Visualizing Hillary Clinton's Emails

Yihe Chen yc3076 | Palmer Lao pol2105 | Daitong Li dl2991 | Ziyue Shuai zs2285 | Eric Zhang ez2232

December 10, 2016

1 Introduction

In light of the recent U.S. presidential election race between candidates Donald Trump and Hillary Clinton, there have been large amounts of interest in the controversy involving Hillary's use of personal email accounts on non-government servers during her previous career as Secretary of State. After a number of Freedom of Information lawsuits were filed against the Department of State, the Department of State released on August 31, 2015 nearly 7,000 pages of Clinton's heavily redacted communications in PDF form.

Subsequent to this release, the data science competition website Kaggle released a sanitized version of the extracted content of the emails for public use and analysis. Our group was interested in analyzing and exploring this data given its connection to current events at the time.

2 Objectives

Our main objective with this project was to reduce the immense dimensionality and volume of the data so that the resulting output could be more easily digested by a data analyst or other interested party. We wanted to find patterns in the data that might be of interest, such as communities of receiving or sending parties that shared certain commonalities.

We tried to discover such patterns through the application of three methods: social network analysis, textrank, and cluster analysis.

However, rather than simply report on the results of such analysis, we believed that a more effective way to communicate results and allow for the generation of new insights would be to create an interactive visualization that would empower the user to both view our findings and explore the simplified data at will. This visualization would show the results of

3 Data Source

Some 7,000 pages of Clinton's emails were released by the State Department in PDF format [2]. Kaggle then scraped the text from these PDFs and hosted them as CSVs and SQL databases on the Kaggle Kernels platform. The dataset contains nearly 8,000 emails sent within Clinton's inner circle from December 2010 to September 2012. In CSV format, the dataset weighs in at just under 7 MB.

Unsurprisingly, the email dataset seems to be Hillary-centric. In addition, the histogram of emails received seems to be much more highly dispersed than the histogram of emails sent, which may suggest that many emails are sent to multiple people.

4 Methodology

4.1 An Overview of our Summarization Procedure for one Document

Because our end goal is to create a visual dashboard for our users, we have to be careful about our bookkeeping while we slice and dice our text data. A brief overview of our procedure from end-to-end for one e-mail is as follows:

1. Split the e-mail into sentences
2. Clean each sentence
3. Represent each sentence as a bag-of-words
4. Run TextRank to get a ranking of the bags-of-words
5. Return the top k sentences from the original e-mail associated with the top k ranking bags-of-words

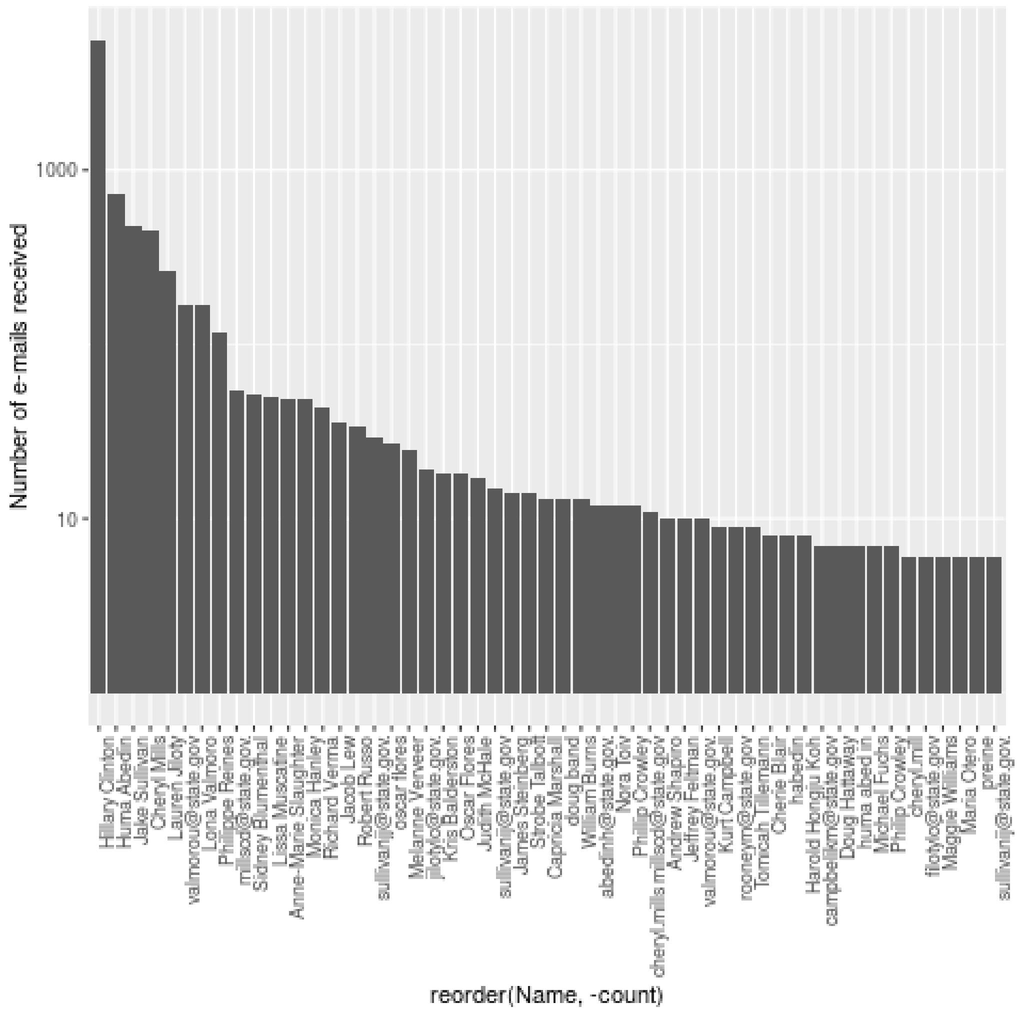

Figure 1: Histogram of how many emails some person has received

Of particular importance to our implementation is keeping track of how the bags-of-words are associated with the original sentences from the e-mail, as the bags-of-words are not nearly as informative to the user.

4.2 The TextRank Algorithm

We treat each e-mail as a self-contained document and attempt to summarize each document using TextRank [3], which is a derivative of PageRank [4] adapted to units of text instead of web pages.

The PageRank algorithm is a way of ranking web pages based on the way that they link to each

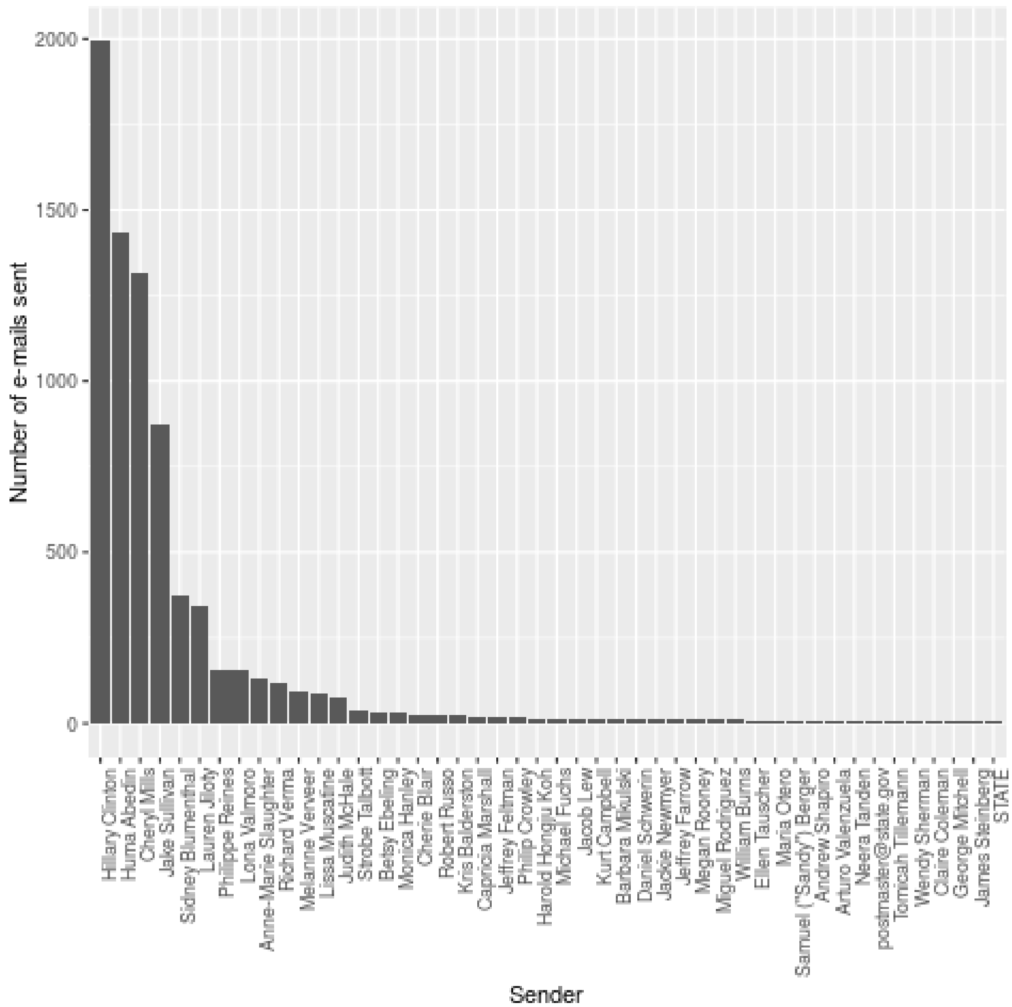

Figure 2: Histogram of how many emails some person has sent

other. The algorithm is based on a random walk model of the typical internet user. We assume the user starts on some arbitrary web page, and then either randomly clicks links on the current page or with some re-seeding probability, d, navigates to another random web page that was not necessarily linked to from the current one. The process of such a user's browsing history is clearly Markovian under this assumption. The addition of the re-seeding probability makes this Markov process irreducible and aperiodic, or equivalently, ergodic [5]. Ergodicity allows us to conclude that there exists a so-called stationary distribution over the set of all web pages. Moreover, this stationary distribution is unique and is the limit of a certain quantity.

The stationary distribution of an ergodic Markov chain has many interpretations. One is that the distribution assigns probability mass to each state, and this mass is equivalent to the probability of the Markov process to be observed in some state after the chain has mixed or lost its dependence on its initial position. The other is the "time-average" interpretation: the average proportion of time-steps a process will spend in some state tends to this state's mass in the stationary distribution.

Web pages that are assigned a large probability mass are thought to be relatively important in the sense that a random web surfer is more likely to visit it. One can sort the pages by the amount of mass assigned to them by the stationary distribution to get a ranking of pages by this notion of importance.

To turn this idea into something that can summarize text, we need to decide on two things. The first is the notion of an item to rank, or in our case, units of text. The second is the notion of how all items pairwise indicate each other's importance.

A typical choice of unit of text is either sentences or phrases. We choose to rank sentences for the sake of simplicity.

For text, we might imagine that a sentence that is highly similar to many of the other sentences is important in the sense that it may mix the contents of the most sentences to provide an insightful summary of the entire document. Naturally, we might choose to define the transition matrix of the Markov chain over sentences by the similarity between sentences, which now reduces our problem to figuring out a measure of pairwise similarity between sentences.

A naive view of sentences is to think of them as sets or multisets of words. An easy measure of similarity over sets S_1, S_2 is the Jaccard similarity:

$$J(S_1, S_2) \triangleq \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

which can easily be extended to multisets [6].

We opt to use the recommendation of the original TextRank paper [3]:

$$\text{Similarity}(S_1, S_2) \triangleq \frac{|\{w_k | w_k \in S_1 \& w_k \in S_2\}|}{\log(|S_1|) + \log(|S_2|)}$$

although there are many other variations that we did not have the chance to evaluate [1].

To actually compute the TextRank of a document represented as a collection of bags-of-words, we perform the following procedure:

1. For each pair of sentences S_i and S_j, compute their similarity Similarity(S_i, S_j) and store it in a matrix as entry $\tilde{M}_{ij}$.

2. Normalize $\tilde{M}$ to get a transition matrix M, representing a Markov chain over sentences.

3. Starting with an initial guess of the stationary distribution R_0, iteratively compute

$$R_{t+1} = dMR_t + \frac{1-d}{N}\mathbf{1},$$

   where d is the re-seeding probability, N is the number of sentences, and $\mathbf{1}$ is a column vector of ones.

4. After enough iterations, R_t will be close enough to the true stationary distribution of our damped Markov chain, and thus can be regarded as the TextRank of our document.

Running an entirely analogous algorithm on the entirety of the web (a graph consisting of about 322 million edges) took the authors of PageRank approximately 52 iterations for their ranking to converge, and about 45 iterations on a graph half that size [7]. They concluded that the number of iterations to convergence should be approximately logarithmic in the size of the network. Since documents tend to create relatively small and dense similarity graphs, we set our number of iterations to a highly conservative 10 iterations.

4.3 Cleaning the e-mails

Many similarity functions [1] as well as the one we implemented rely on metrics defined on bag-of-words representations of text. While the bag-of-words representation is a popular one, it's well known to have many deficiencies that are exacerbated by carelessly processed text.

One consequence of using a bag-of-words is that words that are not a character-for-character match will not be counted as the same. Much of our text cleaning effort is dedicated to ensuring that words that are indicative of sentence similarity are mapped to the same word, and that words that are not indicative of sentence similarity are removed.

The very first part of cleaning a sentence is simple: we set all of the characters to lowercase, as the capitalization of a word should, in most cases, not change its meaning.

Similarly, we remove any nonalphanumeric symbols from our text, as it seems that the method used to extract the text from the e-mails into a database left many wayward symbols (e.g.`\n`, the newline symbol). Any nonalphanumeric symbols were turned into spaces, and any extraneous whitespace was subsequently deleted.

The next transformation we apply is removing stop words. Stop words are words that almost every valid sentence in the English language contains, such as “the”, “a”, “as”, or “for”. Because stop words are so common, leaving them in the sentences would likely inflate the similarity between many bags-of-words. Leaving stop words in our text would effectively dilute our measure of how important a sentence is.

Our last problem is that different conjugations of words such as “run” and “running” will be counted as different although they ostensibly are referring to the same activity. We can attempt to mitigate this problem by applying a popular text cleaning procedure called stemming [8], which attempts to reduce all words to their roots. For example, “running” would be reduced to “run”, and “run” would stay the same. This allows us to measure text like “I am running for president” and “a run for office” as similar despite the fact that (after removing stopwords) none of the words are a character-for-character match. The `R` implementation we used is from the `tm` package, which implements the standard Porter stemming algorithm.

4.4 Network Visualization

In this subsection, we will present varied ways to visualize the network and some techniques to improve the visualization by using the attributes of the edges and nodes.

One key factor for effective visualization of a network is the graph layout. We have compared 15 different types of layout and decided to display our network with the Multidimensional Scale layout. The algorithm behind each layout scheme is beyond the scope of our discussion in this report, but

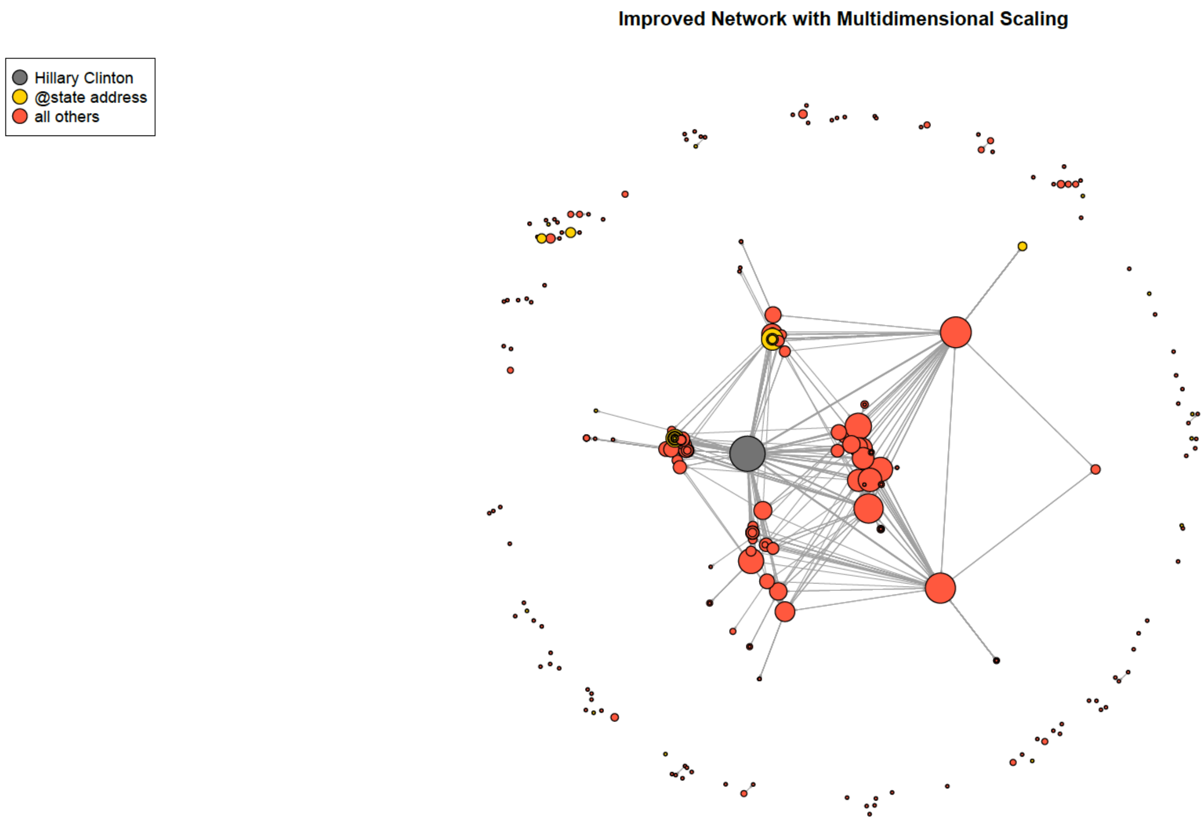

Figure 3: Improved Network after Deleting Low-weight Edges

we do invite the readers to see Appendix $\boxed{\text{APPENDIX REF}}$ and compare all 15 layouts we have tested.

The natural choice of color and size for nodes is based on the values of variables "`person_type`" and "`active_size`". See the legend in Figure ?? to understand the colorcode. Table 4 in the previous subsection suggests the distribution of node size is highly skewed, hence we need to rescale it to make it i) more reasonable as iGraph object input as the default is 15 and ii) have less variance. Inspired by the variance-stablizing transformation, we devised the following rescaling scheme in Equation (1). The side-by-side histograms in Figure 4 demonstrates that this rescaling scheme is effective. The similary log-transform rescaling was also applied to the edge weight, which is also highly skewed.

$$\texttt{rescaled active\_size} = \log \texttt{active\_size} + 1 \tag{1}$$

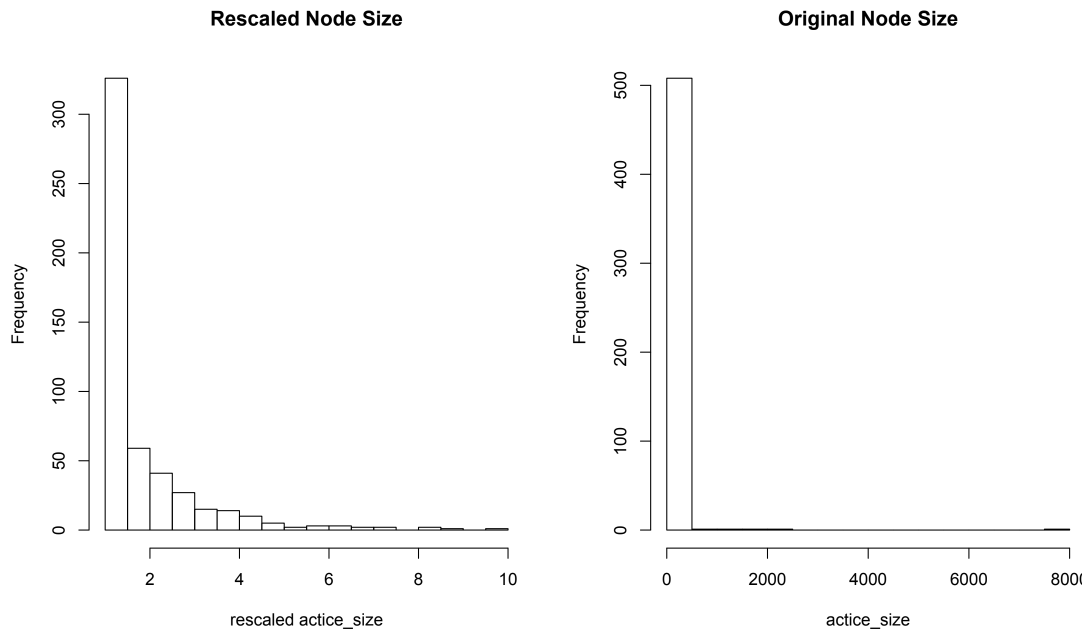

Figure 4: Rescaling of Node Size

Given a HIllary-centric network, we are also interested in examing the network by the e-mails she received and sent separately. Figure 5 shows the side-by-side sub-network by the node type.

4.5 Network Descriptives and Community Detection

We examine the Hillary Clinton Email network by studying properties of the smaller units - dyads (pairs) and triads (triangles).

We first calculate the **density** of this network, which gives us an idea about the extent to which the entire network is connected. The density of a directed network is calculated as Equation 2.

$$\text{density} = \frac{\text{total number of edges}}{\text{number of edges if all nodes were connected}} \tag{2}$$

The density of the HC e-mail network is $0.0028 < 0.01$, which indicates that this network is not at all well-connected. That is, less than 10% of the potential pairs within this network are connected.

A dyad is the smallest possible social group in a network. Such simplistic unit is worth studying in sociology, because two people in a dyad can be linked via some rather exclusive and intimate type

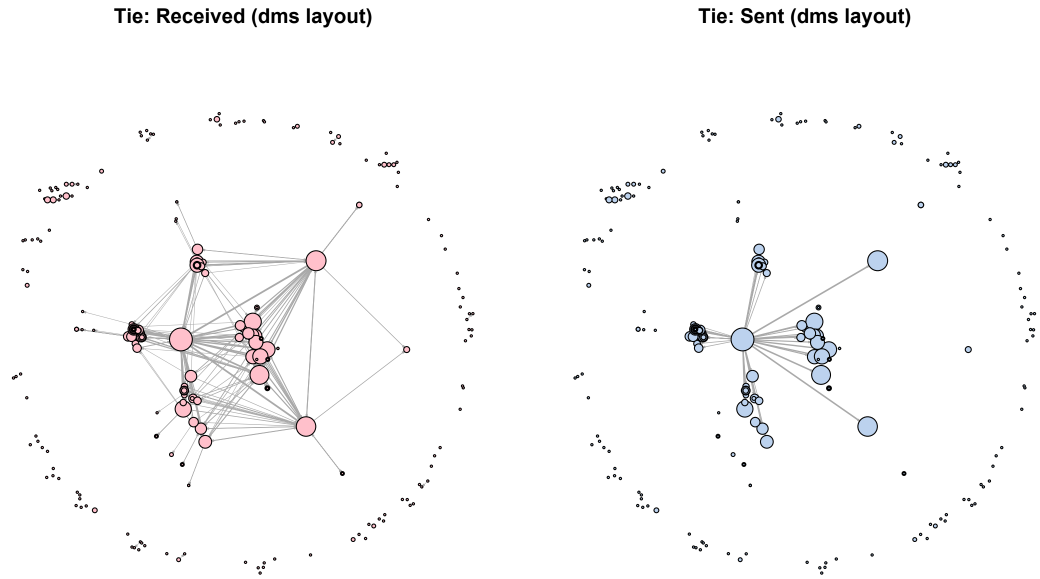

Figure 5: Visualizing Hillary Clinton Received and Sent E-mails

of connections, such as "romantic interetst, family relation,[... ,] partners in crime" [**?**]. Among all the 739 edges in the HC e-mail network, there are 106 mutual dyads (i.e. double directional links between two individuals), and 527 asymmetric dyads (the connection only goes one-way). Thus, the proportion of mutual dyads is $\frac{2\times106}{739} = 0.2869$. This property is called **reciprocity**, which in the context of an e-mail network, gives a rough idea about the dynamic of people's interaction as well as information flow in this network. That is, among all the connected people, about 30% of the communication went two-way.

The study of triads is also significant in sociology, as this type of human group is conceptualized to bear more communicational interactions than a dyad. "For example: adding an extra person, therefore creating a triad, this can result in different language barriers, personal connection, and an overall impression of the third person." [**?**]. Since the goal of our project is to use SNA as a auxiliary tool to extract important information (both e-mail content and people's association), the study of triads is beyond the scope of our report. However, we do include a brief frequency summary of different forms of triads in our network in the Appendix.

Based on our network, we can use clustering algorithms to find subgroups in the network

ID	Size	Individuals
11	4	"Bill Clinton" "Chelsea Clinton" "Tsakina Elbegdori" "dad mom"
4	6	"Betsy Ebeling" "Bonnie Klehr" "Doug Hattaway" "Robert Russo" "abdinh@state.gov" "bonnie klehr"
5	4	"Kris Balderston" "Mark Penn" "Marty Torrey" "Michael Fuchs"
3	9	"Harold Hongju Koh" "Jeffrey Feltman" "Jennifer Robinson" "Megan Rooney" "eichensehr kristen e" "hooke kathleen h" "johnson clifton m" "townley stephen g" "jake.sullivan h"

Table 1: Communities Detected by Modularity Optimization

and thus detect the communities. Because we are analyzing a one-person-centered network, we expect saliently large weight on this person's node and some of the associated edges, which may be coufounding for community detection. And as we are more curious about the unknown in the network, we propose to exclude those HC originated edges from the network. Recall that the sub-networks in Figure 5 - the one for eceived e-mails (Left, Figure 5) preserves the structure of the whole network (Figure **??**), while there are less edges coming out of the network center.

We applied the Fast Greedy Modularity Optimization algorithm [**?**] for the details) to the received E-mails sub-network to find community structure. The algorithm yields a lot of the single-node community and one community of extremely large size - this was all expected as we are dealing with a one-individual centered network. However, we can examine the communities of size 3 to 10 and we do obtain some informative communities that can pass on to the later stage of our project. Table 1 shows some of the communities we obtain through Modularity Optimization.

From Table 1, we see that with the exception of one individual "Tsakina Elbegdori" (the President of Mongolia), people in Community 11 are linked to Hillary Clinton (and each other) via family relation. Community 4 involves people who are related to Hillary Clinton's public relations and communications strategy matters. For example, Betsy Ebeling is an old and close friend of her, who has played an important role in building a positive public image for HC and antidoting the attacks that Hillary Clinton is untrustworthy. Within the same network, we see Bonnie Ward Klehr (with duplicated labels), who not only is a high school friend of Clinton's, but also designs jewelrys for HC to wear on campaign trail. And we also see people of the similar nature in this community, Doug Hattaway, who was in charge of strategic communications for HC's

2008 presidential campaign, and Robert Russo, Director of Correspondences and Briefings for the Hillary for America campaign this year.

With these communities we detected in our network, we hope to gain deeper understanding of Hillary Clinton's e-mails from a social interaction perspective and also help us efficiently browse this large corpus of documents/e-mails to extract as much important information as possible.

5 Results

TextRank proves to be quite useful when summarizing long e-mails. When prompted for the top 3 sentences of the following text, TextRank picks out the bolded sentences.

> Assume the following two pieces from Dawn in Karachi reached you through other channels, but just to be sure. The first is exactly the bad story I was worried about. The second is the Administration's attempt to shoot it down. **Bottom line (another four-letter word): "Whew!" FIRST STORY Talks under way for N-deal with US: Haqqani By Zulciernain Tahir Monday, 15 Feb, 2010 I 06:02 AM PST I LAHORE: Pakistan's Ambassador to US Husain Haqqani has said the government has started negotiating with the United States for an agreement on nuclear technology.** "The US is not sceptical about our nuclear programme. **Talks between Pakistan and the US for cooperation on atomic programmes are under way and we want the US to have an agreement with us like the one it had with India on civil nuclear technology," Mr Haqqani said at a reception hosted by Punjab Governor Salmaan Taseer on Sunday.** He said Pakistan would get 16 latest F-16 aircraft in June. He said although the expectations of Pakistan and the US with each other usually did not fulfil, both were indispensable for each other. "We have to largely depend upon the US for our defence related matters. "India is our main concern as it is buying weapons worth $100 billion from five countries, including China, and to balance it our relations with the US are very significant," he said and added that India had 5,500 tanks and there was a question against whom they would

be used. "We cannot be assured by statements that India will not wage a war against us." Giving a reason as to why Pakistan had to look towards the US for enhancing its military capacity and capability, Mr Haqqani said the European countries did not offer soft terms for buying weapons. "Ties with the US are important for a secure, stable and prosperous Pakistan." Mr Haqqani said Pakistan had also made it clear on the US that it should ensure a strengthened and Islamabad-friendly regime in Kabul before leaving. He said Pakistan had sought drone technology from America. "On one hand our innocent people are losing their lives while on the other Taliban leaders like Baitullah Mehsud get killed in such attacks," he said. Mr Haqqani said the US wanted to strengthen democracy in Pakistan and aid under the Kerry-Lugar Bill had started coming from January. In reply to a question, he said Pakistan's embassy in the US was working on diplomatic and legal aspects in Dr Aafia Siddiqui case and was making efforts for securing her release or transfer of her case to a Pakistani court. **SECOND STORY No nuclear deal with Pakistan, says US By Anwar lqbal Monday, 01 Mar, 2010 I 07:15 AM PST I WASHINGTON: The Obama administration has told Pakistan it would not get an atomic power plant or a civilian nuclear deal from the United States.** A senior US official, while briefing Indian journalists in Washington, said the United States was working closely with Pakistan to help meet its growing energy needs. "But nuclear power is not currently part of our discussions," and the United States had conveyed its decision to Pakistan, the official said. He said the administration had also told Pakistan that "there is no way they can get a civilian nuclear deal similar to the one the Obama administration has signed with India". The lndo-US civilian nuclear deal, the official said, was "specific to India only and there is no thinking going on in the administration to create a template for it."

In our opinion, the three sentences are fairly representative of the entire text. The entirety of the e-mail is about two stories, and the summary produced captures both headlines as well as another sentence from one of the articles.

6 Model Validation

6.1 Choice of hyperparameters for TextRank

There are a few parameters to choose for TextRank. For any use of TextRank at all, one must certainly decide on the re-seeding probability, d, as well as some criterion to decide to stop the iterative estimation of the stationary distribution.

To better understand the effects of the re-seeding probability, d, we first decompose our Markov chain into two Markov chains, one of which exclusively uses the re-seeding effect, and one of which never uses the re-seeding effect. Let M be the Markov chain which we eventually compute the stationary distribution of. Then M has the form

$$M = [M_{ij}]_{i,j=1}^{N} = [\mathbb{P}[X_{t+1} = j | X_t = i]]_{i,j=1}^{N}.$$

Denote the event that a random walker is re-seeded $\mathfrak{R}$. A trivial statement of probability is to decompose the above matrix as

$$\begin{aligned} M_{\mathfrak{R}} &= [\mathbb{P}[X_{t+1} = j | X_t = i, \mathfrak{R}]]_{i,j=1}^{N} \\ M_{\mathfrak{R}^C} &= \left[\mathbb{P}\left[X_{t+1} = j \middle| X_t = i, \mathfrak{R}^C\right]\right]_{i,j=1}^{N} \\ M &= M_{\mathfrak{R}}\mathbb{P}[\mathfrak{R}] + M_{\mathfrak{R}^C}\mathbb{P}\left[\mathfrak{R}^C\right] \end{aligned}$$

However, because the decision of a random-walker to be re-seeded is just an independent Bernoulli trial before she takes her next step, we find that $M_{\mathfrak{R}}$ is the matrix where every entry is $1/N$ and $M_{\mathfrak{R}^C}$ is merely the normalized matrix of similarities

By definition, $\mathbb{P}[\mathfrak{R}] = d$ and therefore

$$M = M_{\mathfrak{R}}d + M_{\mathfrak{R}^C}(1 - d).$$

At the extreme where $d = 0$, the ranking is the ranking of just the normalized similarity matrix if such a ranking exists. If $d = 1$, then at every step the random walker is choosing a random state from the entire state space with equal probability. It's easy to intuit in this scenario that

Rank 1:	4	4	4	4	4	4	4	4	4	4	4	4	4
Rank 2:	19	19	19	19	19	19	6	6	6	6	6	6	6
Rank 3:	6	6	6	6	6	6	19	19	19	19	19	19	19

Figure 6: Top 3 ranking sentence (represented by their index) for 10 TextRank iterations for values of $0.1 < d < 0.9$.

the ranking is a tie across the entire state space. A more rigorous explanation of this phenomenon invokes the concept of time reversibility. The eager reader can refer to [5] for a more detailed exposition of this concept.

For $0 < d < 1$, the ranking is both guaranteed to exist, and is typically not a tie across the state space in most applications. Therefore, it suffices for us to investigate how often or how wildly the ranks change for various settings of d. Most articles about the PageRank algorithm or derivatives thereof recommend setting $d = 0.85$. For our application, we find that the overall ranking is not very sensitive to d. The overall ranking is typically quite similar for a wide range of d with a few transpositions of adjacent ranks. One example is included in Figure 6, where the second and third most important sentences exchange ranks around the middle of our range of d.

The question of how quickly our initial guess converges to the correct stationary distribution is one that can be handled in a surprisingly theoretical fashion [**?**], and the proposition by the authors of PageRank that only logarithmically many iterations are necessary is quite prescient. Indeed, a central result in the theory of mixing times for discrete Markov chains is that for for any initial distribution x and ergodic Markov chain with transition matrix M with eigenvalues $\lambda_1, \lambda_2, ...\lambda_n$ and stationary distribution π,

$$\frac{\|xM^t - \pi\|}{\|x\|} \leq \max\left(|\lambda_2|, |\lambda_n|\right)^t$$

7 Conclusions

References

[1] Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. Variations of the

similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606* (2016).

[2] Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks.

[3] Kaggle. Hillary clinton's emails, 2016. [Online; accessed 08-December-2016].

[4] Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times.* American Mathematical Soc., 2009.

[5] Mihalcea, R., and Tarau, P. Textrank: Bringing order into texts. Association for Computational Linguistics.

[6] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: bringing order to the web.

[7] Ross, S. M. *Introduction to Probability Models, Tenth Edition*, 10 ed. 2009.

[8] Wikipedia. Dyad (sociology) — Wikipedia, the free encyclopedia, 2016. [Online; accessed 09-December-2016].

[9] Wikipedia. Multiset — Wikipedia, the free encyclopedia, 2016. [Online; accessed 08-December-2016].

[10] Wikipedia. Pagerank — Wikipedia,, the free encyclopedia, 2016. [Online; accessed 08-December-2016].

[11] Wikipedia. Triad (sociology) — Wikipedia, the free encyclopedia, 2016. [Online; accessed 09-December-2016].

[12] Willett, P. The porter stemming algorithm: then and now. *Program 40*, 3 (2006), 219–223.

8 R Commands and Data Configuration

8.1 Data Configuration for SNA

In the parlance of SNA, the nodes represent all 513 individuals involved in Hillary Clinton's e-mails based on the ``Persons.csv'' file from Kaggle. Each person has a distinctive Person ID, and some of the intuitive key players and their IDs are the following:

Table 2: Key individual by intuition

```
#    id           name
#    80 Hillary Clinton
#    81 Huma Abedin
#    87 Jake Sullivan
```

We also assign a type and a weight to each node, which are the `person_type` and `active_size` variables in Figure ??. The node type captures the characteristic of the person and is set up as below (see Table 3 for a simple overview for `person_type` by counts):

- `person_type = 3`, node `name` is Hillary Clinton;
- `person_type = 2`, node `name` contains "`@state`".
  That is, the person name is an governmental email address;
- `person_type = 1`, all the others,
  including people with full names, fragmented name, or unidentifiable aliaises.

Table 3: Overview of Node Type

`person_type`	1	2	3
count	355	157	1

The weight of each node measures the level of activeness of each individual. The weight for Person i is calculated as

$$\texttt{active\_size} = \text{frequency Person } i \text{ as Sender} + \text{frequency Person } i \text{ as Receiver} \tag{3}$$

`active_size` has to be at least 1 to appear in Hillary's e-mails. And a brief summary of the Node Size is shown in Table 4

Table 4: Overview of Node Size

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	1.00	1.00	33.32	2.00	7580.00

From Table 4, we see the distribution of `active_size` is highly skewed, as the quantiles are extremely small and close to each other, while the mean and maximum are extremely large. And we can in fact identify some key individuals by the node size extrema alone - four people with `active_size` $>$ 1000 are the three people in Table 2 and Person 32: `Cheryl Mills`.

To better describe the interaction, we use directed graph to depict the network based on the `''Receivers.csv''` and `''Emails.csv''` from Kaggle. Hence, we set up variables `from` and `to` in the Edges file in Figure **??** to capture direction of the email flow. The Edges file keeps track of a total of 9306 pairs of one-to-one interaction in 7945 e-mails. The discrepancy is caused by e-mails with multiple receivers (Hillary is one of the receivers or Hillary sent an e-mail to multiple people).

The edges also have two attributes: `weight` and `type`. The edge type is labeled as below

- `type` = "received", if the corresponding e-mail was received by Hillary (and other people);
- `type` = "sent", if the corresponding e-mail was sent by Hillary (to one person or more);
- `type` = "other', if the Sender is marked as "NA" in the original Kaggle data file.

Table 5 shows that Hillary Clinton's inbox had more incoming ("received") e-mails than outgoing ones. A side-by-side network graphs by edge type will be supplied in Subsection **??** in order to visually compare these two types of interaction.

Table 5: Overview of Edge Type

`type`	other	received	sent
count	13	6549	2744

The edge weight is also devised to identify different interaction pattern. The idea is to accumulate weights as the frequency of e-mail exchange between two individuals increases. But we also want to reward exclusivity of two individuals, so we lower the weight if the corresponding e-mails between two individuals involves other people. Therefore, we came up with the following weighting scheme for edge j where $j \in \{1, 2, \cdots, 9306\}$.

1. Start with initial weight, $\texttt{weight}_j = 20$;
2. Find the corresponding e-mail ID for edge j, $\mathrm{ID}_j = k$ where $k \in \{1, 2, \cdots, 7945\}$;
3. Count the number of Receivers for e-mail k, N_{kr} and the number of people Cc'ed, N_{kc};
4. Final weight for edge j is calculated as

$$\texttt{weight}_j = 20 - N_{kc} - N_{kr} \tag{4}$$

Before building the network, we collapse all the edges between the same two nodes by summing their weights and ended up with 739 distinct directional edges[1].

Table 6: Overview of Edge Size

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.0	17.0	18.0	225.8	49.5	25640.0

Since this is a one-person-centered network, summary statistics of edge weight in Table 6 also have an extremely large maximum comparing to the mean and 3rd quartile. While visualizing the network, we make use of the 1st quartile as the cut-off value and cull the edges to make the graph more informative.

[1] "Directional" in the sense that edges $A \to B$ and $B \to A$ were not collapsed.

9 Supplement Graphs and Tables

We tested 15 different layouts on our Hillary Clinton E-mails network

Figure 7: Exploring different network layouts