# DNC Email Analysis

MATTHEW DUPONT, Marquette University, United States
JAMES WANG, Marquette University, United States
JOE CHUDZIK, Marquette University, United States

An exploratory study of the 2016 Democratic National Committee email leak through network analysis and topic modeling. We analyze various network statistics such as node strength and degree distribution to discover important nodes within the email network,and then perform Latent Dirichlet Allocation to create a statistical model that provides more insight and context to what topics the email network was talking about.

Additional Key Words and Phrases: DNC, data science, machine learning, topic modeling, wikileaks

**ACM Reference Format:**
Matthew Dupont, James Wang, and Joe Chudzik. 2019. DNC Email Analysis. 1, 1 (December 2019), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The 2016 Democratic National Committee email leak is a collection of emails stolen by an alleged Russian intelligence agency [10]. More than 20,000 emails were published in June and July of 2016. The published leak contains emails from many different key DNC staff members. Within some of the emails leaked, there was DNC off-the-record correspondences with reporters from many different news outlets, including Politico and the Washington Post. Many other researchers and teams have performed analysis on this dataset to find that the DNC had an ongoing bias against Bernie Sanders' campaign, leaning in favor of Hillary Clinton. Our goal in this analysis is to find other underlying features within both of the network analysis and primarily the text of the emails.

We scraped the email text, header, recipient and sender information from WikiLeaks in order to attain text information and network level information. We performed topic modeling on this text component to provide a more meaningful and contextual addition to the dataset. Finally, we created subgroups of different parts of the dataset and performed topic modeling on those subgroups to effectively compare the subgroups across each other.

In this project, a novel method is proposed to analyze email networks. The way is a combination of network analysis and text analysis so more details can be detected in this process. One of the benefits is that the real topics can be found, so researchers can know whether the topics reflect the roles of the people in this network. Thereby, researchers can dig deeper into the networks.

Authors' addresses: Matthew Dupont, Marquette University, 1250 W. Wisconsin Ave. Milwaukee, United States, matthew.dupont@marquette.edu; James Wang, Marquette University, 1250 W. Wisconsin Ave. Milwaukee, United States, y.wang@marquette.edu; Joe Chudzik, Marquette University, 1250 W. Wisconsin Ave. Milwaukee, United States, joseph.chudzik@marquette.edu.


XXXX-XXXX/2019/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

### 1.1 Specific Problem

This paper focuses on using two different approaches to analyze the structure of the DNC network as a whole, as well as the differences between subgroups within the network. Since it is already found that there was a committee bias against Sanders' campaign, the purpose of this project is to find important people within the network and to detect different subgroups along with the email topics of each subgroup.

The research questions are generally answered in this paper. In the second section, a review of the related research is presented. In the third section, the research methods are described. The results are presented in the fourth section. The fifth section is the discussion. The last section is the conclusion and future works.

## 2 LITERATURE REVIEW

Emails have been used many years, and email network analysis is a very popular research topic because emails contain a lot of information. Emails can readily be represented as a social network, describing particular accounts as nodes and emails themselves as edges. This allows for a substantial amount of analysis of the underlying structure of the organization from which the email network was derived through application of Social Network Analysis (SNA). Node centrality metrics can be applied to identify significant nodes. Cluster detection algorithms also can be applied to try to discern structural groups underlying the email networks. Examples of applications of these techniques on other email networks include Rowe, et. al. , who apply SNA node centrality techniques to identify hierarchical structures in the Enron email dataset [6], and Bird et. al. who apply similar techniques to analyze an email network of Apache developers, classifying status and relative importance of developers by application of centrality metrics [1].

In addition to structural and role-based identification in email networks using SNA, email bodies are notable for their rich text availability. This allows for a different angle of analysis, in the form of Latent Dirichlet Allocation, to identify latent topics in the text of email bodies. While SNA provides useful in describing the structural elements of an organization, LDA provides context as to the domain-specific information discussed in the network. Blei et. al. describe this technique, its usage, and its application in detail in their paper of the same name [2].

In this paper, we use a "combined arms" approach, utilizing both techniques. The approach used allows for multiple layers of analysis of the target network. LDA analysis of important topics provides initial insight and overview of the content of what was important to the users of the network. SNA then contextualizes the how the network was structured, which individual addresses were important in the discussion, and how individual nodes grouped together to discuss topics. LDA is then applied to individual sub-structures identified by SNA, providing for a more nuanced understanding of individual subgroups and providing a deeper layer of understanding of the structure of the network. In some respects, SNA acts as both a blueprint of the network, as well as a lens, directing LDA, while LDA classifies and colors the network, bringing relevance the structures outlined by SNA.

Limited information was available via Google Scholar and other literature reviews with respect to application of SNA or LDA specifically on the DNC email network. Some informal analyses were available, such as that by GitHub and blogspot user tweinzirl, which provided examples of useful visualizations [8]. Additionally, academic application of similar techniques on other email datasets were identified, to serve as a comparison and inspiration for the techniques to be used for this dataset. In particular, the Enron email dataset, examined by Shetty et. al. and Rowe et. al, provides another excellent example of a leaked email dataset on which SNA has been applied. These papers provide excellent comparisons with respect to overall network characteristics. This holds particular

relevance to our dataset with respect to verification of authenticity - as the DNC email dataset was likely produced by Russian intelligence agents with intent to affect the United States' political process, a greater degree of scrutiny should be applied to how representative these emails are, and whether they demonstrate abnormal characteristics suggestive of tampering.

Bird, et. al. provides an excellent example of preparatory data cleaning in their discussion of analysis of the Apache developer network. In particular, they noted difficulties in identifying specific email addresses, as often many users in the Apache developer network have multiple email addresses. Additionally, addresses may be represented in emails with different capitalization, or may be mistyped. Bird et. al. describe a series of approaches to normalize addresses, including applying a case-insensitive filter to addresses, removing aliases from addresses, and clustering similar appearing emails on the basis of the text content [1].

Bird et. al. also provide an excellent body of classifying their network through analysis of degree of each node, divided as indedgree and outdegree, in a degree distribution. They classify these distributions as displaying "power-law" distributions, normal to the network. Similar analyses are performed by Shetty et. al on the Enron email dataset, demonstrating the same "power-law" distribution [7], [1].

With respect to data visualization, Blogspot user Tweinzirl provides an excellent usage of igraph to display the network structure overall, as well as techniques for readily visualizing the clusters in the DNC email network, and was a strong source of inspiration for this project [8]. Shetty provides similar visualizations on the Enron email network, although Shetty applies the technique of coloring important nodes identified by outside context as holding significant importance in the Enron business structure [7]. Rowe et. al. additionally classify the relative hierarchy of the nodes in the network, identifying important nodes and describing their power relationship in a tree structure.

In this paper, a different approach, which is a combination of network analysis and topic model, is used, so it is easy to find the relations between people and their discussions. This approach presents more meaningful results for researchers.

### 2.1 Research Questions

(1) **RQ1:** Does the DNC email network compare to similar other email networks?
(2) **RQ2:** Who are the most important people within the network, and which roles did they hold?
(3) **RQ3:** Can particular structures be identified in the DNC email network, and their structural roles be inferred from their structure?
(4) **RQ4:** Can we classify the content of what the DNC was discussing in the email networks, identifying particular topics?
(5) **RQ5:** How are topics distributed throughout the DNC email network? Are particular topics localized to particular groups?
   (a) Was discussion of working against Bernie Sanders widespread or localized to particular subgroups?

## 3 METHODS

### 3.1 Data Collection

Initially, a DNC co-recipients email network was found from KONECT [4]; however, this network lacked the email body text required to perform topic analysis. To allow for topic analysis and provide more options with respect to network visualization, the full email dataset (n = 22456 emails) was scraped from WikiLeaks [9] using cURL[3]. A substantial limitation to this approach was the time required to scrape the entire email set. An initial approach instructed the script to retrieve emails

as quickly as possible was completed in approximately 6 hours, but failed to validate results to retry invalid HTTP responses. A second approach, using an exponential backoff on requests when a bad response was returned, was able to produce the complete email set, but took approximately 4 days, and frequent monitoring of the process.

Python was used to parse the emails and extract useful information for this project. Email sender, recipient, header, and body information were extracted using two packages - email.parser[5] and email_reply_parser[11], to extract email headers and to filter email bodies for only the top reply in reply chains, respectively. R was used to visualize the network, get the centrality matrices, cluster the network, build predict models, and create topic models.

### 3.2 Network Analysis

For network analysis, one of the most import methods is visualization. The visualizations can clearly show the structure of a network and important information, such as degree, strength, weight, etc., can be easily marked on plots. For these reasons, the visualization is also a very important part of this project. Besides that, centrality matrices are also important for network analysis, such as degree distribution and strength distribution. Visualizations are especially important for their utility in displaying complex mathematic network analyses in an intuitive format.

### 3.3 Text Analysis

Latent Dirichlet Allocation (LDA) [2], a topic modelling technique was used to detect the overall topic and subgroup topics in the DNC email set. LDA is one of the most common algorithms for topic modeling. It uses statistical methods to extract word distributions as topics and topic distributions as documents based on a word and document matrix. It is an unsupervised learning algorithm, so the number of topics k is set by users. Because it is one of the most common to get topics and the results are good enough, it is used in this project. The LDA function in R is very easy to use and get results; however, there were many data preparation challenges in this project.

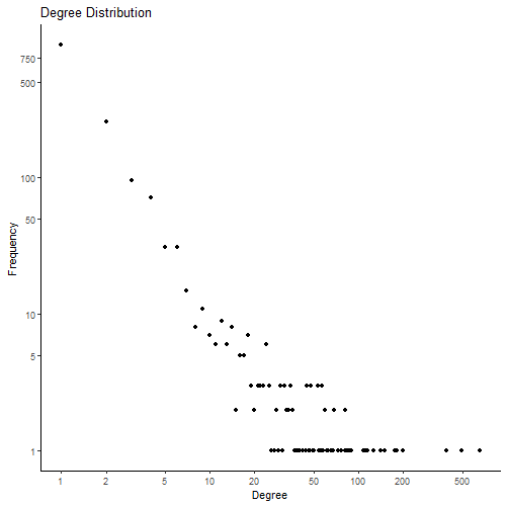
Many emails in the dataset contained full email reply chains in their body. To ensure appropriate analysis of individual documents by LDA, the first challenge was to extract the top email bodies from email files. It was easy to read email bodies using email parser in email package, pre-installed with Python. Then "email_reply_parser"[11] package was used to parse the email bodies and get the top bodies. It did not perfectly work, but it worked in most cases. The email IDs, from addresses, to addresses, time stamps, subjects, and top bodies were read and written into a CSV file that was used for topic modeling.

The second challenge was cleaning data. Because these are emails, the words are not formal, the punctuation includes different formats, and many web links are included. In this case, many special characters were replaced with white-space firstly, such as /, ‘, ’, etc. The reason is that in many cases, they are not common punctuation, so if they are removed directly, some words will be connected. Then common cleaning functions were used, including remove punctuation, remove, numbers, remove stop words, strip white-space, and stem documents. After these steps, many documents were empty, so the empty documents were removed. After a few tests using these steps, the outputs were acceptable.

## 4 RESULTS

### 4.1 Network Characteristics

To characterize the network broadly, distributions of node degree, node strength, and edge weight were calculated.




[H]

Fig. 1. log-log Strength Distribution

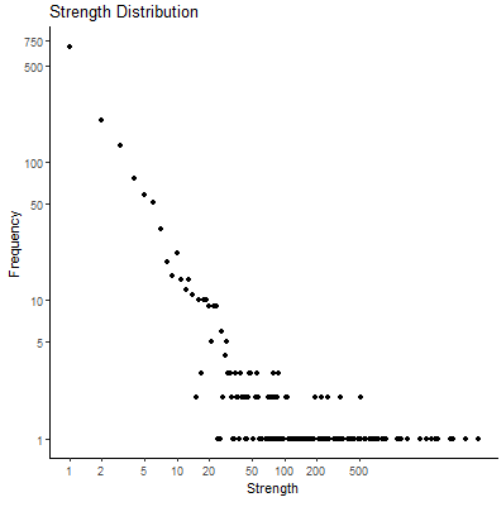



Fig. 2. log-log Strength Distribution

[H]

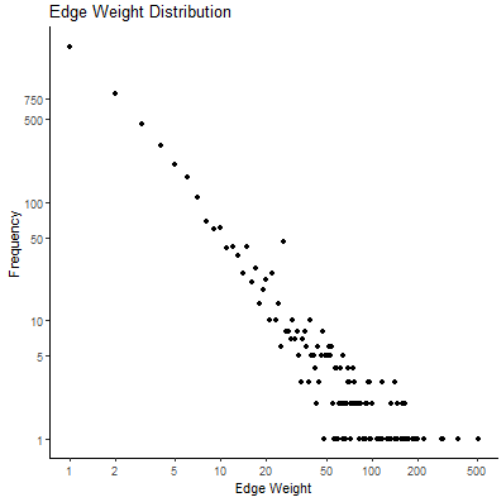



Fig. 3. log-log Edge Weight Distribution

These metrics demonstrate an apparent linear pattern when graphed with log-log scaling. Of minor note are the apparent tails on the high distribution of degree and strength, with a few email addresses in each case apparently above the curve. Outside of these findings, these distributions are comparable to similar distributions found in other email datasets. [1][7]

### 4.2 Network Structure

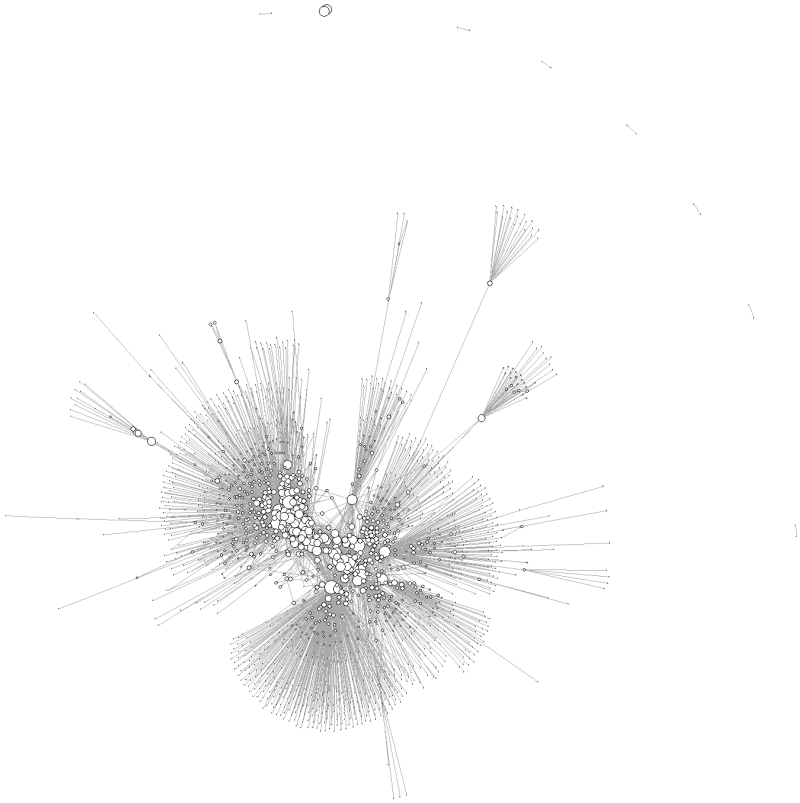


Fig. 4. DNC Email Network

Initial visualization of the social network was plotted using a Fructerman-Reingold layout for points, to position points relative to each other on the basis of the weight of any edges between them (Figure 4). Nodes were scaled logarithmically by strength, such that nodes with high strength appear larger in the visual. Of the 1609 unique email addresses identified, all but 8 pairs were linked in a single central component. Of the 8 outlying pairs, only one was notable for having an edge weight of 505, which was an edge from "noreply@messages.whitehouse.gov" to "dncpress@gmail.com".

A majority of the nodes with high strength appear in a central region of the graph, as would be expected of a Fructerman-Reingold layout. Also notable was the presence of several apparent clusters, displayed distant from the center of the graph. These clusters each demonstrate a selection of small nodes, all of which are connected exclusively or near exclusively to a single central node.

### 4.3 Important Nodes

Three centrality metrics were calculated for each node - degree (number of neighbors), strength (the sum of incident edge weights), and betweenness (number of shortest paths passing through this node). For each metric, the 10 nodes with the greatest score on each metric were selected. This produced a resulting set of 18 nodes, as some nodes had high centrality metrics in multiple categories. Of note, 4 nodes were in the highest 10 of each centrality metric - "mirandal@dnc.org", "comers@dnc.org", "kaplanJ@dnc.org", and "parrishd@dnc.org", who help positions of Communications Director, Finance Chief of Staff, Director of Finance, and Finance Director of Data and Strategic Initiatives, respectively. Additionally, a single node, "comm_d@dnc.org", was notable for having an outdegree of 0, suggesting it only received emails despite its high centrality.

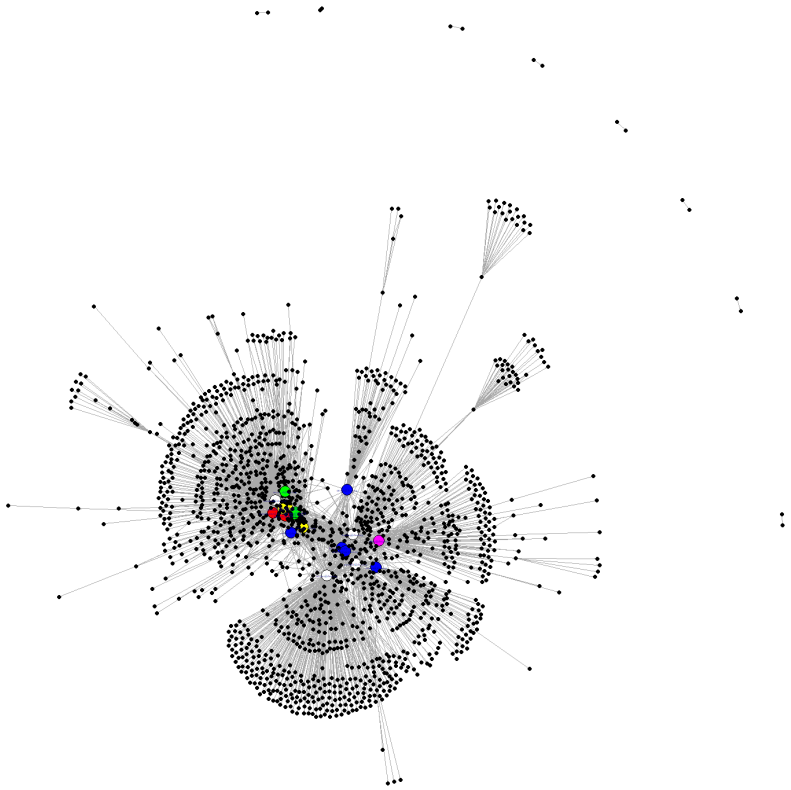


Fig. 5. Important Nodes in DNC Email Network

The graph was visualized to color each of the 18 highest centrality nodes (Figure 5). Each metric was assigned a color: Degree = Red, Strength = Green, Betweenness = Blue. Nodes in the top 10 of multiple centrality categories were colored the RBG combination of each color: Degree + Strength = Yellow, Degree + Betweenness = Magenta, Strength + Betweenness = Teal, and All 3 metrics = White.

### 4.4 Source/Sink Analysis

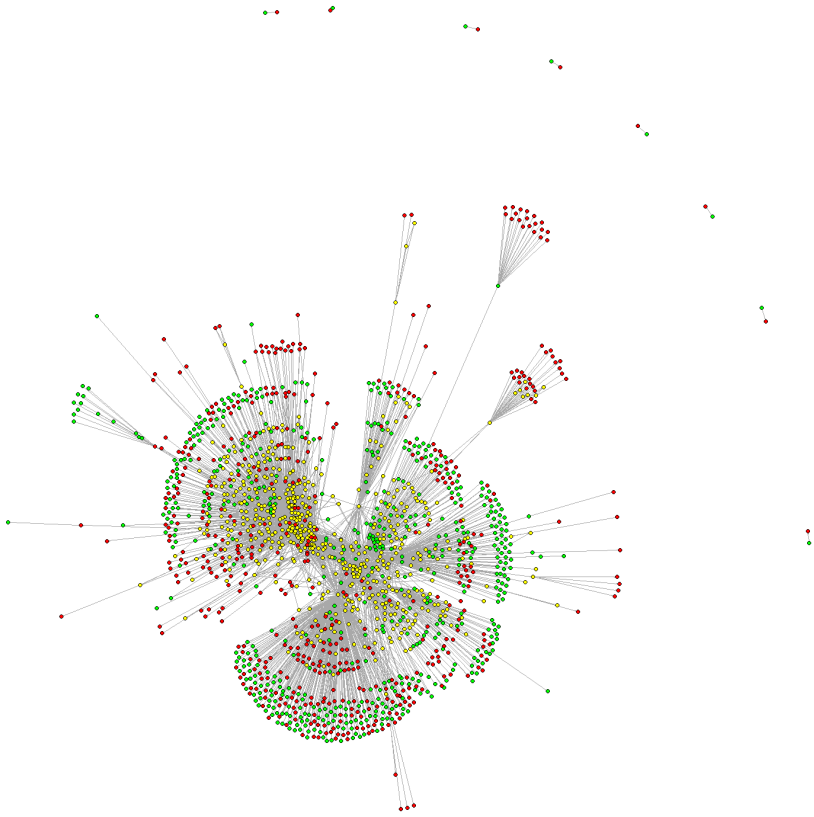


Fig. 6. Sources and Sinks in DNC Network

To identify patterns in data transmission, nodes were categorized on the directed nature of their communications. Nodes which only received emails were categorized as "Sinks", while nodes which only sent emails were categorized as "Sources". When subdivided on these criteria, 509 nodes were identified as Sources, 557 nodes were identified as Sinks, and 543 nodes both received and sent emails. The network was then visualized to display these differences - Sources were colored green, while Sinks were colored red, and nodes who both sent and received emails were colored yellow.

Of note, nodes serving as sources or sinks tend to agglomerate on the edges of the visualization, while the center of the visualization contains the majority of nodes both sending and receiving emails. There also qualitatively appears to be functionally distinct clusters - many of the clusters previously observed in Figure 4 demonstrating a single central node surrounded by a cluster of leaf nodes display the pattern of all leaf nodes being either sinks or sources, dependent on cluster.

### 4.5 Cluster Analysis

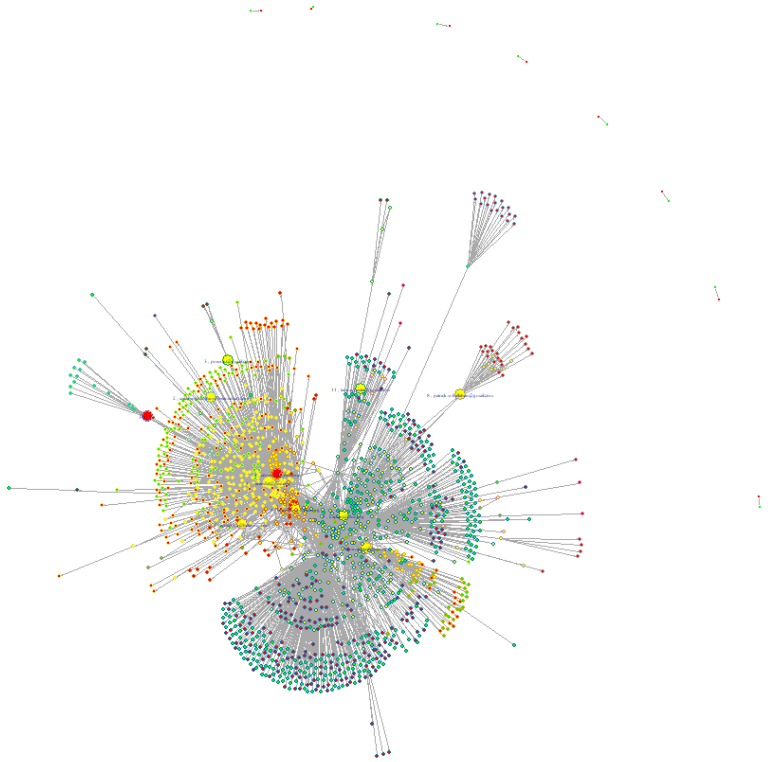


Fig. 7. Clusters in DNC Email Network

Clusters were identified in the network using the Spinglass algorithm, k=12, generating groups with sizes between 9 and 816. Node boundaries were then colored a unique color on the basis of which cluster they belonged to, while node centers were colored on the basis of Sink/Source identification as described in Figure 6.

Cluster identification confirmed the previously qualitative clusters illustrated by the Fructerman-Reingold pattern. Several groups were clearly identified by their patterns of behavior - groups with many Source nodes providing to a single central Sink or mixed node, suggesting an information gathering group, or a group with many Sinks connected to a central source or mixed node, suggesting an information dispersal or executive group.

### 4.6 Text Analysis

For the whole dataset, four topics were identified, and for each of the 12 clusters, two topics were found. The LDA algorithm was used for the whole text mining process.

For the overall topic modeling, after evaluating the LDA algorithm for several number of topics, four topics produced the most distinct topics, so four topics were selected as the final result. Figure 8 presents the four topics, top word distribution, of the overall dataset. Of immediate note is the relatively high gamma for "trump" in topics 2 and 4, which matches a high overall prevalence of the word "trump" in the dataset.

Topic 1 likely represents DNC meta-communication or public relations - specific political candidates are notably absent in the top terms, while words related to communication ("email", "press", "twitter", "communication", "call") place relatively high. Topic 2 likely represents discussion about Republican primaries - it specifically references "kasich", "trump", "cruz", and "primari" while references to democratic candidates are absent. Notable as well is the inclusion of "candid", suggesting accumulation of gaffes and political "mud" on the opposition for campaigning. Topic 3 likely represents the Democratic Primaries, with Republican candidates notably absent from top terms while "bernie", "sanders", "hillary", "clinton", and "primari" are prominent results. Topic 4 likely represents discussion of the general election for the presidency, containing both "donald", "trump", as well as "obama", "clinton", and "presidenci", while lacking references to "primari"


Fig. 8. Overall topic model

For the cluster topic modeling, the topic number was set as two to avoid missing important topics. No significant extra topic was found in any cluster. Two topics of each cluster were kept for future reviewing. The purpose was to detect differences between the clusters, so different roles of the clusters could be found during the comparison.

In cluster one, many news articles, audio clips, and videos were cited. Many of these emails contained text describing potential weaknesses of Trump's campaign - for example, gaffes, hypocrisy, and ties to politically extreme positions and groups. Topic 1 was a particularly large topic, suggesting some combination of high interest in the DNC to accumulate potentially damaging information about their opponent and/or a high amount of damaging behavior on behalf of Trump. Figure 9 shows the top terms of this cluster, and Table 1 shows the exemplar sentences in this cluster.

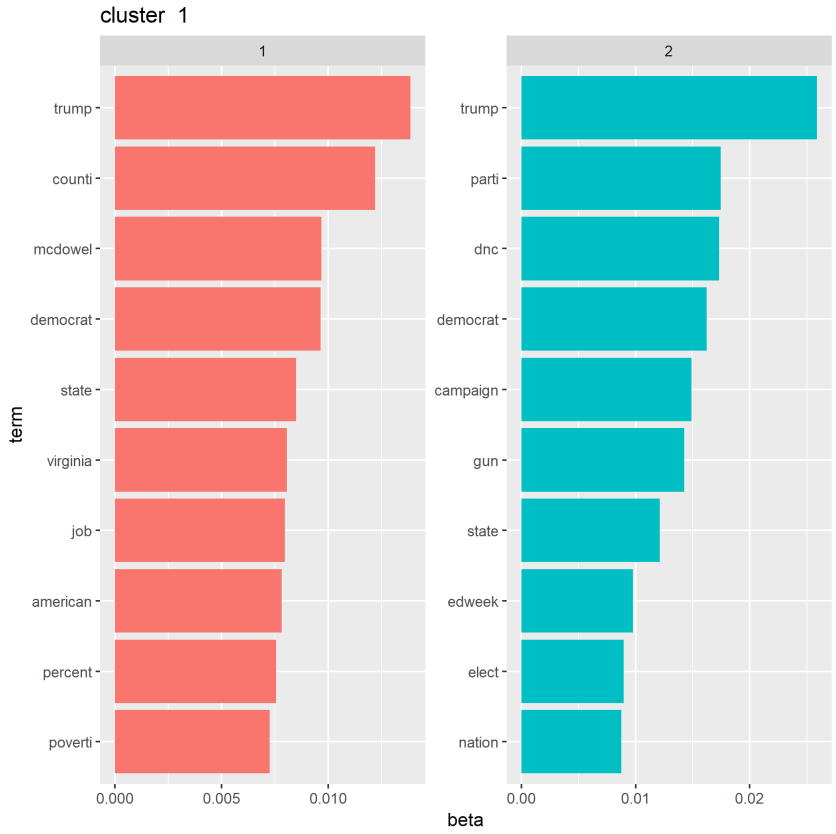



Fig. 9. Cluster 1 topic model

Table 1. Cluster 1 exemplar sentences

| |
|---|
| "One of the counties we focused on at this hearing was McDowell County and I was very pleased that Sabrina was one of the witnesses at this hearing." |
| "The Audio Of Trump Discussing Blacks Vs. Whites." |
| "Trump criticized Clinton for using teleprompters this morning, currently plans to use one himself today." |
| "Randi Weingarten: Donald Trump's Rhetoric Has Contaminated Schools By Andrew Ujifusa." |

In cluster two, the emails are all regarding political news and principal travels. This is a general DNC discussion topic, so this is also a big cluster. Figure 10 and table 2 show the topic of this cluster.

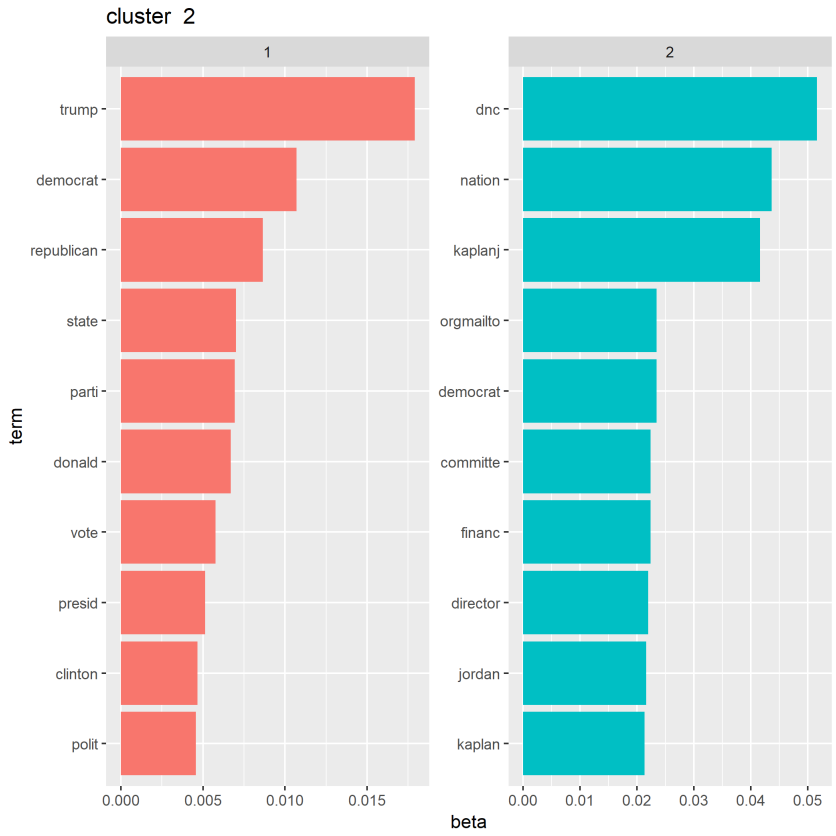



Fig. 10. Cluster 2 topic model

Table 2. Cluster 2 exemplar sentences

| |
|---|
| "U.S. Directs Public Schools to Allow Transgender Access to Restrooms" |
| "Obama arrives in Germany, facing a Europe strained by the migrant crisis and a slow economy" |
| "President Obama Meets With Child of Undocumented Immigrants on Cinco de Mayo" |
| "This email is intended to provide a brief summary of key Political Department priorities, including recent news and principal travel." |

The cluster three is a fairly small cluster. The news discussed in this cluster was about employment information. Figure 11 and table 3 show the topic of this cluster.

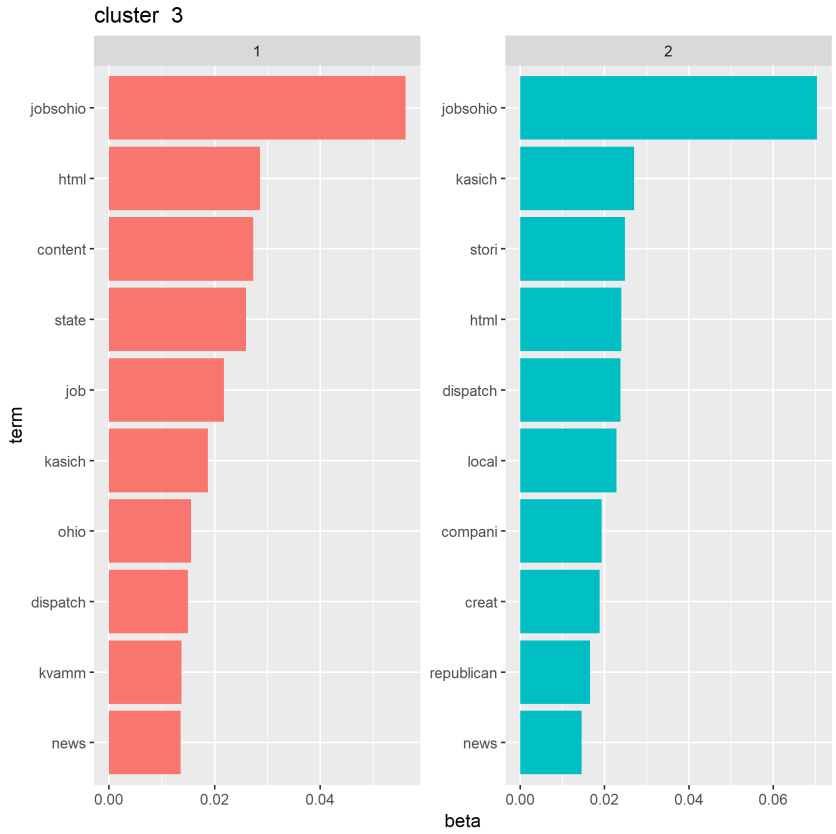



Fig. 11. Cluster 3 topic model

Table 3. Cluster 3 exemplar sentences

| |
|---|
| "FYI – This afternoon, John Kasich will make an announcement that his JobsOhio agency has struck a deal with the CEO of a Thai chemical company." |
| "Great move. Hopefully Henry says the same." |
| "Here is the updated briefing (from Zach rather than from me). Martin gave a little over 25k last year, 20 of which was for Hamilton." |

In cluster four, many support data, numbers, and achievements were mentioned, and the structure of speeches was discussed, so the topic was likely regarding election strategies. Figure 12 presents the top terms, and Table 4 shows the exemplar sentences in this cluster.

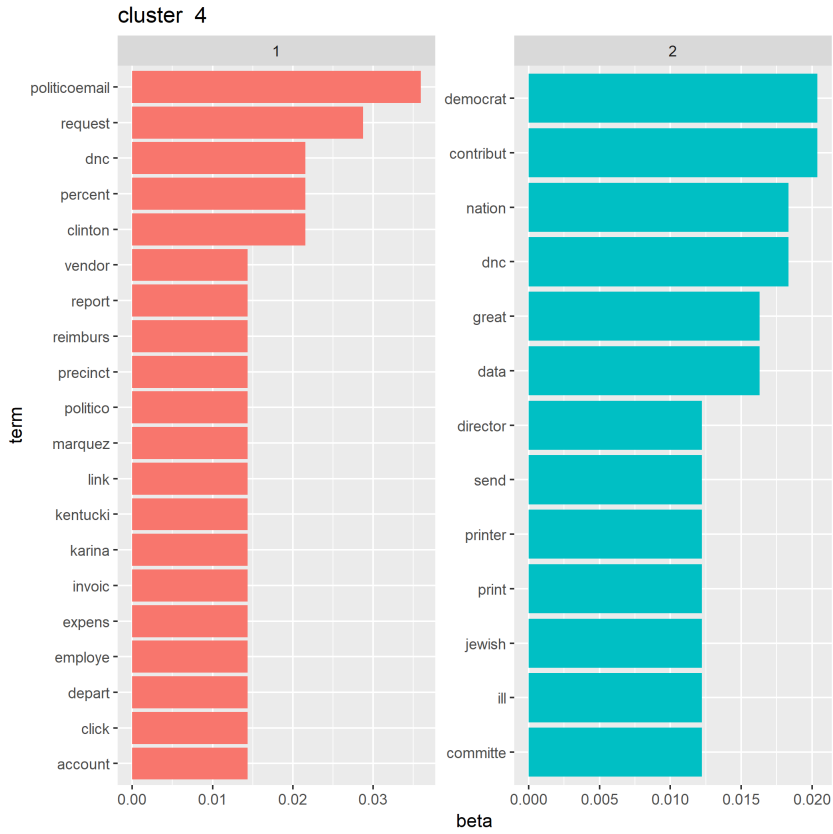



Fig. 12. Cluster 4 topic model

Table 4. Cluster 4 exemplar sentences

| |
|---|
| "Hillary Clinton claimed victory in the Kentucky Democratic primary on Tuesday night, though The Associated Press said the race was too close to call with 99 percent of precincts reporting." |
| "The following request was submitted requiring your department approval." |
| "Do we have any helpful turnout numbers yet?" |
| "Can we pls get this Q & A on DNC clips?" |

The cluster five emails contain a few in-house schedules and some other news, and this cluster is also small. Figure 13 and Table 5 present the topic.

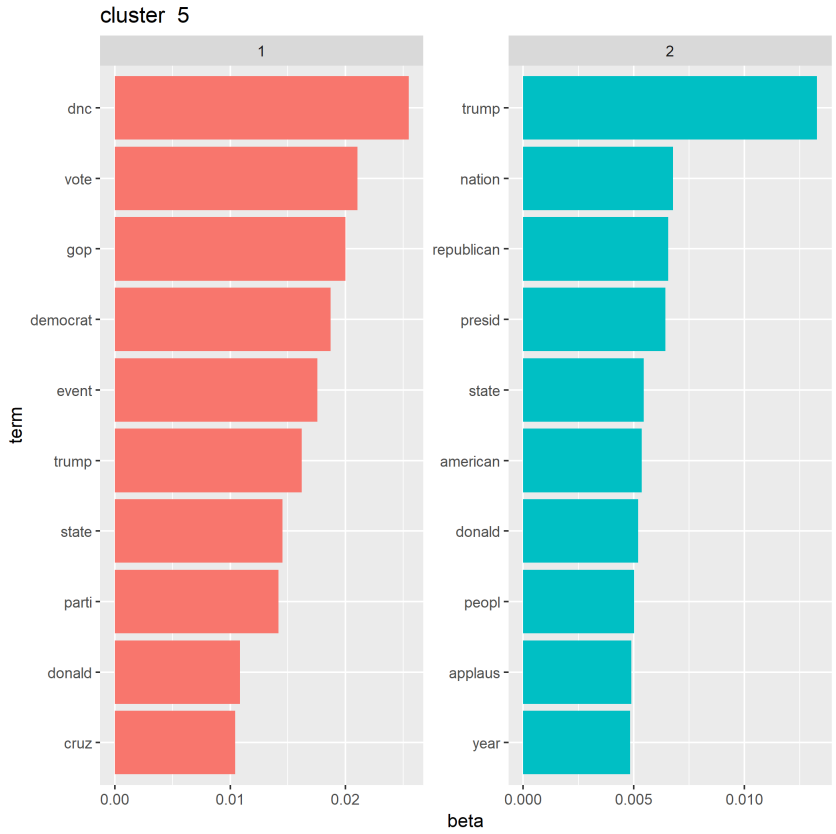



Fig. 13. Cluster 5 topic model

Table 5. Cluster 5 exemplar sentences

| |
|---|
| "Bush the businessman has sometimes lent his name and credibility to money-making ventures that involved dubious characters." |
| "IN-HOUSE SCHEDULE LAST UPDATED 4/25/2016 2:41:06 All times are Eastern Standard Time SIGNIFICANT UPDATES" |
| "REMARKS BY THE PRESIDENT AT CEREMONY HONORING THE RECIPIENTS OF THE NATIONAL MEDAL OF SCIENCE, AND THE NATIONAL MEDAL OF TECHNOLOGY AND INNOVATION" |
| "IN-HOUSE SCHEDULE LAST UPDATED 5/8/2016" |

The other clusters have a similar context, so they are not described in detail. Figure 14-20 and Table 6-12 present the top terms and exemplar sentences.

## 5 DISCUSSION

### 5.1 Reducing Suspicions

Almost any type of information today is available to the public through the internet. An individual may be able to find information on either stock trading, sports statistics, or random survey data. In

the area of political-related data, the idea of credibility is more important. Politicians are always expected to tell the truth and to help the public, but most times that may not be the case. It is a very controversial area, especially with the idea of collusion being accused across both running parties at the time - the Democrats having a leaning bias towards Clinton over Sanders and the Republicans colluding with the Russians to release this data. Either way, through the Mueller investigation, this dataset was said to be leaked by a group of Russian hackers during 2016. This immediately sparks conversation over the idea if this data was manipulated in any way.

Data within the text component of this project may have been easily manipulated to show a potential bias within the findings. Lacking a sophisticated cybersecurity background to identify metadata indicating the emails were fabricated, we compared general network metrics of the DNC email network to other analyses of email networks. Initial findings from plotting the different network characteristics proved that this network is relatively normal - demonstrating normal power-law distributions on important centrality metrics.. Network distributions are expected to follow a negative binomial regression structure, and that is what was found and shown in Figures 1, 2, and 3. From this, we were then able to effectively reduce our suspicion that this is a fabricated network.

Looking at the topic model of the text component as a whole, there are a handful of words that are still shared between the topics. Initially, we thought that the text component may have been altered in some way, but upon further inspection, we noticed that the same words were in different topics as they were being talked about in different contexts. For example, the word "Trump" is listed in topic 2 and topic 4 of the entire dataset. Through further analyzing the other words and sentences in the respective topics, we were able to conclude that topic 2 most closely resembled speaking of the Republican Primaries while topic 4 most closely resembled speaking of the general elections.

### 5.2 Sources and Sinks

Most social networks contain different nodes or people who act as information gatherers (sources) and information collectors (sinks). In some cases, nodes may be able to collect this information and later pass on the information, often connecting at least two different subgroups of nodes. Often, nodes who act as bridges between different subgroups hold the most amount of power in a network. Although, analyzing this network finds that there is not this type of relationship.

The network analysis performed shows that only about a third of the network contained nodes who act as information bridges between other nodes. The majority of the network is comprised of nodes who either collect information from other nodes or forward information to other nodes. In politics today, many politicians have "workers" who are set out to find specific information for that person. That person normally does not forward that information onto another person, but instead may use that information for their own, personal well-being.

### 5.3 Topics

The topic model built across the entire dataset shows that there are different, albeit not completely distinct, topics. Some of the topics contain words that are shared between topics, yet the context of those topics still differ. Through analyzing the topics and searching through top sentences per topic, we have come to the following conclusion for the four overall topics:

- Topic 1: DNC public relations
- Topic 2: Republican primaries
- Topic 3: Democratic primaries
- Topic 4: General elections

We were able to locate and create subgroups of different nodes and then perform topic modeling on these groups. Each cluster has clearly defined goals, shown by the differences in individual topics as well as the top sentences. Here, information aggregation is a top, common theme due to the high presence of articles being passed around. This reinforces the idea of functional groups and focuses on the idea of information distribution rather than two-way discussion of information. Unfortunately, we were not able to pinpoint the exact roles of each of these subgroups. Given our limited expertise in this area, it was too difficult to analyze the specific roles of the different subgroups

There are emails within this dataset that are disapproving of Bernie Sanders, yet it is unlikely that these emails have been communicated across the entire dataset. Through topic modeling, we have shown that the presence of emails criticizing Sanders is not indicative of an organization-wide bias against him. This would indicate that the bias against Sanders was apparent in key central nodes who are most likely to have gathered or dispersed the informaiton in the first place.

## 6 LIMITATIONS/FUTURE WORKS

Our biggest limiting factor was the credibility of this dataset. Through network analysis, a clear social graph is able to be created which gives us reason to believe that this is a true network. Although, the text component to this project was leaked data, which gives us very little information on details of its collection. The text component could have easily been fabricated to contain false or misleading information. Furthermore, there was very little information on how this data was specifically collected. There could have been extra information contained in this dataset that was withheld when it was originally released. All we know was that the emails were leaked and made available to the public through the internet.

In the future, spam emails can be detected and remove from the dataset and see what the differences are after that. Moreover, the roles of the key persons in the email set can be analyzed. The roles of the persons in the DNC can be found, so the differences can be checked between the two roles for the same person.

## 7 CONCLUSION

This network analysis has shown that the emails being sent between people are primarily functional in nature. Much of the communication is the sharing of news articles rather than discussion between individuals, and most individual nodes in the network do not connect with most of the rest of the network. The role of gatekeepers is less powerful in this network as a whole. Any given email, or collection of emails, is more likely to belong to a set of emails being collected by a particular individual, or distributed to a subgroup of specific individuals, than it is to be passed through the organization as a whole. As such, while emails critical of Bernie Sanders are undoubtedly present in the dataset, it is unlikely those emails reached a majority of the network. Ultimately, this project was an exploratory analysis of the DNC emails obtained from WikiLeaks. We have tried to be as non-biased as possible in the identification of topics while also being transparent as to where we obtained our data, the possible suspicions and implications of the data, and how we then performed analysis on this data.

## REFERENCES

[1] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. 2006. Mining Email Social Networks. In *Proceedings of the 2006 International Workshop on Mining Software Repositories (MSR '06)*. ACM, New York, NY, USA, 137–143. https://doi.org/10.1145/1137983.1138016

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] cURL. [n.d.]. cURL. https://curl.haxx.se/

[4] KONECT. [n.d.]. DNC co-recipients. http://konect.cc/networks/dnc-corecipient/
[5] Fedele Mantuano. [n.d.]. emailParser. https://pypi.org/project/mail-parser/
[6] Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore Stolfo. 2007. Automated social hierarchy detection through email network analysis. (2007).
[7] Jitesh Shetty and Jafar Adibi. 2005. Discovering Important Nodes Through Graph Entropy the Case of Enron Email Database. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD '05)*. ACM, New York, NY, USA, 74–81. https://doi.org/10.1145/1134271.1134282
[8] Tweinzirl. [n.d.]. Graph Analysis of Leaked Democratic National Committee Emails. https://loveofdatascience.blogspot.com/2016/09/graph-analysis-of-leaked-democratic.html
[9] WikiLeak. [n.d.]. Search the DNC email database. https://wikileaks.org/dnc-emails/
[10] WikiPedia. [n.d.]. Democratic National Committee. https://en.wikipedia.org/wiki/Democratic_National_Committee
[11] zapier. [n.d.]. emailReplyParser. https://github.com/zapier/email-reply-parser

# A TOPIC MODELING

## A.1 Cluster Top Terms

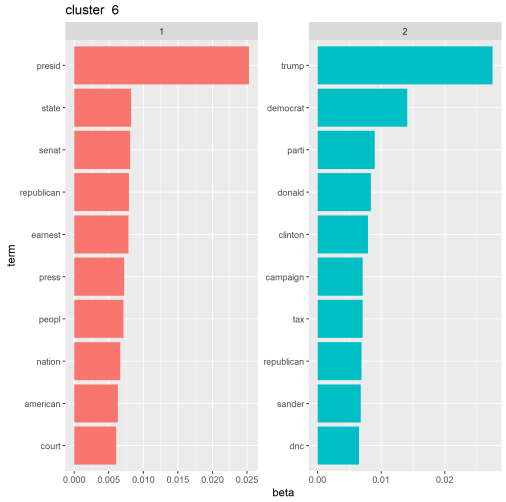



Fig. 14. Cluster 6 topic model

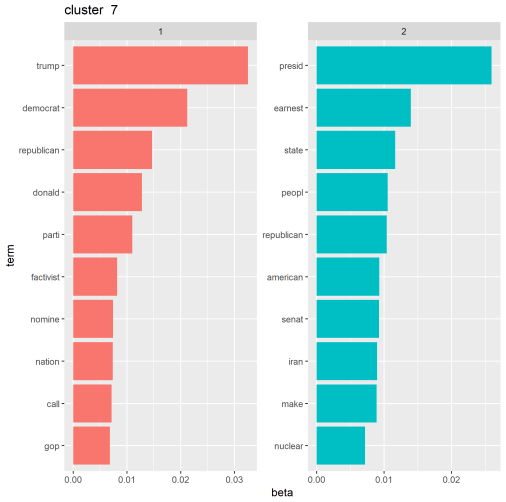



Fig. 15. Cluster 7 topic model

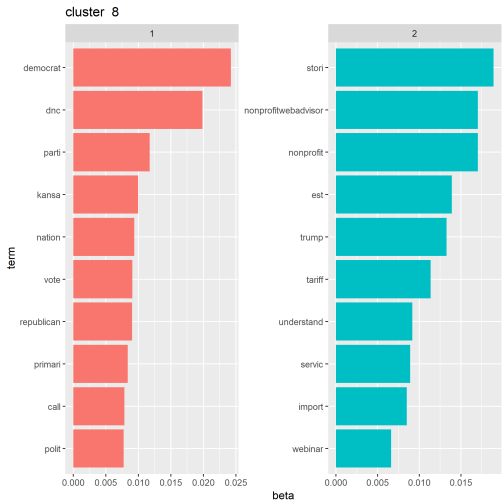



Fig. 16. Cluster 8 topic model

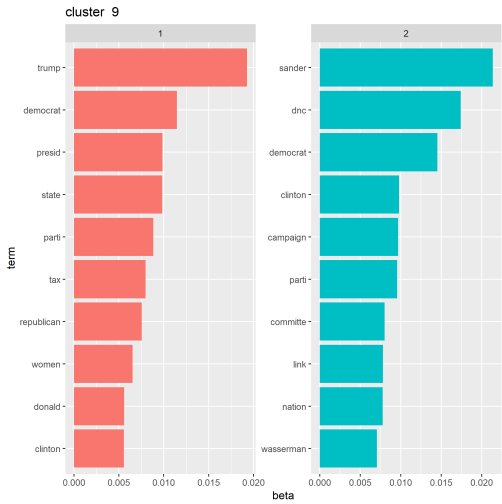



Fig. 17. Cluster 9 topic model

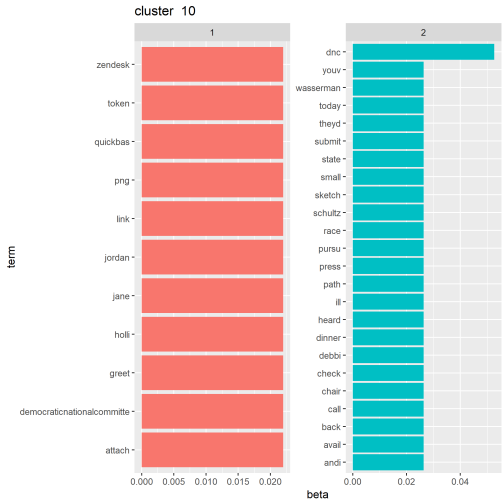



Fig. 18. Cluster 10 topic model

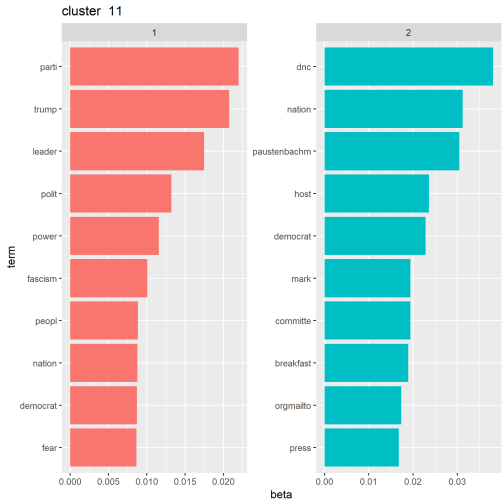



Fig. 19. Cluster 11 topic model

Table 6. Cluster 6 exemplar sentences

| |
|---|
| “THE WHITE HOUSE Office of the Press Secretary For Immediate Release” |
| “Hi everyone, As you may know Clayton, Max and Vaughn have started the Democratic Lawyers Council (DLC).” |
| “Key point: The only window into Trump’s handling of his income taxes came during the 1981 New Jersey report after Trump’s application for a casino license.” |
| “IN CASE YOU MISSED IT Trump’s income tax returns once became public. They showed he didn’t pay a cent.” |

Table 7. Cluster 7 exemplar sentences

| |
|---|
| “THE WHITE HOUSE Office of the Press Secretary For Immediate Release” |
| “It’s been a disastrous week for the GOP. Now that Ted Cruz and John Kasich have both dropped out of the race, Donald Trump is the Republican Party.” |
| “Finance Contributions Status - Yesterday and Today” |
| “Donald Trump’s position on transgender rights is incoherent, not "nuanced"” |

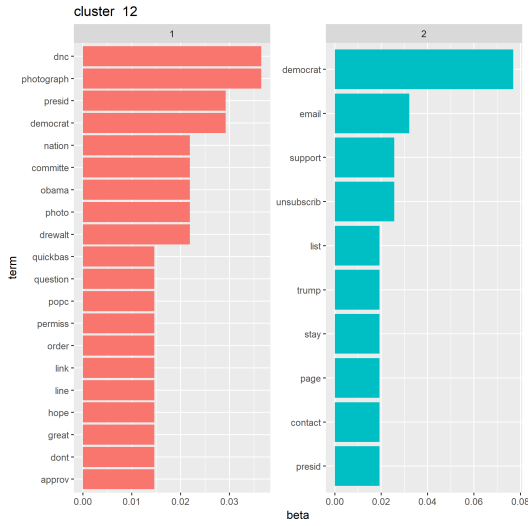



Fig. 20. Cluster 12 topic model

## A.2 Cluster Exemplar Sentences

Table 8. Cluster 8 exemplar sentences

| |
|---|
| “This email is intended to provide a brief summary of key Political Department priorities, including recent news and principal travel.” |
| “Understanding How to Communicate Your Nonprofit’s Story” |
| “Reminder for the New York State Political Briefing to be turned in Friday so you have time to hunt down possible answers and comments after the briefing is reviewed.” |
| “Makes sense, since Trump thinks wages are too high” |

Table 9. Cluster 9 exemplar sentences

| |
|---|
| “THE WHITE HOUSE Office of the Press Secretary” |
| “Key point: The only window into Trump’s handling of his income taxes came during the 1981 New Jersey report after Trump’s application for a casino license.” |
| “In an attempt to head off an ugly conflict at its convention this summer, the Democratic National Committee plans to offer a concession to Sen.” |
| “Trump hired me as a powerful woman. I saw how sexism became his trademark” |

Table 10. Cluster 10 exemplar sentences

| |
|---|
| “DNC Chair Debbie Wasserman Schultz hosts press call on the state of the presidential race” |
| “Strange, try these links” |
| “Dear Scott Comer, Listing of all open invoices requiring approval.” |
| “Thanks, Andy. Let me know when you’ve heard back about which path they’d like to pursue.” |

Table 11. Cluster 11 exemplar sentences

| |
|---|
| “Blastable Key Point: This is how fascism comes to America, not with jackboots and salutes (although there have been salutes, and a whiff of violence) but with a television huckster, a phony billionaire” |
| “2016 CashFlow Report” |
| “Ideas: * Host a Democratic Happy Hour: need a break from the racism and the misogyny?” |
| “Hello Finance, Amy has requested to sit down with each of the departments for a brief discussion of what to expect as we head into the summer.” |

Table 12. Cluster 12 exemplar sentences

| |
|---|
| “Thank you for joining us at the Dinner last week with President Obama.” |
| “Donald Trump is the presumptive Republican nominee for President of the United States.” |
| “Gabe Debenedetti/Politico and Alex Seitz-Wald/NBC both want to attend the ASDC meetings in Philadelphia.” |
| “Nothing really new from the rally.” |