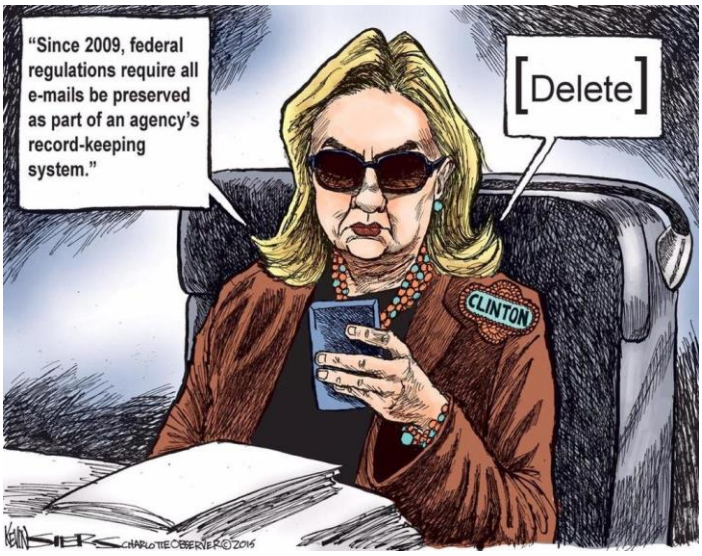
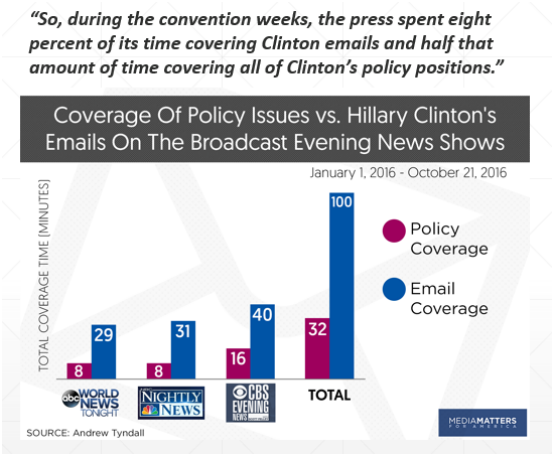
Team Project Report: Samantha Melchiori & Ono Gantsog
IDS 564: Social Network Analytics
April 24, 2018

Hillary Clinton Emails

Context, motivation and research question(s)



The 2016 presidential elections (Hillary Clinton presidential compaign 2016, n.d.) were clouded with the backdrop of the Hillary Clinton email scandal. Starting in 2009, and through her tenure as the US Secretary of State, Hillary Clinton sent thousands of emails from a private email server installed in the basement of her New York home covering issues of government matters later considered classified information (Hillary Clinton email controversy, n.d.). The discovery of this private email server brought up concerns of national security and conduct that became the foundation for several investigations for which Secretary Clinton has often publicly blamed for her defeat. These emails, now nearly 8000 publicly released, provide a rarely seen, intimate view into the personal and professional network of political advisors surrounding a top-ranking political figure. (Annie Karni, 2015) Why do we care? These emails and their public media typhoon represent a scandal that she herself believes could have cost her the presidential election. (Boehlert, 2016)



- What can we conjecture about the significance of those advisors and their relationship to Hillary by understanding the network created from emails sent to and from her personal server??
- What about their relationships to each other?
- How big, how close and how connected was her political advising circle?
- Who can we imagine were the most influential figures and how did the relate to her publicly and politically outside her email?

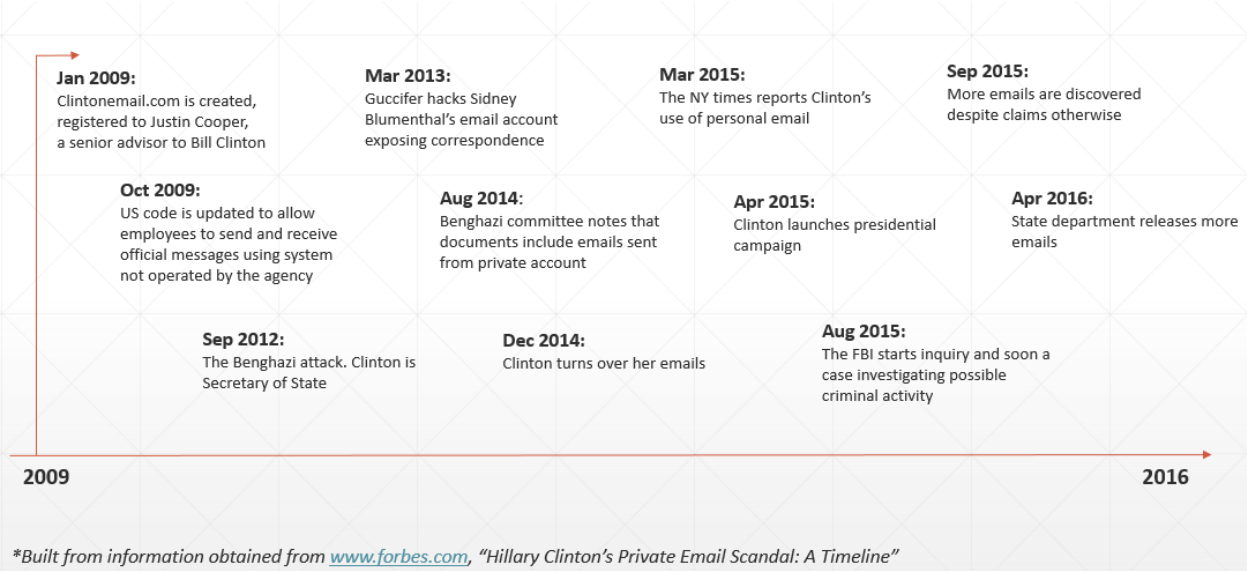
Background on Data

Clinton emails were released in SQL files by the US State Department in August of 2016 and can be obtained through APIs and or files available in several locations including Kaggle and include:

- Original SQL files released
- CSV extracted files with email contents
- CSV files with emails alias conversions

These files represent emails through Hillary Clinton's private emails server from Jan 2009 on. The emails represent primarily communication to and from Hillary and include some unrelated, non-political personal content (Kaggle, 2016). However, they show the importance of several close advisors and within her closest circle, and some of the relationships between those advisors. Below, a high-level timeline from the inception of her server until the State Department release of the emails in 2016:

Object 1: Timeline of Hillary Clinton Email Scandal



(Forbes, n.d.)

Describe the Network Data, Summary statistics and Visualizations

The state department released SQL files containing nearly 8000 emails sent Hillary Clinton's personal email server. These can be accessed from data files located on Kaggle as well as several other locations. Extractions from the SQL files allow for easy access of email contents through CSV files. However, significant cleaning of files is required to create uniformity in data structure, naming conventions and to eliminate NA values. To evaluate the network structure, we will clean this data to be inclusive of only the "to" and "from" information. We import nearly 8000 "to" and "from" records. Eliminating NA and emails with multiple recipients leave us with a data frame containing 7662 observations. Separately we will also extract subject text information to gather see if we can gather insights by applying text analytics.

The nodes in our network represent the individuals sending and receiving emails on Hillary's private server. The edges, when shown in a directed graph, represent the direction of the email (to-from). When represented in an undirected graph, the edges are the represent both (to-from and from-to) or just the existence of an email communication path between two individuals. When represented with varying sizes, the node size represents the number of emails the individual sent and received. When represented with varying edge width, the edge weight represents the frequency of the specific email path.

Some basic information on our network:

- Number of nodes: 303
- Number of edges: 7663

The graph constructed from the imported data is neither weighted nor simple, meaning each individual email sent even between same individuals is represented by a unique edge. For our analysis we will use two basic graphs, a simplified graph of our original data and a simplified and weighted graph from our original data. We will show varying representations and visualization of this data and create several induced subgraphs to help with our analysis.

Our simplified graph network:

- Number of nodes: 303
- Number of edges: 438 directed, 374 undirected

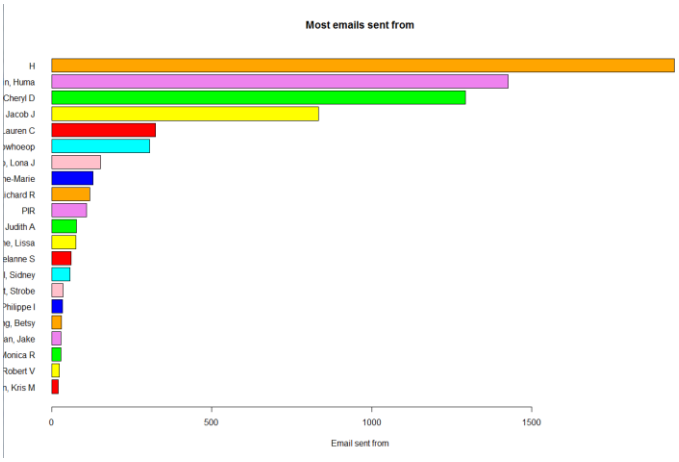
Below is an example of the node format in our data:

```
> V(g_emails)$name
  [1] "H"                    "Abedin, Huma"          "Russo, Robert V"
  [4] "Sherman, Wendy R"     "Sullivan, Jacob J"     "Mills, Cheryl D"
  [7] "Flores, Oscar"        "Hanley, Monica R"      "Schwerin, Daniel B"
 [10] "Nides, Thomas R"      "Jiloty, Lauren C"      "Valmoro, Lona J"
 [13] "Burns, William J"     "michele.flournoy"      "Cheryl"
 [16] "Reines, Phillippe I"  "brian"                 "Mikulski, BAM"
 [19] "ntanden"              "Lew, Jacob J"          "Huma Abedin"
 [22] "rsloan"               "stalbott"              "Balderston, Kris M"
 [25] "Steinberg, James B"   "wburns"                "mhcaleja@state.gov"
```

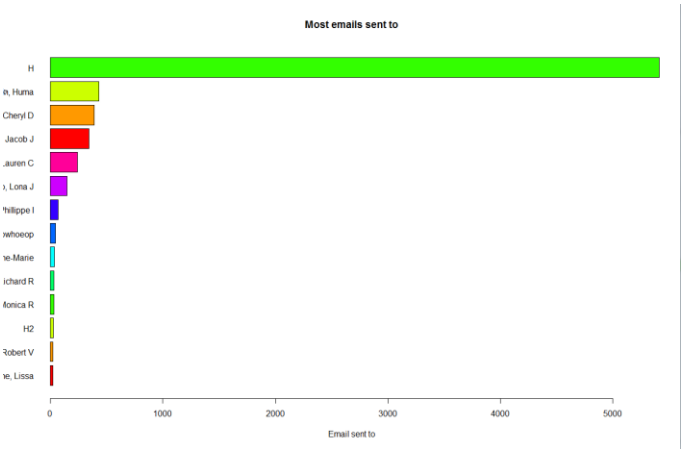
Immediately we can see that we have reduced our edges by greater than an order of magnitude, meaning that a substantial number of our emails are traveling to and from similar members of our group. We should see some significant impact from this in our weighted representations.

First let's examine who are the prominent email senders and recipients with a simple barplot analysis.

Object 2: Barplot of most emails sent from > 20



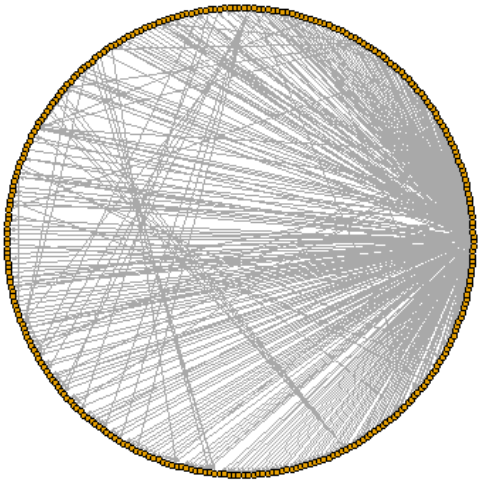
Object 3: Barplot of most emails sent to > 20



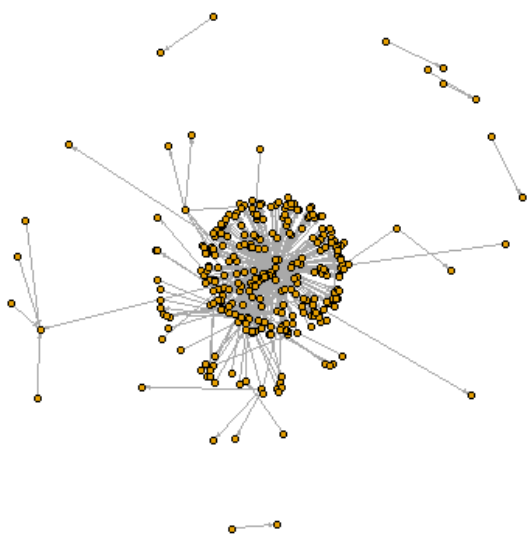
Here we are looking for those individuals who sent or received greater than 20 emails. We can see several names common among both graphs, where substantial communication has occurred to and from an individual. Some of the prominent and common names include: Hillary ("H"), Huma, Cheryl, Jacob J, Lauren and Lona. We can anticipate these individuals will play a central and important role in Hillary's network.

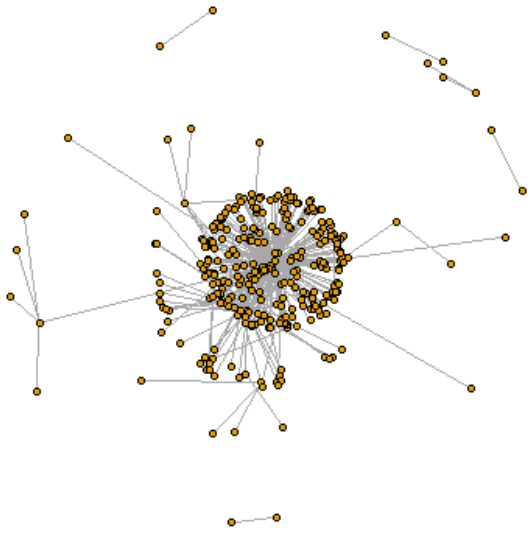
Below we will plot our simplified networks:

Object 4: Simplified network, circular layout



Objects 5 & 6: Simplified network, kamada kawai layout (directed vs undirected)

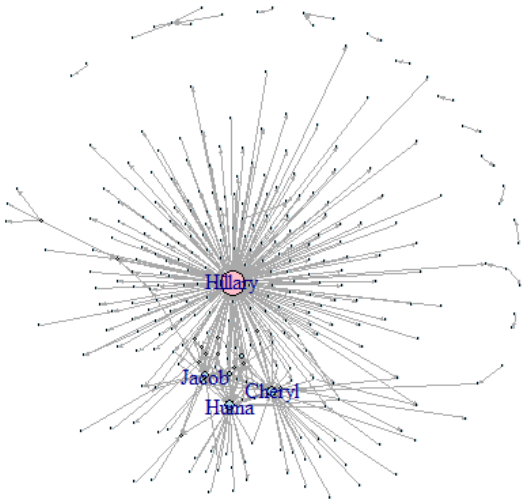




We see immediately the "more" connected and more centered area of our graph and then the unconnected periphery. Noted, these graphs (both directed and undirected) are not connected, not "strongly" and not "weakly". It is also difficult to see the difference between the two graphs. It will be good in our analysis to focus on communities and areas of higher centrality and connectivity and / or higher degree to better understand the inner circle of Hillary's network.

Below is our full network, weighted and directed:

Object 7: Simplified network, kamada kawai layout, scaling adjusted, nodes sized by graph strength, edge width adjusted for edge weight (directed)



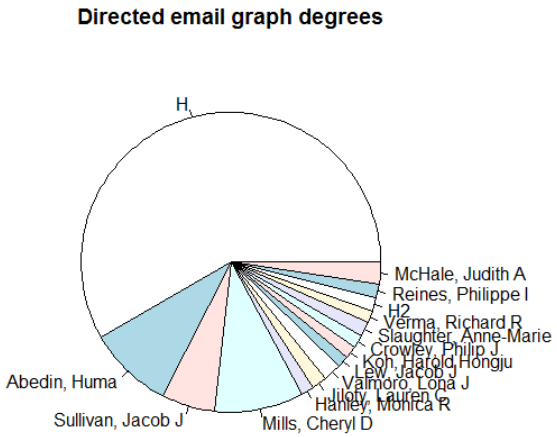
This view allows us to see the expansiveness of the whole network, as well as the limitedness of prominent players. We see only a few nodes of significant size, correlating to those individuals at the tops of our barplots. The majority of emails are sent from and occasionally to Hillary from in volumes less than 20 in total and do not appear indicate a significant relationship.

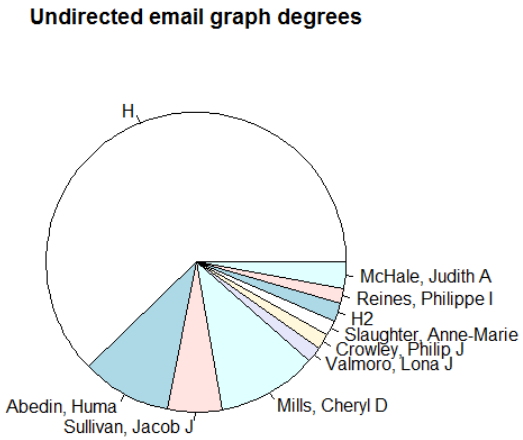
Analysis and results

To gain a slightly deeper understanding of our network, we will evaluate some of its basic characteristics beyond node and edge numbers. The degree of a node in a network (or it's connectivity) is the number of connections or edges the node has to other nodes. Network degree is a reflection of the number of emails

sent and received from each node, so our network degree should resemble our barplots for number of emails sent and received.

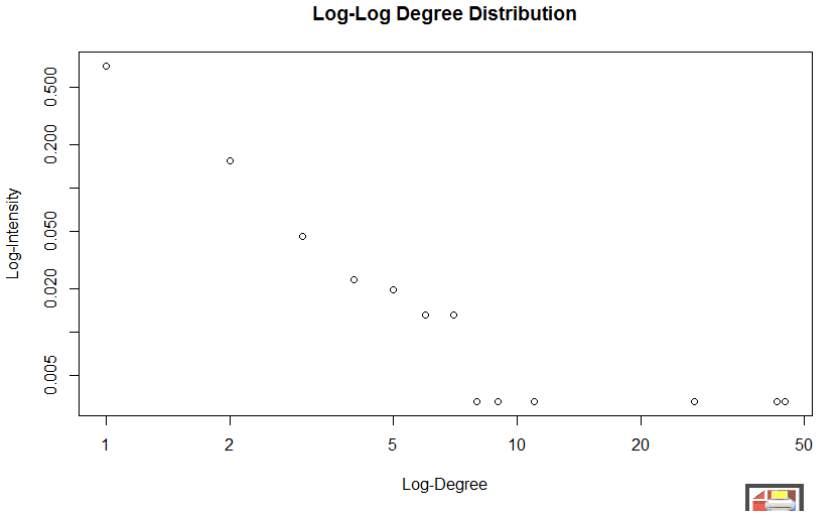
Objects 8 & 9: Pie Chart of Network Degree (Directed vs Undirected)





Again, our results can easily be predicted by our barplots and it is no surprise that degree is most significant for Hillary as the vast majority of emails travel to and from her. We start to see an emerging pattern in the significant presence of players such as Huma Abedin, Jacob Sullivan and Cheryl Mills.

Object 10: Log-Log Degree Distribution (undirected)

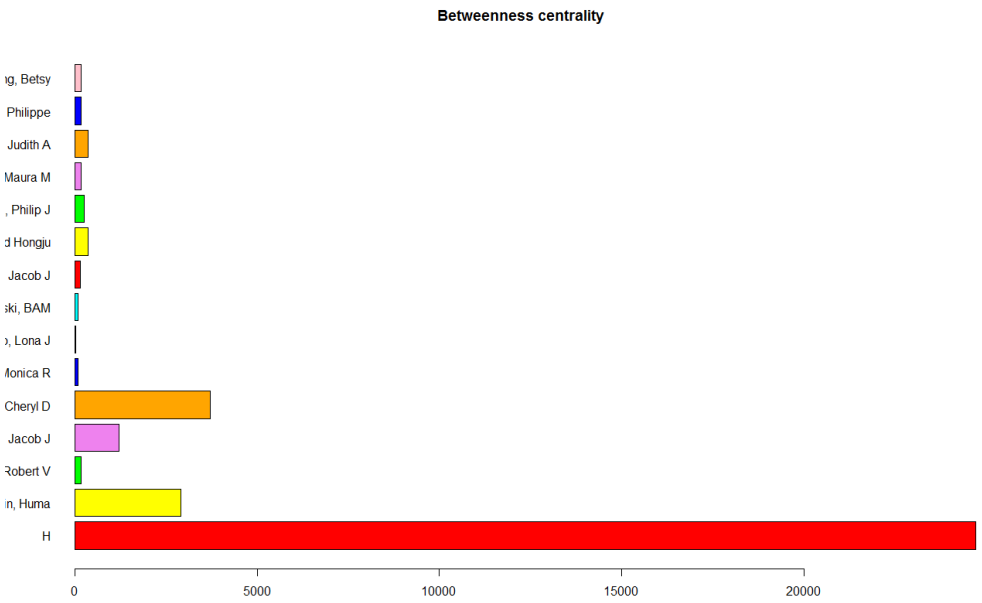


From our graphs, including our log-log degree distribution we see strong evidence of preferential attachment (the rich get richer). Few nodes of very high degree and many nodes of very low degree.

We know that beyond several core members email frequency trails off nearly nothing between pairs of nodes almost immediately.

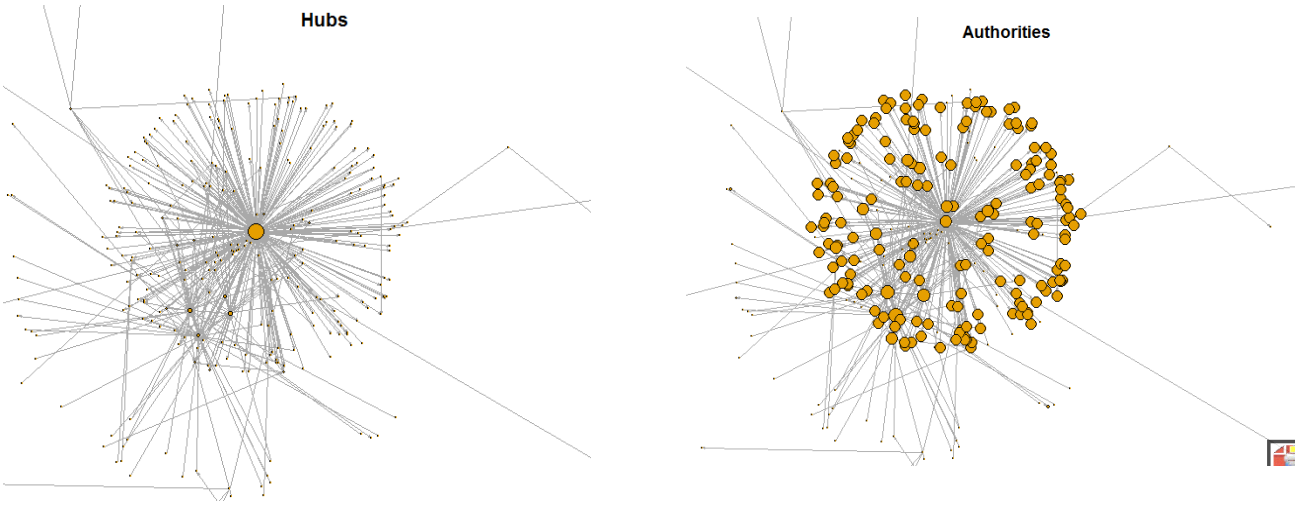
Betweenness centrality measure how much a node is "between" other pairs of nodes. High betweenness indicates the extent to which a node forms a bridge or the singular path between other nodes. Below we can see the nodes with the highest betweenness centrality for our network:

Object 11: 15 nodes with highest betweenness centrality



Similar to network degree we see Hillary, Huma, Cheryl and Jacob stand out with high betweenness centrality. In particular Hillary has a uniquely high betweenness central. If we recall our graph of our network and our understanding that she was the prominent sender and receiver of emails, we understand that she is the center of our network. Functionally she serves as the center node. The point through which nearly all email communication flows. If we examine the hubs and authorities in our network, which should see this represented clearly and we do. We see below Hillary is our clear and most important network hub and her those to who she most commonly emails are the most prominent authorities. We would likely consider these her advisors.

Objects 12 & 13: Network hubs and authorities



Beyond understanding Hillary's centrality to the network and being able to identify her circle of advisors in the most general sense, we further describe the network through the characteristics of average path length, network diameter and transitivity (a measure of global clustering):

- Average path length: 2.32
- Network diameter: 5
- Transitivity: 1.3%

These characteristics are easy to see in the network graphs, but generally the network is often only 1 or 2 degrees from Hillary so a short average path length and diameter are expected and seen. Transitivity is also extremely low. This describes how infrequently nodes are part of clusters, meaning Hillary is sending to and from advisors but rarely are those advisors or individuals emailing each other, indicating that many of them may not have a strong relationship outside of Hillary or the fact that this is Hillary's private server is masking their relationship. Probably the latter is true, given the likelihood that this is a closer knit group political circle than indicated by this metric.

Despite low transitivity, we can find some interesting clusters within the closest inside circles of the network.

Object 14: Table of Cliques by Clique length

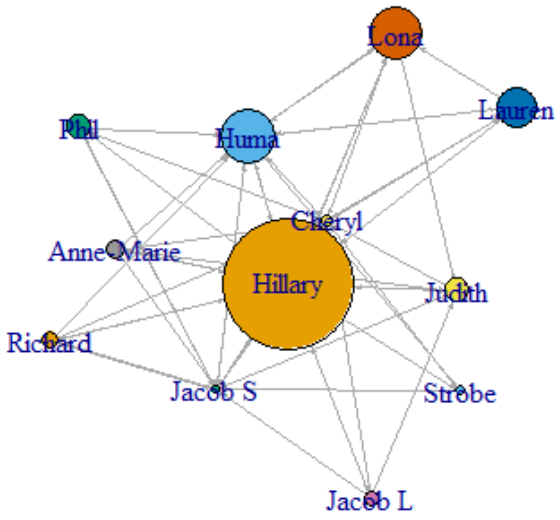
```
> table(sapply(cliques(g_emails_simpl), length))

  1   2   3   4   5
303 374 114  40   7
```

This network only has cliques up to 5 and relatively few cliques greater than 2 given. We expect this given the low transitivity of our network:

- Cliques of 4 == 40
- Cliques of 5 == 7

To examine the more connected portion of the graph structure, we induce a subgraph from only the large cliques ==5 in our original undirected, simple graph.

Object 15: Induced Subgraph from Cliques == 5 in Length

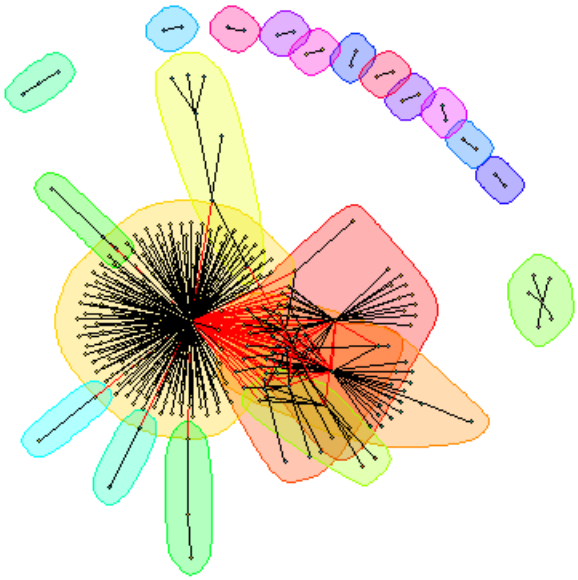


Here we can see some connectivity in the inner circle of Hillary's close advisors. This is some evidence that this group is probably more connected than our graph may represent. Additionally, those in the inner circles are definitively connected. As we expect, here again we see Huma, Cheryl and Jacob.

Similar in concept to identifying cliques, we can examine our network for larger communities. Using the fast greedy community detection algorithm on our complete network (undirected), we can see the large community surrounding Hillary and several subsequent medium communities around her close advisors

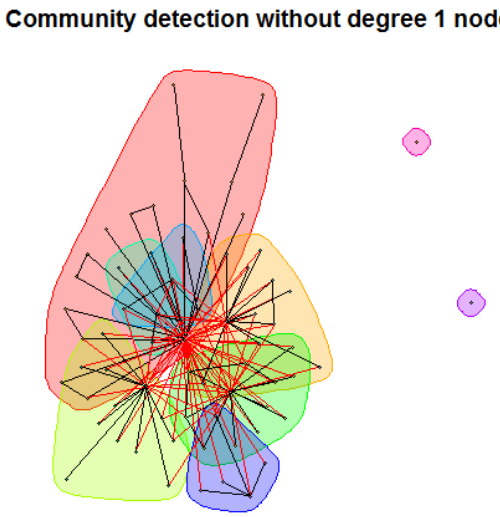
and finally many small communities where we see an email connection between individual users (in particular in the unconnected regions of the network).

Object 16: Fast Greedy Community Detection on full undirected network



However, more insightful we can remove nodes of degree 1 or less to improve our visibility of the more core network and those communities:

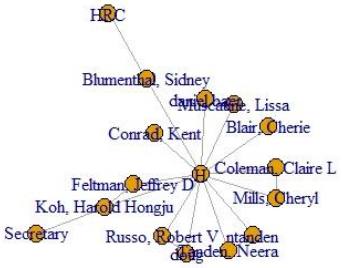
Object 17: Fast Greedy Community Detection without nodes degree 1 or less



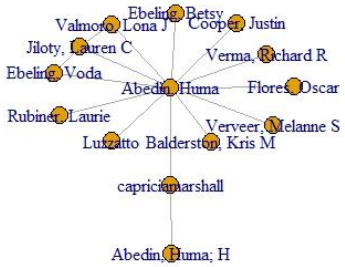
Now we can better see the 5 communities of Hillary and her advisors. Plotting examples from communities 1-3 we can see explicitly the members of Hillary's community and also those belong to the communities of her close advisors Huma and Jacob.

Objects 18-20: Community networks 1-3 with nodes degree 1 or less

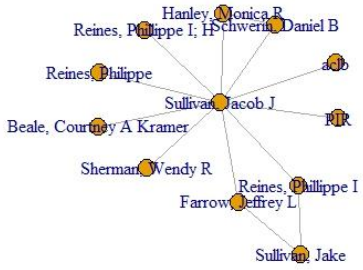
Community 1 (after removing degree 1)



Community 2 (after removing degree 1)



Community 3 (after removing degree 1)



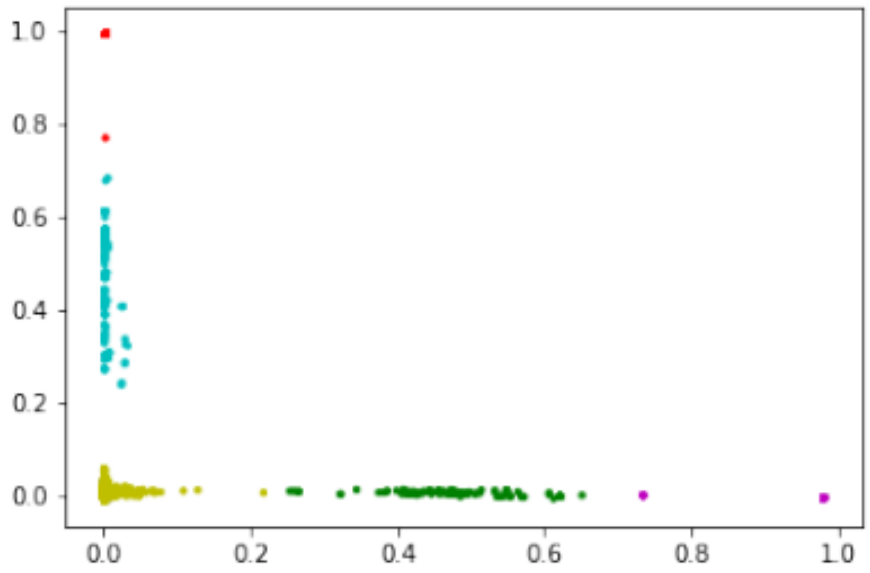
These community plots and the induced subgraph from the large cluster seem to tell us the most about Hillary's direct community of advisors and the communities and connections that surround them.

Community significance

Community	**Significance**
Community 1	p-value = 0.0007193
Community 2	p-value = 0.8749
Community 3	p-value = 0.6127
Community 4	p-value = 0.8414
Community 5	p-value = 0.7506

Last, we evaluate the contents of the sent and received emails through text analytics to understand if we can make some deductions on the probable content or nature of the email communications. This content was given the appearance, in particular through the media coverage, of being highly controversial and potentially covering confidential content on Benghazi and other affairs of Secretary business.

Object 21: Text analytics clustering from email subject contents



Our analytics and clustering is not overwhelmingly conclusive, but we do find some themes among them.

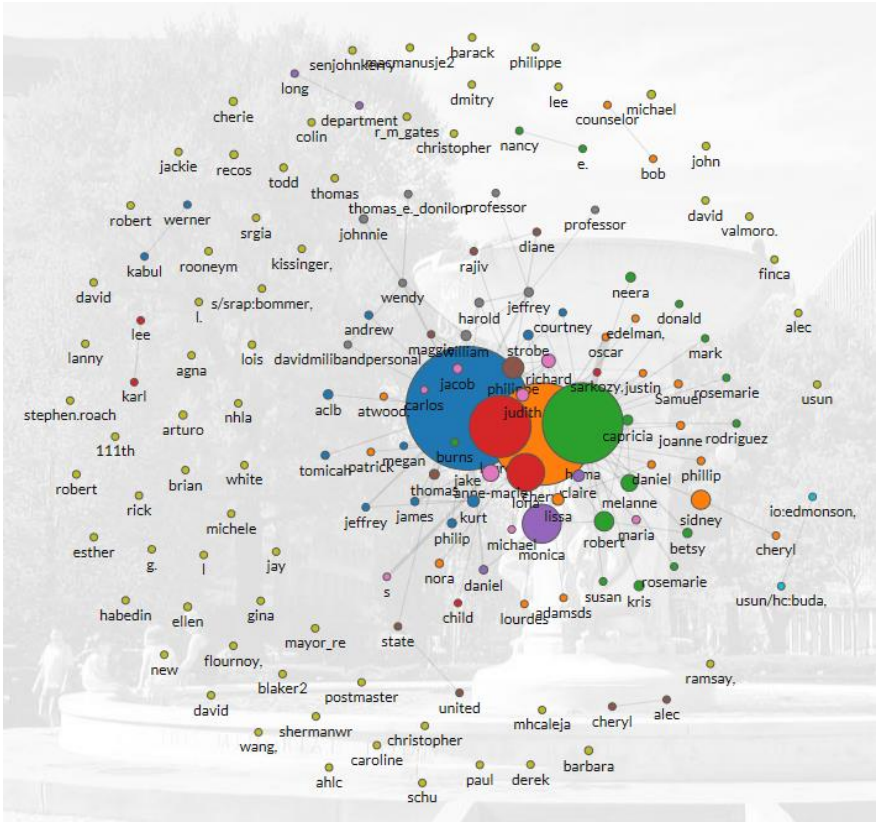
- Schedule: A substantial amount of her correspondence seems to focus on the arranging calendars, events and meetings
- Speech: A number of emails around the construction and development of speeches or talking materials
- Haiti: A number of emails around Haiti, a focus for Hillary around this time.

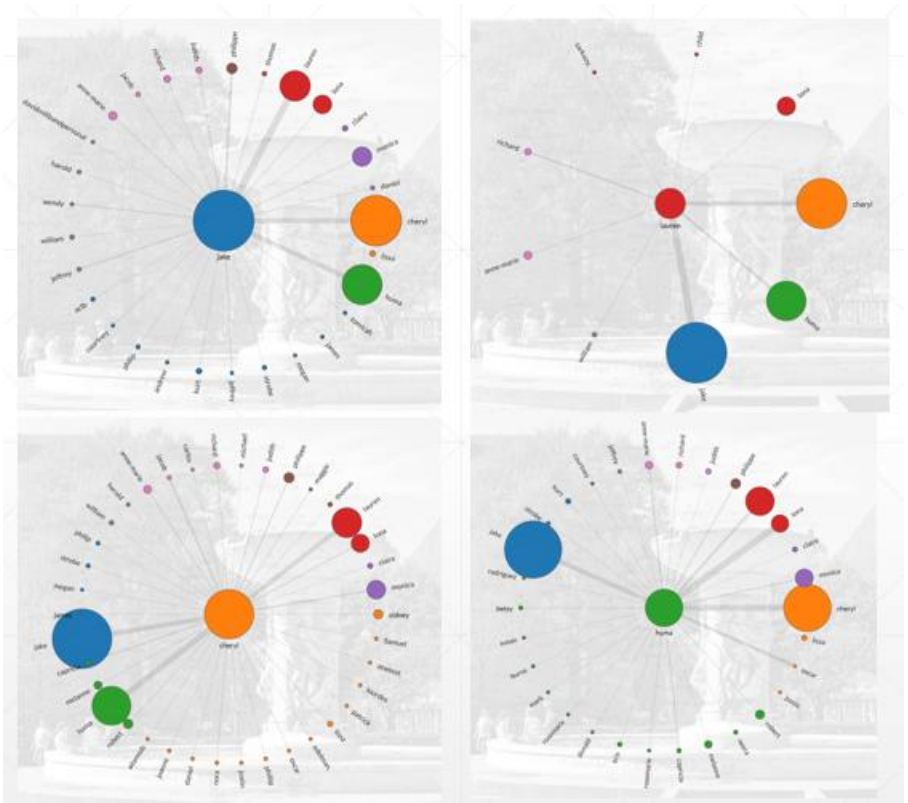
Surprisingly, we don't see Benghazi or Libya, given the amount of media focus on related classified information in her emails. The actual amount of content was likely limited or not in the subject of the email. Perhaps the overwhelming majority of emails were benign and in general over-dramatized by the media. (Bradner, 2016)

Other Visualizations and Research Available

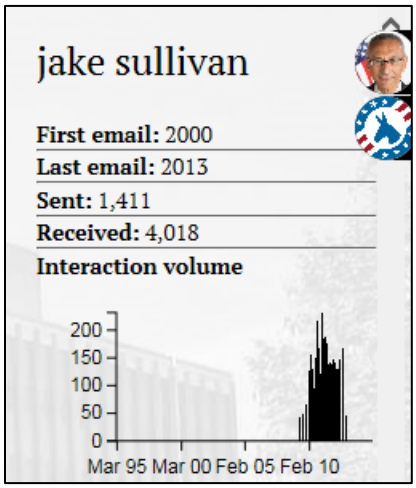
Before reviewing our final conclusions, it is helpful to compare our visualizations others available. A particularly excellent representation has been constructed by MIT. The visualization is similar to ours in the sizing of nodes based on their network strength. Their graph is not however directed nor does it weight edges. However, the network is fully interactive, allowing the user to interact with each individual inside the network, opening up the network community around them. This was comparable in many ways to our community detection. Shown below is the full network and then induced subgraphs for Jacob, Cheryl, Huma and Lauren. Also shown below is the individual user summary for Jacob Sullivan. This provides a very informative timeline view of both volume and time series information on their email participation. (A.Hidalgo, 2016)

Objects 22-26: Hillary Clinton Email Network Visualization (MIT) and induced subgraphs for Jacob, Cheryl, Huma and Lauren



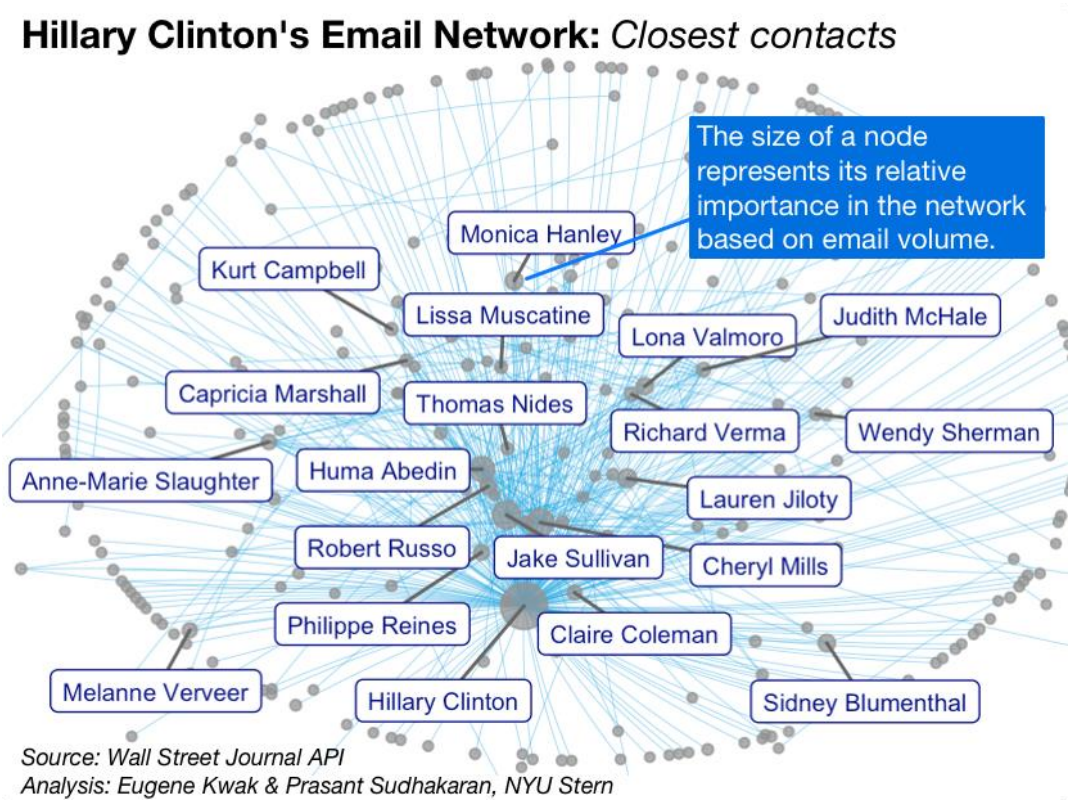


Object 27: Individual User Summary (Jake Sullivan) (MIT)



Similar but also nicely laid out was the Hillary Clinton Email Network constructed by CNBC. This visualization was not interactive but also provided an excellent view of Hillary Clintons closest advisors and their relatively importance as estimated by volume of correspondence. This network was however developed from a slightly different data set and some differences can be observed due to this.

Object 27: Hillary Clinton's Email Network (CNBC)



"Based on a theory that measures how individuals and groups interact within their network, the chart reveals Monica Hanley, Cheryl Mills, Huma Abedin and Sidney Blumenthal as the most important nodes. This group signifies a high degree of centrality, or the individuals that communicated the most in the network besides Hillary Clinton. Blumenthal's role has been especially controversial, because he was not a government employee but appeared to have handled classified information." – CNBC "Hillary Clinton's Email Network" (Sigdyal, 2016)

Insights and conclusions

"Throughout her many years in public life- through all the disappointments and triumph, the scandals real or alleged – Clinton has surrounded herself with protectors: a tightly knit Praetorian Guard, mute and loyal…. The State Department emails provide ample evidence of the hermetic circle that exists around Clinton – a world of gatekeepers and advisers, but favor seekers too." – Vanity Fair (Ellison, 2016)

It should not be surprising to that what we find in our email network is a vast communication network throughout the government but only a few truly important authorities or advisors and that the network between them is highly linked and well connected. Our overall subject analysis supports that in majority this email server was likely used predominantly to schedule and coordinate government and personal meetings and other relatively benign content. However, it is likely that the media craze was not completely unwarranted as it seems likely that it was also used a private means to communicate among her close advisors (Cushing, 2016). We see this in the core analysis when we dig in the more connected center of our network. It is likely here that the concerned information perhaps was shared and the true scandal was born.

We could glean perhaps a greater understanding of messaging content with a deeper view into text analytics and this might enlighten our view on what was contained in these emails. Additionally, it would be interesting to compare her private network to her public network, her actual public and formal business relationships. Additionally, comparison to networks in news or twitter might be revealing. These might both provide a broader and more factual reveal of the network and the importance of the individuals that surround her.

However, we do see some truth in our analysis. Who are those individuals we find and do we know their importance? Huma Abedin (Huma Abedin, n.d.), Cheryl Mills (Cheryl Mills, n.d.) and Jake Sullivan (Jake Sullivan, n.d.) are known to be some of Hillary Clinton's closest confidants and this is clearly illustrated in our network. We know they worked tightly for years in a close-knit circle around her as her main support systems and political advisory committee.

Bibliography

(n.d.). Retrieved from Forbes: https://www.forbes.com/pictures/eglg45hlfmd/january-2009-clintonema/#5c15dd0a3bcd

(2016). Retrieved from Kaggle: https://www.kaggle.com/kaggle/hillary-clinton-emails/data

A.Hidalgo, C. (2016, Nov 4). Retrieved from MIT media lab: https://medium.com/mit-media-lab/what-i-learned-from-visualizing-hillary-clintons-leaked-emails-d13a0908e05e

Annie Karni, G. T. (2015, 12). Retrieved from Politico: https://www.politico.com/magazine/story/2015/10/hillary-clinton-2016-emails-213241

Boehlert, E. (2016, 11 2). Retrieved from Media Matters: https://www.mediamatters.org/blog/2016/11/02/how-media-s-email-obsession-obliterated-clinton-policy-coverage/214242

Bradner, E. (2016, 10). *CNN politics.* Retrieved from CNN: https://www.cnn.com/2015/09/03/politics/hillary-clinton-email-controversy-explained-2016/index.html

Cheryl Mills. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Cheryl_Mills

Cushing, T. (2016, 6 24). Retrieved from SOTT: Signs Of The Times: https://www.sott.net/article/320821-Killarys-private-server-a-massive-security-headache-meant-to-dodge-FOIA-requests

David Easley, J. K. (2010). *Networks, Crowds and Markets.* Cambridge University Press.

Ellison, S. (2016, 11). Retrieved from Vanity Fair: https://www.vanityfair.com/news/2015/10/hillary-clinton-inside-circle-huma-abedin

Eric Kolaczyk, C. G. (2014). *Statistical Analysis of network data with R.* Springer.

Hillary Clinton email controversy. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Hillary_Clinton_email_controversy

Hillary Clinton presidential compaign 2016. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Hillary_Clinton_presidential_campaign,_2016

Huma Abedin. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Huma_Abedin

Jake Sullivan. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Jake_Sullivan

Sigdyal, P. (2016, 3 27). Retrieved from CNBC: https://www.cnbc.com/2016/03/27/hillary-clintons-emails-what-does-the-data-show.html