# Progress Report

**CS448b Final Project**
**Sanby Lee**

The goal of my project is to create a digital tool that enables users to visualize their email archives and draw insights about their social interactions with their contacts. Potential questions to answer might be: Who do I correspond with the most? How frequently do we correspond? What sort of topics do we talk about?

## Literature Review

A review of the related literature shows that a surprisingly large amount of work has already been conducted on the visualization of email communication. We can categorize the previous work into four broad areas, based on the purpose of the work and the type of email data that is visualized.

***Enterprise Analytics.*** These projects are aimed at providing insights that enable users to answer their emails efficiently, particularly in a business context. Perer and Smith created several visualizations that showed enterprise users how effective they are at managing email, including (a) a treemap that organizes a user's contacts hierarchically, based on the domain name and suffix of their email address (b) a scatter plot that plots each contact based on the number of messages received from versus the number of messages sent to that contact (c) a chart that plots the number of conversations initiated by the user versus the number of conversations responded to by the user on a weekly basis [5]. Their user studies showed that the visualizations led to insights that helped users manage their workflow more effectively, such as identifying contacts for whom email was not an efficient way of getting responses. On the commercial side, Gmail Meter provides a monthly report that includes a breakdown of emails sent and received, length of emails, and response time, visualized using bar charts and time series charts [1]. Both of these projects are focused on visualizing statistics about the email messages themselves, and do not show connections between users.

***Social Network Analysis.*** Another group of projects are aimed specifically at showing the network connections between users. Viegas et. al have created *Social Network Fragments*, a visualization that plots all of a user's contacts and draws connections between them based on the emails exchanged. They use a spring system algorithm to position contacts with strong bonds close to each other, and contacts with weak bonds far from each other [6]. *Immersion*, from the MIT Media Lab, creates a similar visualization, grouping contacts into clusters and drawing ties between all of them [2]. *My Map* takes a slightly different approach, by positioning all contacts in a ring, and drawing ties between them [3]. All of these projects rely solely on email metadata, such as the From, To, and timestamp fields, in order to determine the presence of a connection, and do not look at the contents of the message body. Viegas et. al and the *Immersion* team also note privacy concerns, which may have influenced the decision to look at metadata only.

***Communication Patterns.*** The third group of projects focus on studying dyadic communication between the user and one other contact, rather than looking at the social

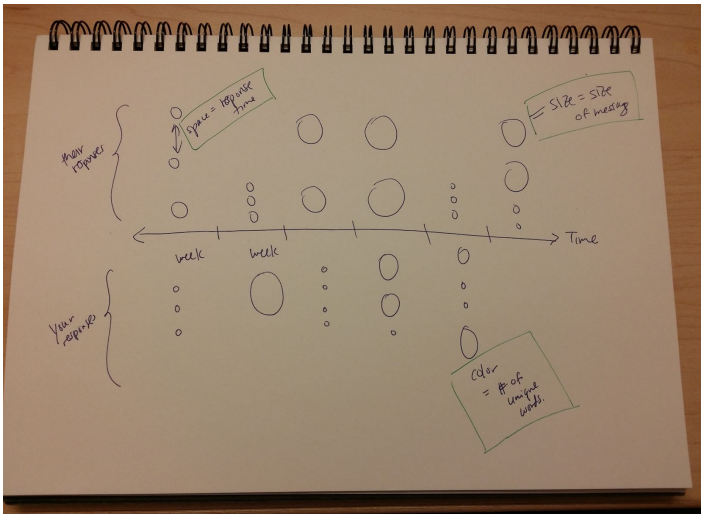
network as a whole. The goal of these projects is to characterize the "rhythm" of communication between two people. Viegas et. al have created *PostHistory*, which uses a calendar format to visualize the volume of email that a user sends or receives on any given day [6]. Days with high email traffic are represented as large squares on the calendar, while days with low email traffic are represented as smaller squares. Users can also filter the visualization by specific contacts. Perer et. al created time series plots of one user's email volume over many years [4], with a separate plot for each contact. They used k-means clustering to group the 76 most active relationships into 9 groups, but did not find any noticeable correlation with user-defined groups. Perer et. al also found that just 2% of the dyadic relationships accounted for 31% of all email volume, suggesting a power distribution of dyadic relationships. As in the ***Social Network Analysis*** category, both of these projects relied on email metadata to visualize communication patterns, and did not look at the content of the email.

***Content Analysis.*** The final category of work does look at the content of emails in creating visualizations. Viegas et. al created *Themail*, which analyzes the email exchanges between the user and another contact, to find the most commonly used words by month and by year [7]. The words are plotted on a timeline by month, with the most common words for each month being listed at the top in larger and brighter font. Case studies with 16 users show that the users find these visualizations personally meaningful and start reminiscing, and mention that the visualization serves a purpose similar to that of a photo album.

A review of the literature shows that work on visualization of email communication falls into four broad categories. For my project, I have chosen to focus on communication patterns. Previous work on visualizing communication patterns has focused primarily on the volume of emails that are exchanged in order to determine the "rhythm" of the relationship between two users. As a further step, I am interested in visualizing the rhythm based on additional attributes, such as (a) the response time between each email (b) the length of each email (c) the word choice of each email. Furthermore, previous work treated all email exchanges between two people as one unit. In my work, I plan to break out the messages originating from the user, versus the responses from the other contact, in order to analyze who is initiating and/or responding. My goal in bringing additional attributes into visualizing "rhythm" is to detect more nuanced patterns, such as two people who exchange short, quick responses (e.g. setting up a meeting) or lengthy, measured responses (e.g. a debate over an article).

## Project Plan

A rough storyboard for my intended visualization is shown below. Each circle represents one email message. The visualization encodes 4 attributes: The responses from the other person are shown above the line and the responses from the user are shown below the line; the distance from the line shows the response time; the size of the circle shows the length of the message; the color of the circle shows the number of unique words.




My project completion plan includes the following parts:

- Download email data in structured format using Gmail API - done
- Preliminary analysis of basic stats such as who is emailed most often - done
- Parse email and reformat data to aggregate it at the user-user level
- Create visualization in d3 using reformatted data
- Build UI using JavaScript, allowing user to interactively create visualization for each of his/her contacts
- Package application so that each user can link it to his/her own email data

## References

1. Gmail Meter. http://www.gmailmeter.com/
2. Immersion. MIT Media Lab. https://immersion.media.mit.edu/
3. My Map (A Self-Portrait). http://christopherbaker.net/projects/mymap/
4. Perer, A., Shneiderman, B., & Oard, D. (2005) Using rhythms of relationships to understand email archives. Paper presented at the 22nd Annual Symposium of the Human-Computer Interaction Laboratory, University of Maryland, College Park, MD.
5. Perer, A., & Smith, M. (2006) Contrasting portraits of email practices: visual approaches to reflection and analysis. In AVI, ACM Press.
6. Viegas, F., boyd, d., Nguyen, D., Potter, J., & Donath, J. (2004) Digital artifacts for remembering and storytelling: *PostHistory* and *Social Network Fragments*. In Proceedings of the 37th Hawaii International Conference on System Sciences.
7. Viegas, F., Golder, S., & Donath, J. (2006) Visualizing email content: Portraying relationships from conversational histories. In CHI, ACM Press.