Visualizing Hillary Clinton’s Emails

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

December 6, 2016

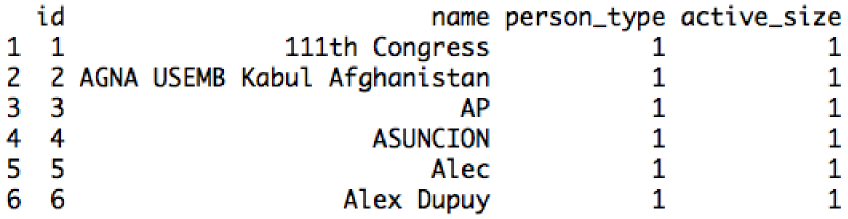
1 Social Network Analysis

Instead of exploring the e-mail content, we are also interested in examining the interaction and connections through the Hillary Clinton’s e-mail exchange. We focus on the receivers and senders of all the e-mails and build a social network based on the correspondence using Social Network Analysis (SNA) techniques from the study of Sociology. In subsection 1.1, we will walk through the data configuration catered to the process of SNA. The visualization of the network will be presented in Subsection 1.2. Following Subsection 1.2, we dive into the properties of the network using some common metrics in SNA and further explore the subgroups in the network to detect potential communities.

1.1 Data Configuration

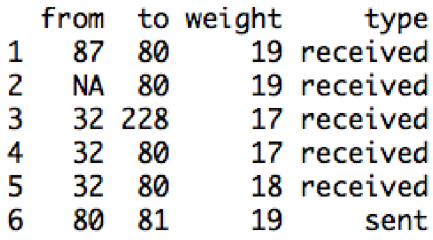
Wikipedia summarizes that SNA is the process of investigating social structures[1]. The structure consists of two parts, individuals and interactions, graphically characterized as nodes and edges in the network. And the data sets for our SNA are set up in such format (see Figure 1 and 2 for snapshots of the data sets).

Figure 1: Nodes file snapshot



```
  id                        name person_type active_size
1  1               111th Congress           1           1
2  2 AGNA USEMB Kabul Afghanistan           1           1
3  3                           AP           1           1
4  4                     ASUNCION           1           1
5  5                         Alec           1           1
6  6                  Alex Dupuy           1           1
```

Figure 2: Edges file snapshot



```
  from  to weight     type
1   87  80     19 received
2   NA  80     19 received
3   32 228     17 received
4   32  80     17 received
5   32  80     18 received
6   80  81     19     sent
```

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 1: 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 1. The node type captures the characteristic of the person and is set up as below (see Table 2 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 2: 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

$$\texttt{active\_size} = \text{frequency Person } i \text{ as Sender} + \text{frequency Person } i \text{ as Receiver} \tag{1}$$

`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 3

Table 3: 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 3, 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 1 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 2 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 4 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 1.2 in order to visually compare these two types of interaction.

Table 4: 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

$$\texttt{weight}_j = 20 - N_{kc} - N_{kr} \tag{2}$$

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 5: Overview of Edge Size

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

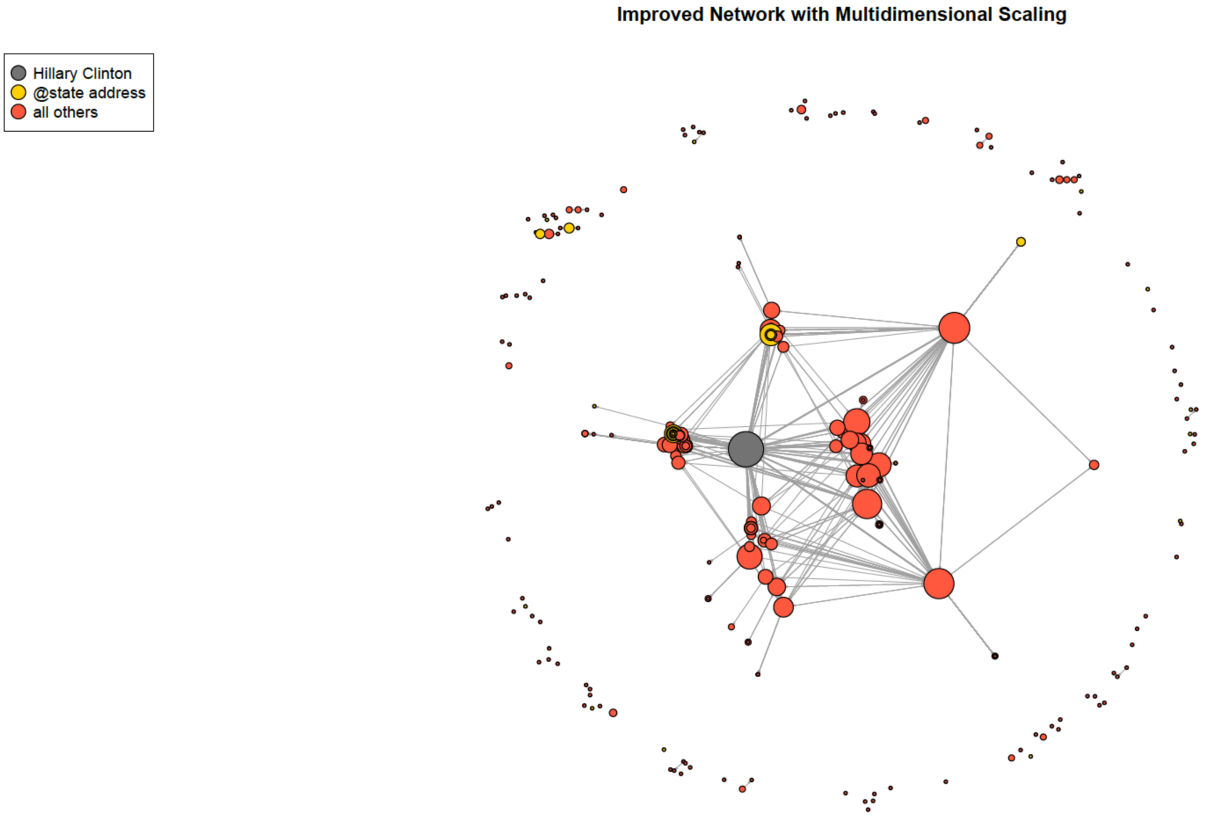
Since this is a one-person-centered network, summary statistics of edge weight in Table 5 also have an extremely large maximum comparing to the mean and 3rd quartile. In the upcoming subsection, we make use of the 1st quartile as the cut-off value and cull the edges to make the graph more informative.

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

1.2 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.

Figure 3: Improved Network after Deleting Low-weight Edges



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 we do invite the readers to see Appendix A Figure empty 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 3 to understand the colorcode. Table 3 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 (3). 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.

$$\texttt{rescaled active\_size} = \log \texttt{active\_size} + 1 \tag{3}$$

Figure 4: Rescaling of Node Size

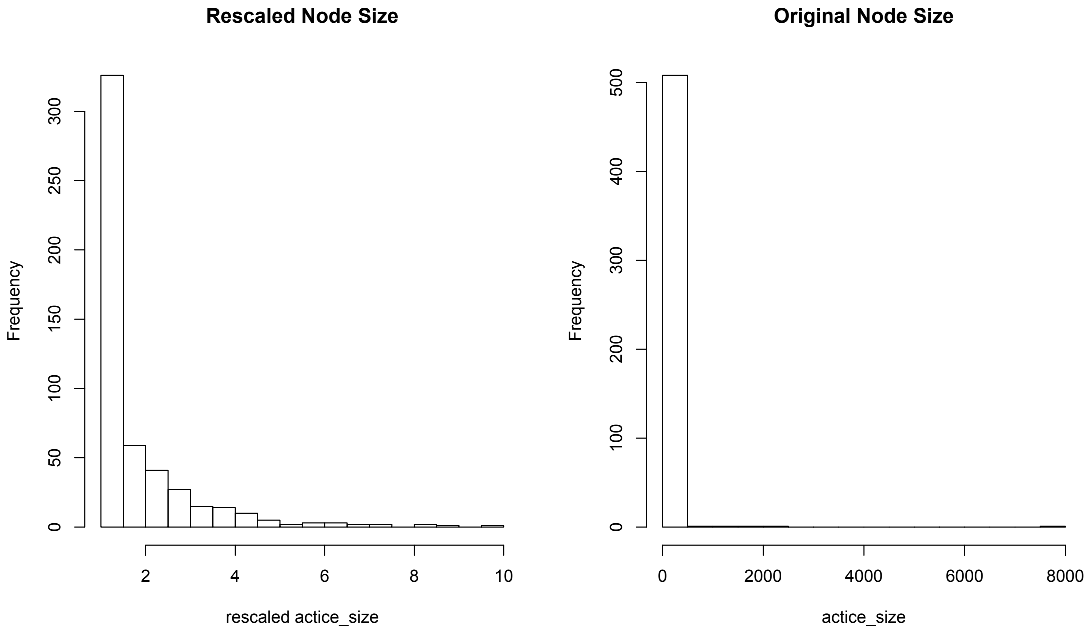
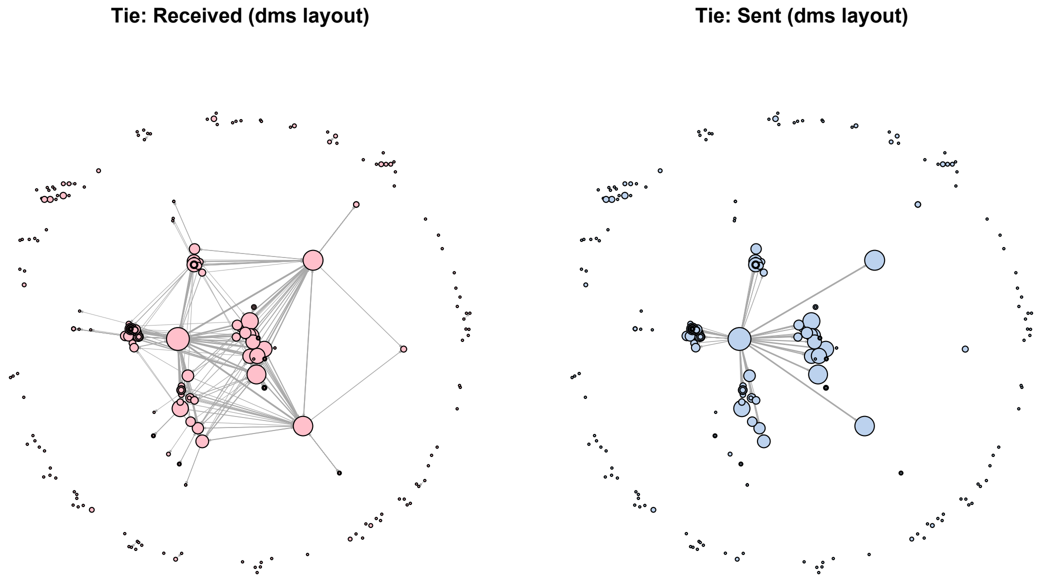


Figure 5: Visualizing Hillary Clinton Received and Sent E-mails



As we have mentioned in Subsection 1.1, we can split the network by the edge type.

1.3 Network Descriptives and Community Detection

Some statistics are calculated to describe the Hillary Clinton network. We first exam the **density** of this network

$$\text{density} = \frac{\text{total number of edges}}{\text{number of edges if all nodes were connected}} \tag{4}$$

By Equation (4), 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 potentially related pairs were actually connected.

A R Commands

B Supplement Graphs

References

[1] https://en.wikipedia.org/wiki/Social_network_analysis