.

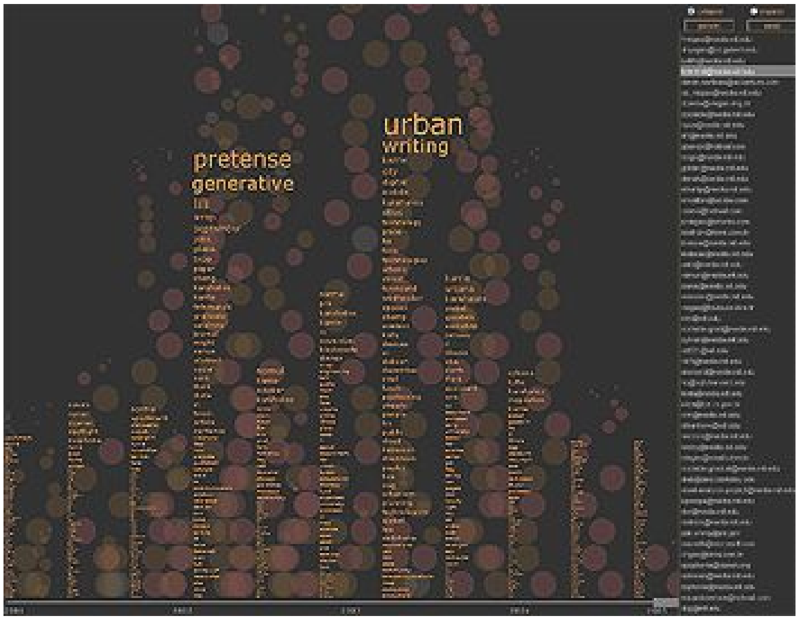


Themail is a visualization of the contents of a person's email archive. The application is designed to be used by the owner of the email archive and it addresses two main questions:

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?

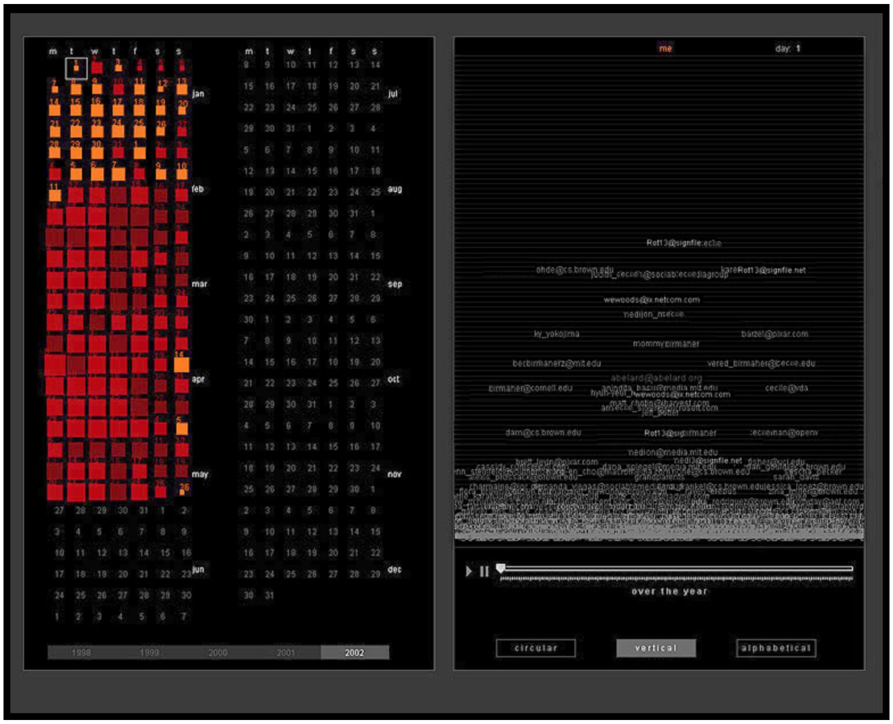
As in the figure, each column of words refers to emails exchanged in previous months with the selected person. The more salient and unique a word is in your conversation with a specific person, the bigger that word appears in the visualization. Each circle represents an email message you have exchanged with the selected person.

We were inspired by this visualization and decided to abstract keywords from contents to help client to have a better look at the topics.

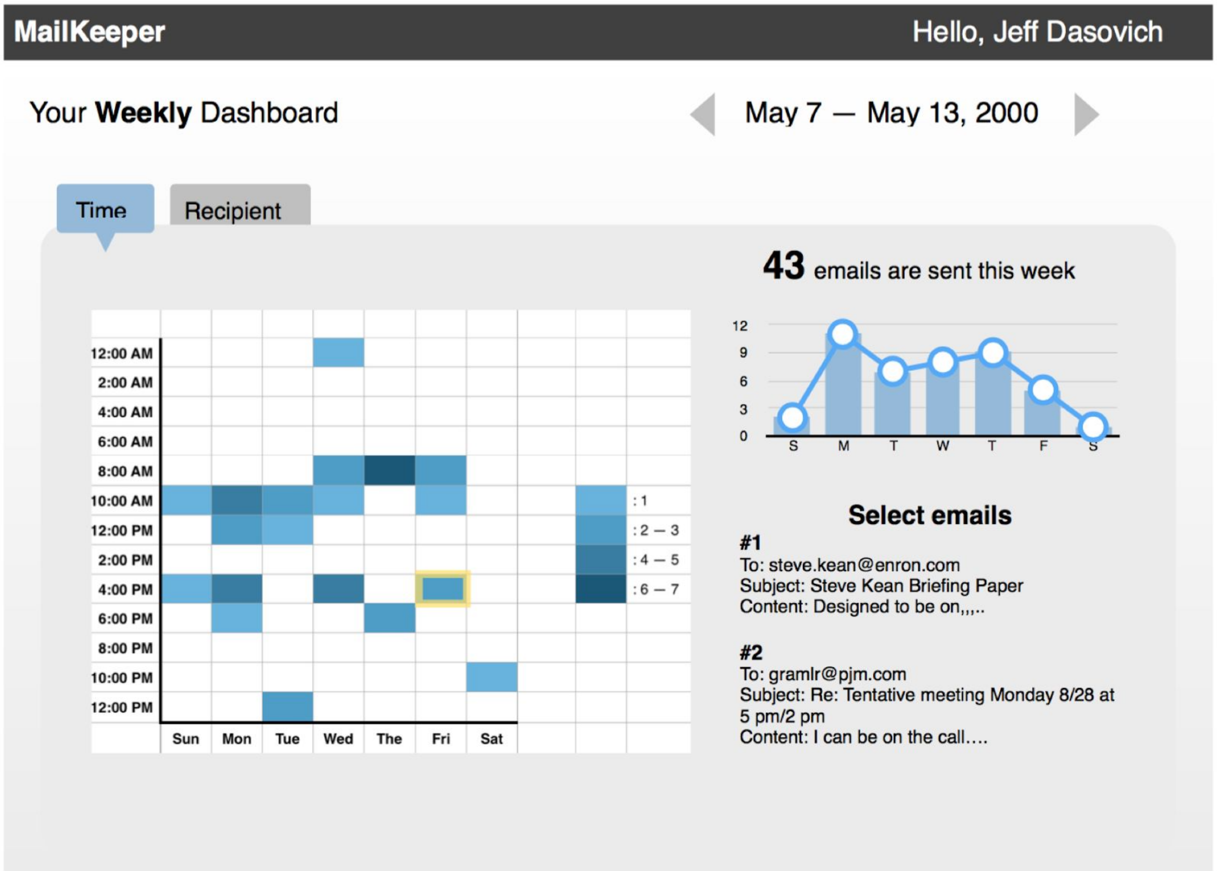


PostHistory depicts quantitative aspects of a user's email activity on a daily basis. It differentiates between headers and interpret what they mean in terms of social network constructs as well as in terms of formal social structure. It represents for reflection and insightful monitoring of fundamental 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.

Our prototype with heatmap is much similar to this app. We want to create a view to show the frequency of the sending behaviors.



Initial Mockup



MailKeeper

Hello, Jeff Dasovich

Your **Weekly** Dashboard

May 7 — May 13, 2000

43 emails are sent this week

Select emails

#1
To: steve.kean@enron.com
Subject: Steve Kean Briefing Paper
Content: Designed to be on,,,,.

#2
To: gramlr@pjm.com
Subject: Re: Tentative meeting Monday 8/28 at 5 pm/2 pm
Content: I can be on the call….

Figure 1

How to read it:

Figure 1 shows the weekly time trend of emails sent by the client. There are two charts. 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 client 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 client 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.

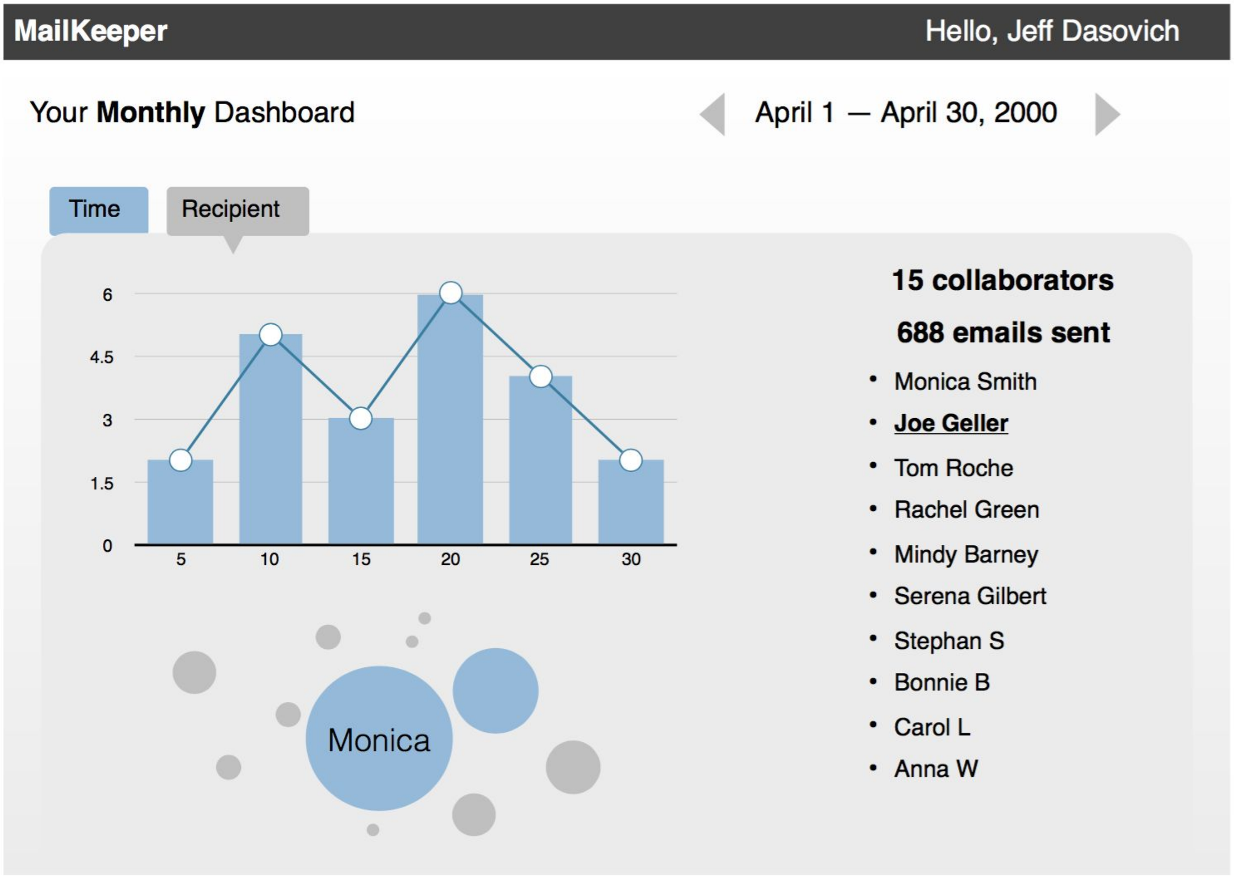
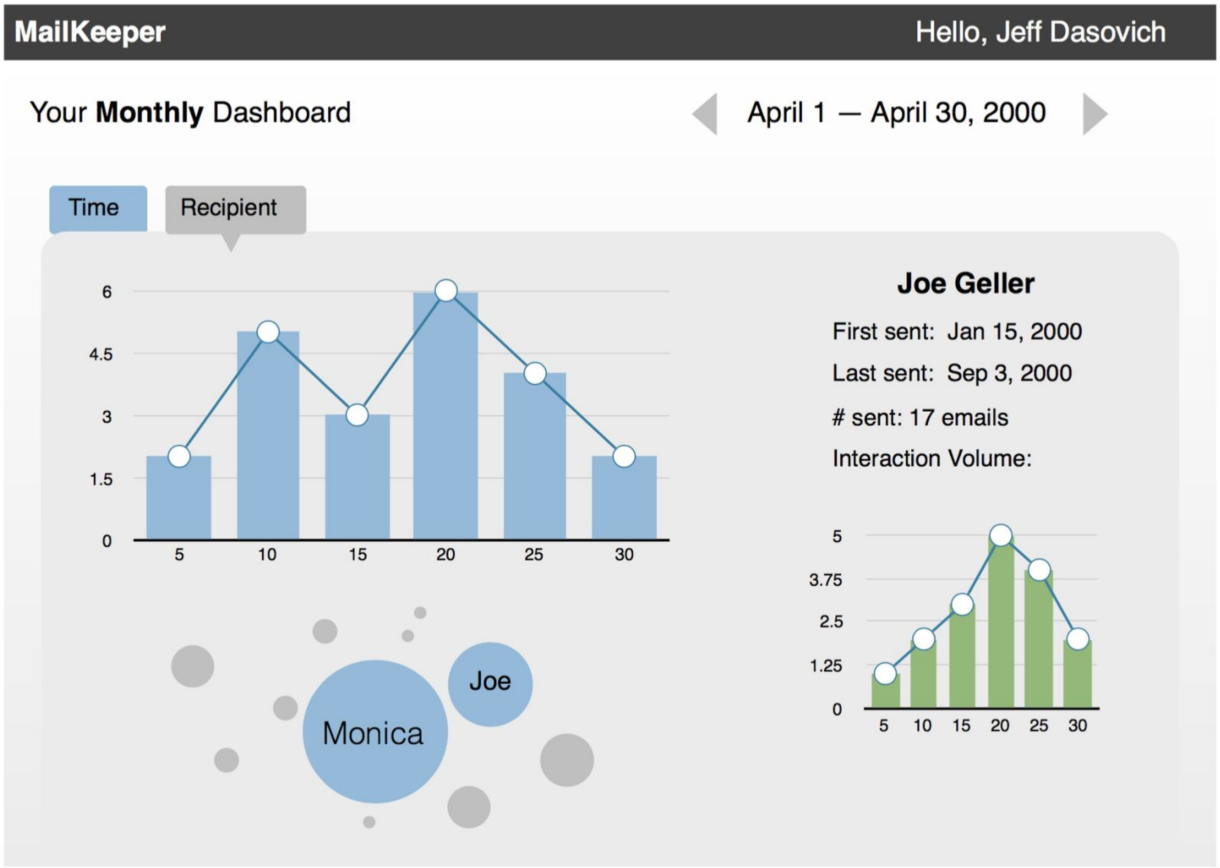


Figure 2

How to read it:
Figure 2 shows the recipients the client 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 client 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 client 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.



MailKeeper

Hello, Jeff Dasovich

Your **Monthly** Dashboard

April 1 — April 30, 2000

Joe Geller

First sent: Jan 15, 2000

Last sent: Sep 3, 2000

sent: 17 emails

Interaction Volume:

Figure 3

How to read it:

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 client sent to this recipient. Figure 2 and 3 together show how the recipients changes during a period, who is the one that the client communicates with the most and the trend of this relation during a period.

What did not work and how to improve it:

- Issue: Focus on weekly views first. You can do monthly or yearly later.
  Solution: Change the main view to weekly and put monthly as an option.

- Issue: Change the order of the axes: hours on the x-axis and days on the y-axis.
  Solution: Put hours on the x-axis and days on the y-axis.

- Issue: Don't use color to show volume, use bars and encode volume with their height/length.
  Solution: Use bars that grow with volume.

- Issue: Rather than showing the volume by day in the bar on the right (which is too redundant) create a new view as follows. A list with multiple bar charts. Each one represents one week, the same way you do for the bar chart you currently have on the right. Put in the list the whole set of weeks you have in the data. This way the list can be used to quickly grasp the whole data and jump to "interesting" weeks

  Solution: Remove some redundant sections and add week list on the left side to make the view clearer and easier to read and interact.

- Issue: You should extract and visualize keywords for the current week. This is quite easy. Take all the words in the week, remove stop words and rank them according to the TF/IDF method (https://en.wikipedia.org/wiki/Tf%E2%80%93idf) where document frequency is the frequency of words in the whole collection of emails and term frequency is the frequency in the current week.

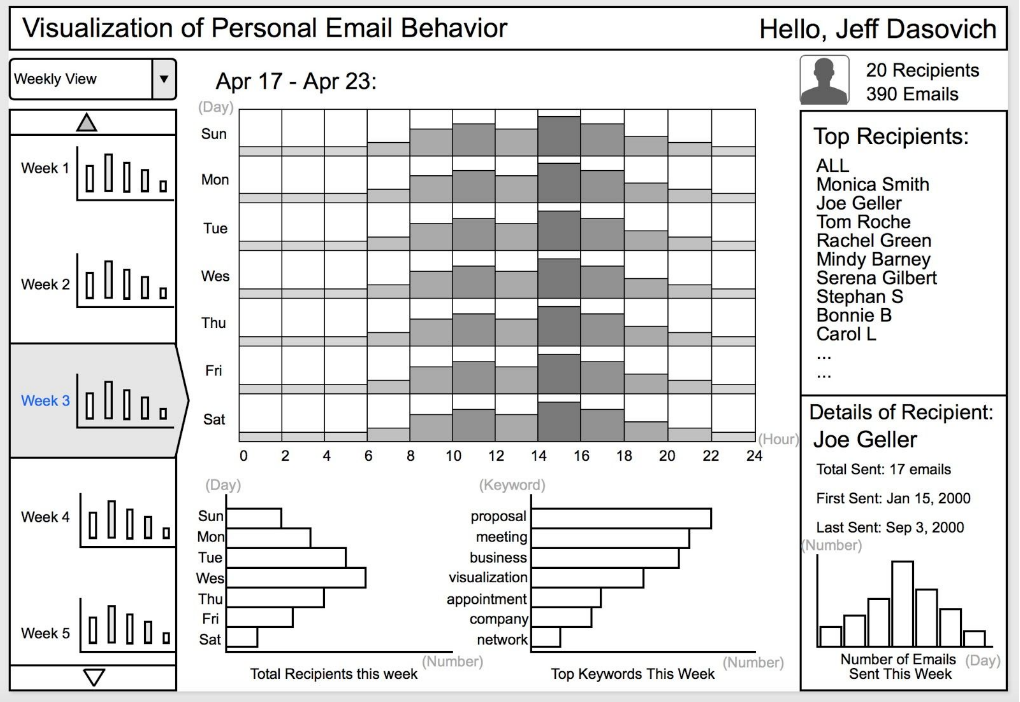
  Solution: Add a bar chart, listing top seven keywords that are frequently used in the current week.

- Issue: I suggest to merge the view you have in Figure 1 and Figure 2. can you add recipients directly in the timeline view? I believe you have enough space. Just make the view of recipients a list, not a series of bubbles. Use horizontal bars to show the frequency rather than bubble size.

  Solution: Put timeline view on top, and charts of recipients and keywords at the bottom. The recipient list is at the right side so that client can have a better view.

Project Update

Sketch of the updated version:



Visualization of Personal Email Behavior

Hello, Jeff Dasovich

Weekly View ▾

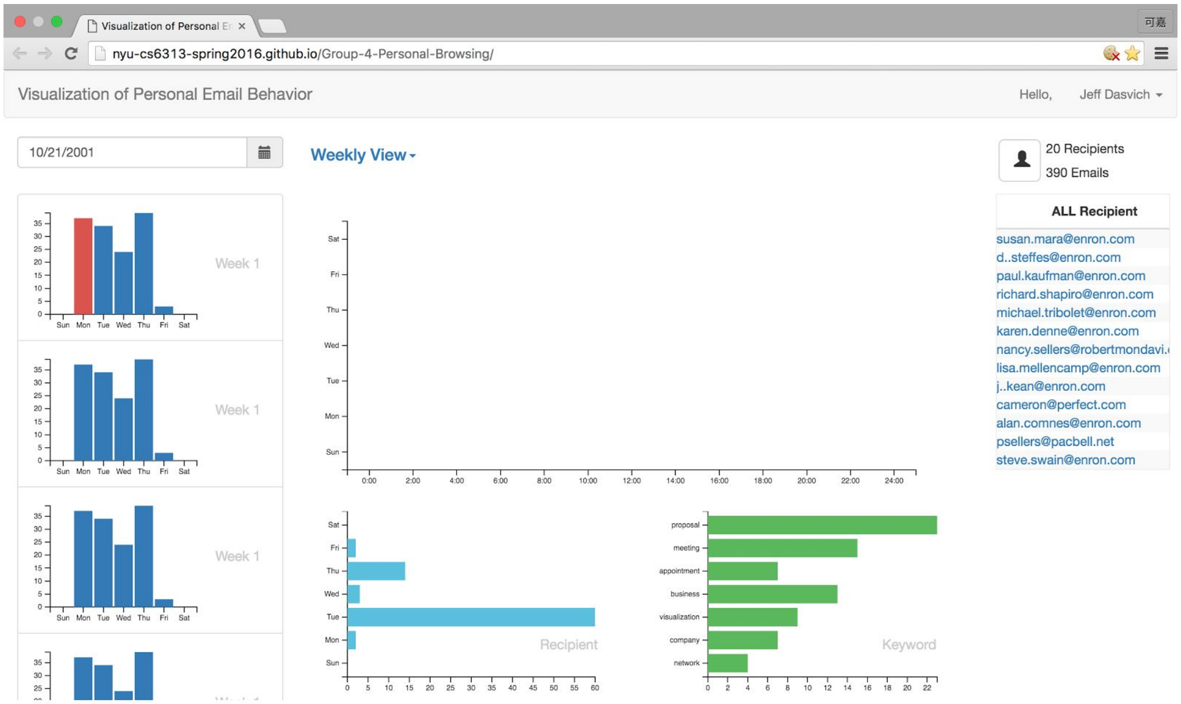
Apr 17 - Apr 23:

20 Recipients
390 Emails

Top Recipients:
ALL
Monica Smith
Joe Geller
Tom Roche
Rachel Green
Mindy Barney
Serena Gilbert
Stephan S
Bonnie B
Carol L
...
...

Details of Recipient:
Joe Geller
Total Sent: 17 emails
First Sent: Jan 15, 2000
Last Sent: Sep 3, 2000

Implementation of the updated version:



Visualization of Personal Em ×

nyu-cs6313-spring2016.github.io/Group-4-Personal-Browsing/

Visualization of Personal Email Behavior

Hello, Jeff Dasvich ▾

10/21/2001

Weekly View ▾

20 Recipients
390 Emails

ALL Recipient
susan.mara@enron.com
d..steffes@enron.com
paul.kaufman@enron.com
richard.shapiro@enron.com
michael.tribolet@enron.com
karen.denne@enron.com
nancy.sellers@robertmondavi.
lisa.mellencamp@enron.com
j..kean@enron.com
cameron@perfect.com
alan.comnes@enron.com
psellers@pacbell.net
steve.swain@enron.com

How to read it:

The visualization indicates weekly view of personal email behavior. The navigation banner shows the name of the project and of the client.

Below the banner on the left side, there is a selector so that client can select the week directly by date. Below the selector is the list of weeks with a bar chart for each week. Client can simply select a week of interest and all information for that week will be shown in the middle.

There are three charts in the middle. At the top is day-hour chart to reveal the frequency of emails sent by client during the current week. This chart has hour as the x-axis and day as y-axis, with hours displayed bi hourly, and days displayed from Sunday to Saturday. At the bottom left is the number of recipients the client sent emails to during the current week. This chart has number as the x-axis and day as the y-axis, with days from Sunday to Saturday. At the bottom right is the top seven keywords the client mentioned in his emails during current week. This chart has number as the x-axis and list of keywords as the y-axis.

Below the banner on the right side, the account overview is displayed at top right corner. Below that is the list of recipients the client sent emails to during the current week, sorted by frequency.

What did not work and how to improve it:

- Issue: In left panel, weekly preview are way to big; scale for every week should be the same; date selector is redundant. Weekly list should be using real data.
  Solution: Reduce the space of left panel to spare more for the main chart; maintain the scale for all week; remove the date selector and add label to each week preview. Implement real data.

- Issue: Main chart should be implemented, and narrow shape could be more reasonable; start time of a day should be marked; total number of emails of a day is needed, could use bubble with certain size and number inside it.
  Solution: Implement the main chart and use one hour as the grid interval. Shadow the nighttime, make it obvious. Add bubble with text at the left side of the main chart showing the number of emails of a day.

- Issue: Recipient bar chart at the bottom is redundant; abstract keywords from the content is needed.
  Solution: Remove the bar charts from the center display. Merged the recipient chart with the recipient list on right hand side using a background bar to show the number and trend. The keywords bar chart is relocated as a second tab on the right hand side; keywords are abstracted and put in the list.

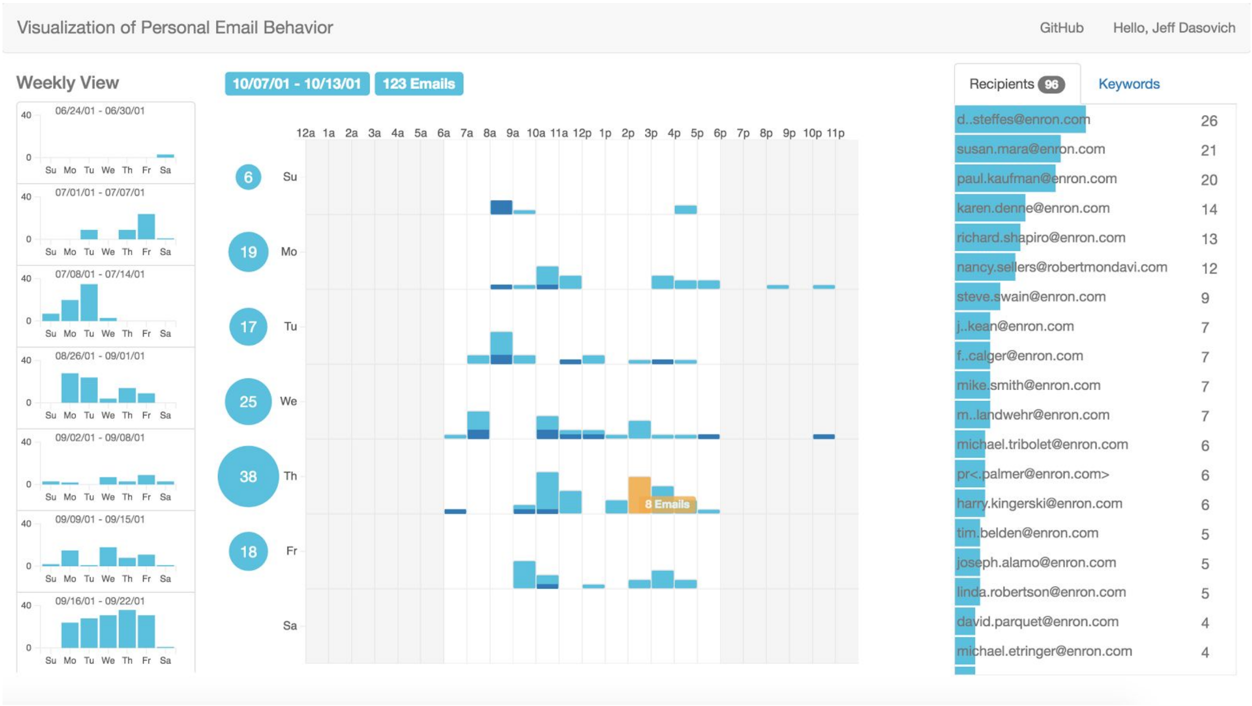
- Issue: The relationship between the charts are necessary.
  Solution: Implement interaction between the data shown in charts.

- Issue: Colors don’t seem to make any sense.

Solution: Remove unnecessary color, remain certain color to show the basic information and highlight informations.

Final Visualization

Final implementation of the project:



d..steffes@enron.com	26
susan.mara@enron.com	21
paul.kaufman@enron.com	20
karen.denne@enron.com	14
richard.shapiro@enron.com	13
nancy.sellers@robertmondavi.com	12
steve.swain@enron.com	9
j..kean@enron.com	7
f..calger@enron.com	7
mike.smith@enron.com	7
m..landwehr@enron.com	7
michael.tribolet@enron.com	6
pr<.palmer@enron.com>	6
harry.kingerski@enron.com	6
tim.belden@enron.com	5
joseph.alamo@enron.com	5
linda.robertson@enron.com	5
david.parquet@enron.com	4
michael.etringer@enron.com	4

How to read it:

The visualization indicates weekly view of personal email behavior. The navigation banner shows the name of the project and name of the client.

Below the banner on the left side, there is a list of weeks with a bar chart for each week. Client can simply select a week of interest and all information for that week will be shown in the main chart in the middle.

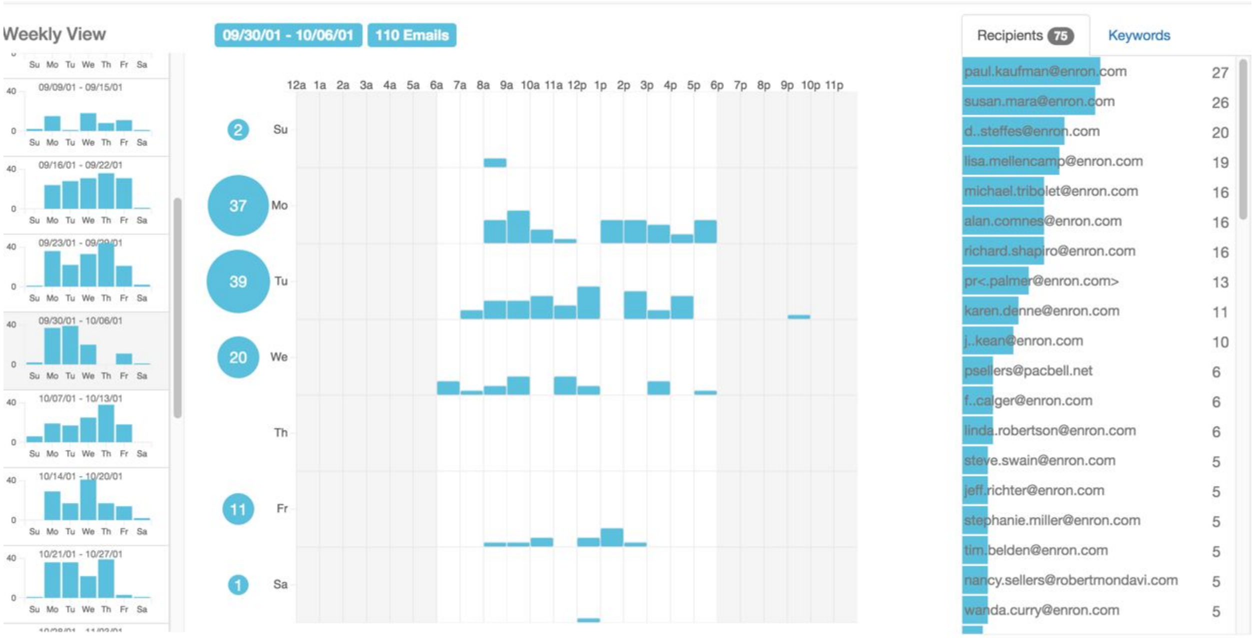
On top of the middle, it shows the selected week and the total number of emails of the certain week. The main chart then is a day-hour multi-bar chart to reveal the frequency of emails sent by client during the selected week. This chart has hour as the x-axis and day as y-axis, with hours displayed from 12am to 11pm, and days displayed from Sunday to Saturday.

Vertically-arranged bubbles on the left of main chart show the total number of emails client sent on certain day from Sunday to Saturday. When client hovers onto certain bar, a tooltip can show how many emails did client send that hour that day of the selected week.

On the right side, there is a two-tab panel shows recipients and keywords respectively. They are both sorted by total number of the selected week. Beside the “recipient”, there is a number shows how many recipients client has sent to in this week. When client clicks certain recipient, in the main chart, highlighted bar will show when did client send to this recipient and the number of the emails.

Data Analysis

1. When did Jeff send emails during certain week? On what day and at what time in a day did he send? What is the frequency like?

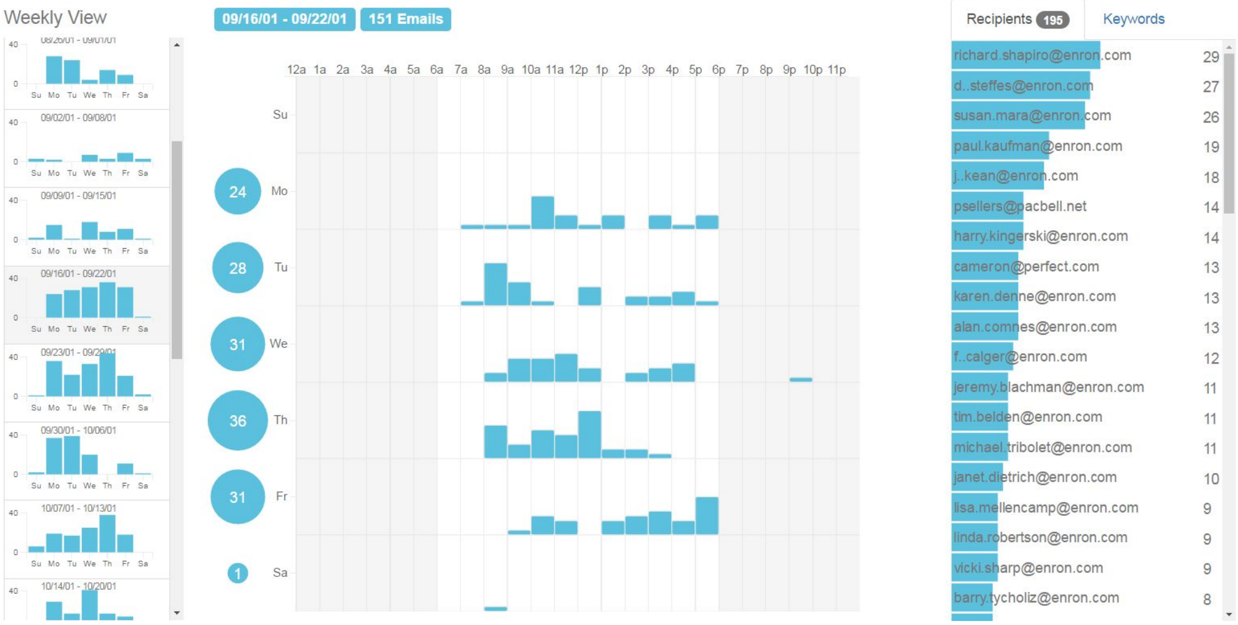


paul.kaufman@enron.com	27
susan.mara@enron.com	26
d..steffes@enron.com	20
lisa.mellencamp@enron.com	19
michael.tribolet@enron.com	16
alan.comnes@enron.com	16
richard.shapiro@enron.com	16
pr<.palmer@enron.com>	13
karen.denne@enron.com	11
j..kean@enron.com	10
psellers@pacbell.net	6
f..calger@enron.com	6
linda.robertson@enron.com	6
steve.swain@enron.com	5
jeff.richter@enron.com	5
stephanie.miller@enron.com	5
tim.belden@enron.com	5
nancy.sellers@robertmondavi.com	5
wanda.curry@enron.com	5

You can see from the screenshot that Jeff sends the majority of his emails between the hours of 6am to 6 pm. You can also observe, both from the weekly list as well as the main chart, that most emails were sent on the weekdays. On this particular week, only 4 emails out of 110 total were sent outside of this time range. This information suggests that Jeff has normal work hours, during which he sends the majority of his emails.

You can also observe that Jeff had a spike in emails sent during some weeks in September and October. This raises the obvious question: Why are was there this spike in emails sent in these weeks and who were they to?

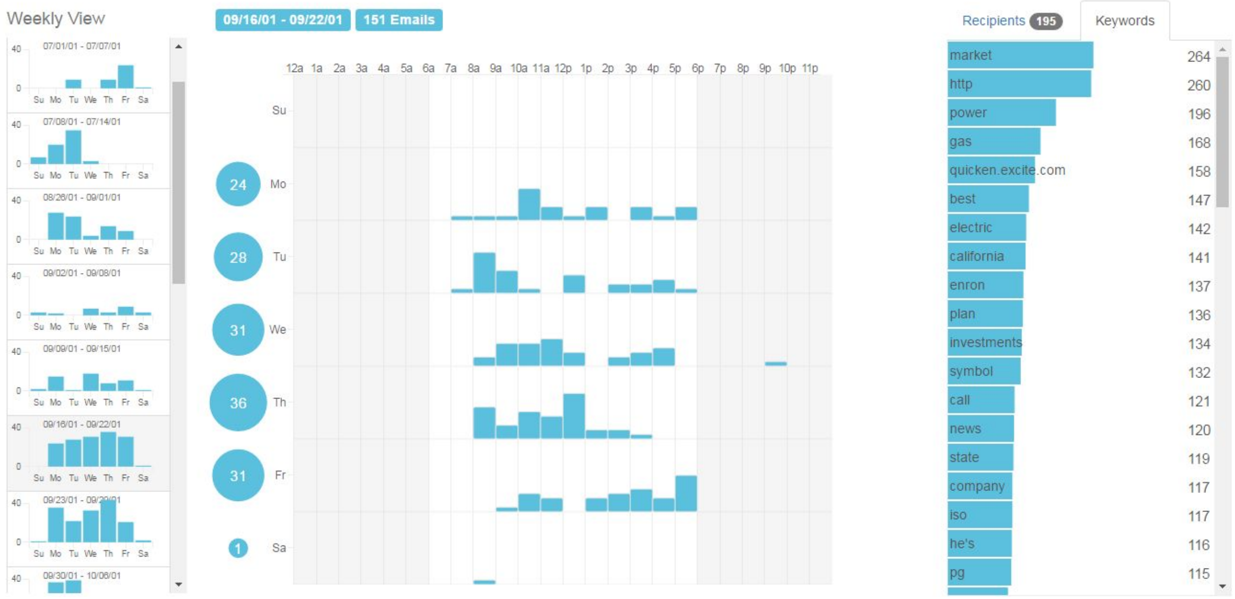
2. Which recipients did he sent to most on certain weeks?



richard.shapiro@enron.com	29
d..steffes@enron.com	27
susan.mara@enron.com	26
paul.kaufman@enron.com	19
j..kean@enron.com	18
psellers@pacbell.net	14
harry.kingerski@enron.com	14
cameron@perfect.com	13
karen.denne@enron.com	13
alan.comnes@enron.com	13
f..calger@enron.com	12
jeremy.blachman@enron.com	11
tim.belden@enron.com	11
michael.tribolet@enron.com	11
janet.dietrich@enron.com	10
lisa.mellencamp@enron.com	9
linda.robertson@enron.com	9
vicki.sharp@enron.com	9
barry.tycholiz@enron.com	8

We can observe from the above image that during the week of September 16 to September 22, jeff sends many emails to Richard, Steffers and Susan. This is not very different from other weeks, as many other weeks also feature these three recipients on the top 5 recipients for the week. This, along with the information that there are more recipients than emails for this week, tells us that Jeff just simply had a big spike in email usage for the week, particularly emails that are sent to multiple recipients at once.

3. What are the most frequently used keywords during certain weeks? How did they vary with weeks?

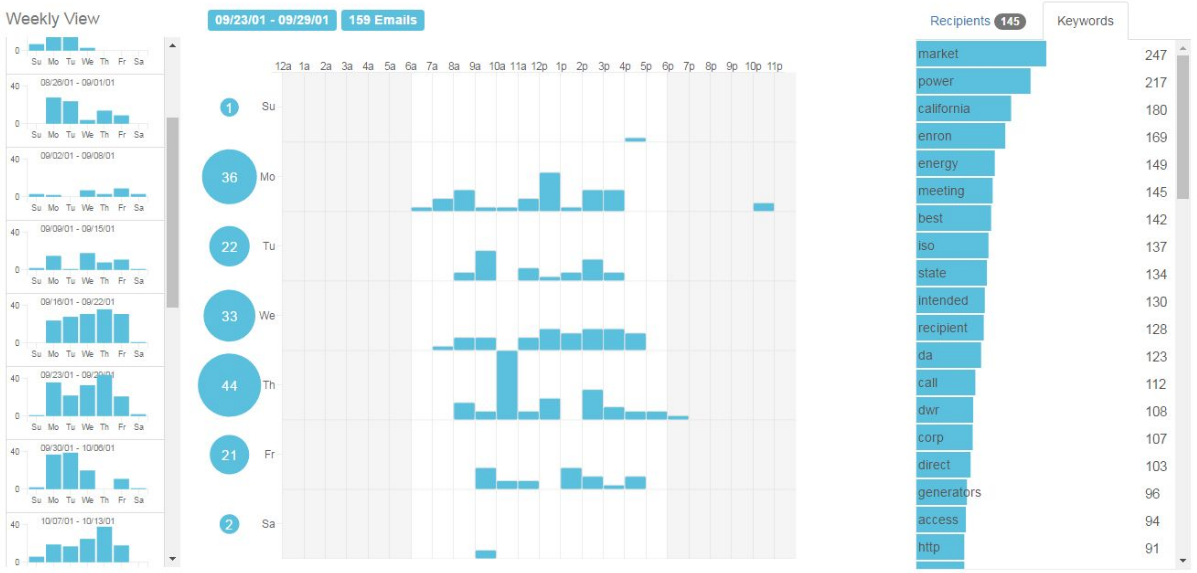


market	264
http	260
power	196
gas	168
quicken.excite.com	158
best	147
electric	142
california	141
enron	137
plan	136
investments	134
symbol	132
call	121
news	120
state	119
company	117
iso	117
he's	116
pg	115

From keywords section, we can see that Jeff was sending emails which used keywords such as market, power, gas, electric many times. This tells us that the majority of his sent emails during this week were work related, also indicating that Jeff probably works at an energy company.

Jeff Dasovich was actually a state government affairs executive for Enron, an energy, commodities and services company. Even without this prior information, many details about his work life and what he did can be accurately inferred from the email data.

Another interesting observation that can be made for the week of September 16 to September 22 is that the keywords 'http' and 'quicken.excite.com' are among the top 5 keywords for the week. If we compare to the other weeks in the visualization, we can notice that these keywords are seldom used in other weeks. With this information, we can safely assume that this new web page, https://quicken.excite.com is either the cause of the spike in emails for this week, or the main topic for the emails sent this week.



market	247
power	217
california	180
enron	169
energy	149
meeting	145
best	142
iso	137
state	134
intended	130
recipient	128
da	123
call	112
dwr	108
corp	107
direct	103
generators	96
access	94
http	91

Limitations and Future Works

- Currently, our project can only answer questions related to the weekly view of sent emails, which may not be so informative. We would like to make it more complete by adding monthly view and maybe yearly view in the future.
- Another limitations is that we do not have an interface for user to upload their own data and get a personalized visualisation from our project web page. From this stage, we manipulate the sample data from Enron dataset offline and use only a certain formatted JSON file as the data source to get the chart implemented. In the future, we may design an interface or API to provide a portal to access the functions of our project application. For this purpose, we plan to set up a server and a database to support advanced functions such as enable searching keywords in content, showing complete details of each email, etc.