Hillary Clinton's email Collaboration

Shreenath Iyer

February 20, 2018

1 Objective

The objective of this visualization is to help visualize the network and activities of all the collaborators in the Hillary Clinton email scandal. This dashboard has been implemented using the dataset of the released emails available on Kaggle.

This visualization aims to provide information regarding the web of connections between the communicating parties, as well as draw knowledge of their behaviour with each other based on NLP techniques. Using textblob as the library, the dashboard shows the polarity of each sender to respective receivers based on the sent emails.

This visualization is not based on any political affiliation or interest; it merely tries to provide visual aid to a scattered dataset.

2 Implemented Views

The dashboard contains 4 discrete graphs, each meant to provide separate information regarding the collaborators. The graphs are as following:

1. A bubble chart containing all the senders
2. A force directed network graph showing all the receivers for any sender
3. An isometric joyplot showing the sentiment analysis for a given sender to a particular receiver
4. An interactive bar chart to display the number of emails sent per year by a sender

2.1 Bubble chart containing all senders

This is an interactive bubble chart containing all the senders listed in the database who have sent one or more than one email. The size of the bubbles is directly proportional to the amount of emails sent. This size of the bubble is indicative of that individual's amount of involvement in the collaboration. Hovering over a bubble gives the name of the sender, the sender id as listed in the database and the number of emails they have sent between 2009 and 2013. Clicking any bubble opens the force directed network graph for that individual.

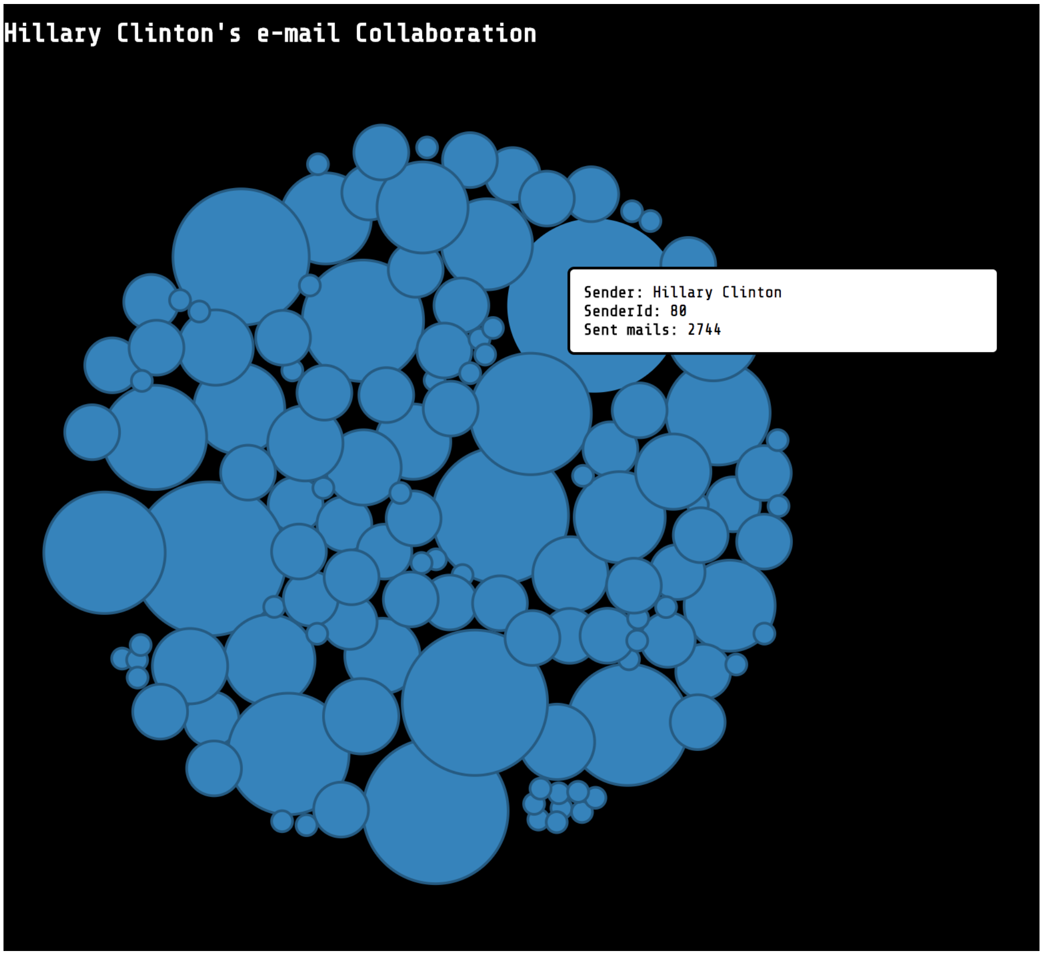


Figure 1: Bubble chart containing all senders

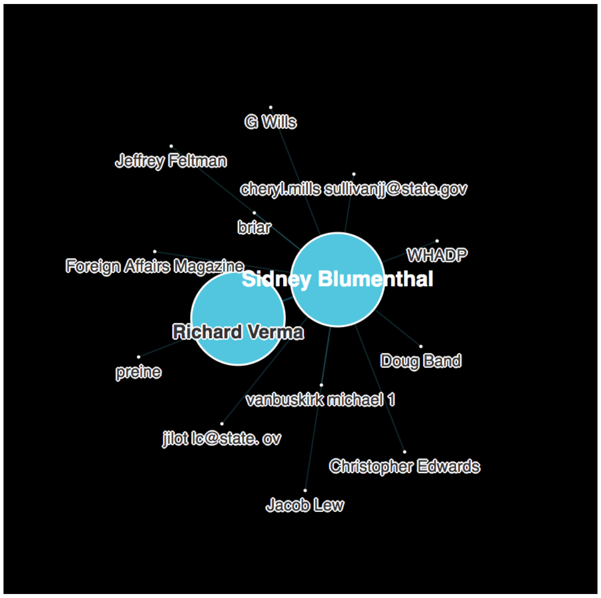


Figure 2: Force directed network graph

2.2 Force directed graph depicting the network

This graph contains the network of all non null recipients that a sender has sent emails to. This size of the node is proportional to how important that individual is to the entire scheme. While there are senders with more than 200 recipients in their list, there are also senders who have got no listed recipient as recorded in that database.

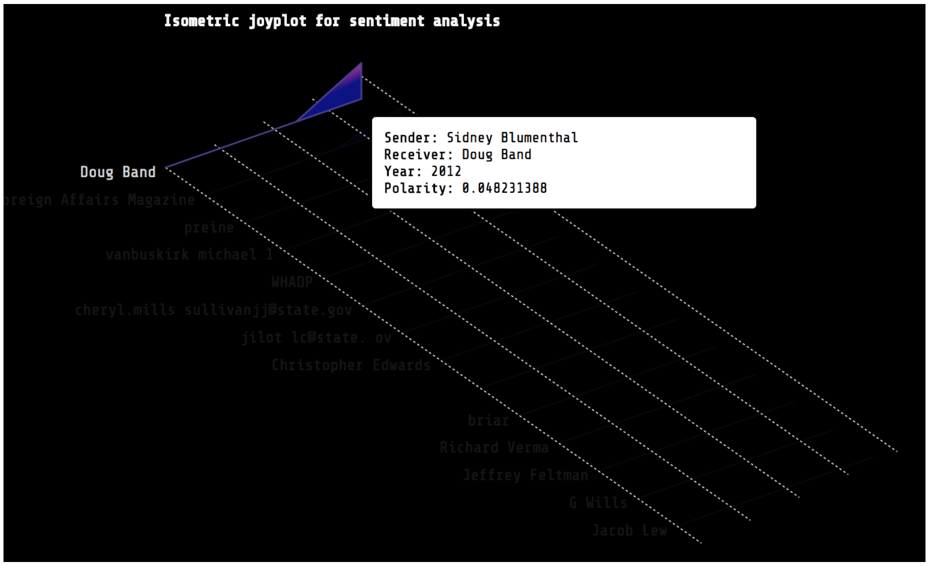
2.3 Isometric joyplot for sentiment analysis

This plot contains the sentiment analysis for a sender to any of their recipients. This is a key part of the visualization as it helps identify the relationship a sender shares with their recipients. This information is derived from the emails they have sent over a period of time and running those emails through Python's textblob library.

Hovering over any axis, gives the maximum or minimum polarity the current sender shares with the selected recipient.

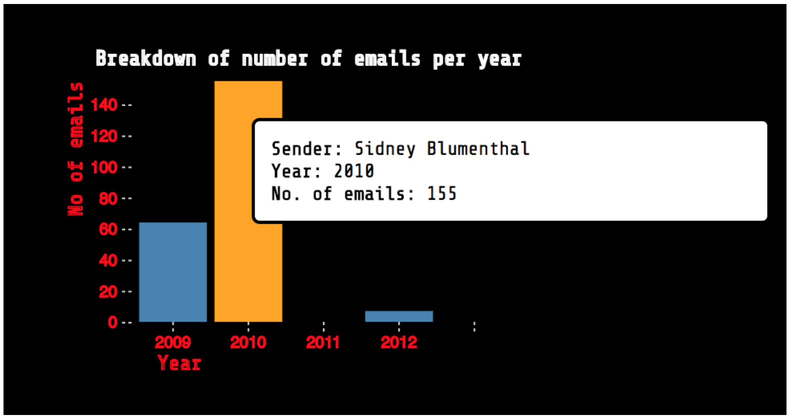
2.4 Bar chart containing number of emails per year

This graph breaks down the total number of emails for the selected person into the number of emails sent per year between 2009 - 2012. Hovering over any bar gives the number of emails sent by that person that year.



Sender: Sidney Blumenthal
Receiver: Doug Band
Year: 2012
Polarity: 0.048231388

Figure 3: Isometric joyplot for sentiment analysis



Sender: Sidney Blumenthal
Year: 2010
No. of emails: 155

Figure 4: Bar graph showing number of emails

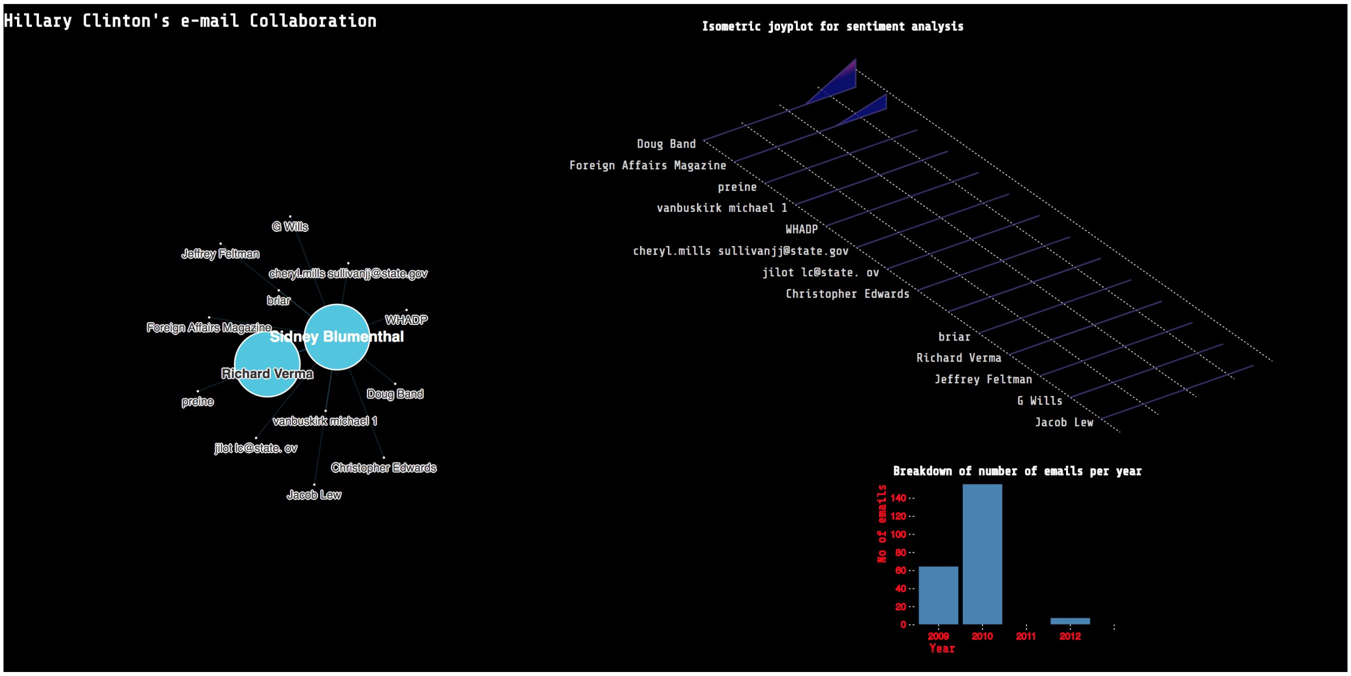


Figure 5: Network graph, isometric joyplot and bar graph

3 Usage

The dashboard opens by loading a bubble chart containing all the senders in the database who have recorded at least one email to an non-null entity. On clicking on any of the senders, a network of their contacts opens up on the same svg. It also opens another isometric joyplot graph that contains the sentiment analysis for each person to a respective recipient and a bar chart containing the number of emails sent by that person in a given year.

Clicking any node in the force directed network graph regenerates the bubble chart again and also removes the isometric joyplot.

4 Selection of Charts

I selected the bubble chart because I needed to display just the senders in the beginning. Since this was one dimensional data, I decided to use a simple bubble chart with the tooltip functionality for added information.

Next, I decided to use the force directed network graph because I wanted to show a web of all the recipients of any selected sender. I wanted to use as little space as possible without compromising on the amount of data retrieved. Hence, I chose to toggle between two svgs on the same division.

I chose to implement the isometric joyplot for sentiment analysis because the joyplot curves the slope and that was important for me to display something gradual like emotions. Another reason to go with this graph was that I could afford the data looking compact in the end without making the graph too messy. As I could always focus on the selected element, the amount of recipients did not matter.

Finally, I chose to implement a bar chart because it was the most straight forward and the cleanest way to represent the 2 dimensional data of the number of emails sent by a selected person per year.

5 Insights

The Hillary Clinton email database contains 513 listed people and 850 aliases. There are nearly 8000 emails exchanged that have been recorded in this database, and this isn't even the entire database of all the emails; this is just what the FBI chose to release.

Initially, it was impossible for me to figure out the relationship between two entities just by looking at the database. However, when I collected the tuples I needed and ran it through textblob, sentiments and relationships became evident. It was easy for me to visualize who contacted whom frequently and how the tone of their emails generally was. Such information is extremely valuable to the agencies interested in these conversations, for instance the DoJ or the FBI.

Not only does visualization save considerable effort and time, but it also conveys additional information that would have been impossible to extract from the database by just looking through it.

Over the past two projects, if there is anything I have learned, it's that choosing the right graph for the data is key. I have found myself running through the github page looking at examples after examples just to find the right match for my data. My idea is not to mould my data into any existing visualization, but rather find a visualization or a combination of visualizations that match the blueprint of how I want the data to be visualized in my head.

I have also tried to be miserly this time with the amount of space required for each visualization, like using the same div for two svgs. I realized that having the first visualization with the network on the same dashboard was just increasing redundancy.

6 Future Scope

Right now, the visualization does not validate for null values. This results in the ladder chart sometimes getting skewed out of the division. The visualization could also include information such as the top collaborators for any given sender, their most frequently used words, the frequency of sending emails to a specific receiver and so on. Right now, the database lacks a lot of information to make some of these visualizations possible.

7 Student Evaluation

My project was reviewed by Prerit Auti. Prerit asked me to add a few labels that I have duly added. She also asked me to reduced the collision strength in the bubble graph, reduce the node size and she also pointed out a bug in my visualization. I have incorporated all these changes.

She also asked me to align the ladder chart within the div for null values, something that I did not get the time to do.