Finding Topics within Communities of the DNC Email Hack of 2016

MAGNUS FYHR*, XINMING WEI*, and TOM KISSANE*, Marquette University

Network analysis and topic modeling is a powerful modern-day tool to uncover group dynamics within a network. The following paper seeks to take the emails from the hacked 2016 Democratic National Convention and identify communities within the sent and recieved emails. Based on these communities we will use Latent Dirichlet Allocation to extract topics from the emails to determine if there are conversations happening that are unique to specific communities.

Additional Key Words and Phrases: network analysis, community detection, natural language processing, topic modeling

ACM Reference Format:
Magnus Fyhr, Xinming Wei, and Tom Kissane. 2019. Finding Topics within Communities of the DNC Email Hack of 2016. 1, 1 (December 2019), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The DNC email hack of 2016 in some ways had a profound effect on how the nation sees their electoral system. The general consensus is that it was the work of Russian hackers who wanted to influence the outcome of the election, although this has never been confirmed. Nonetheless, it is a great networked dataset to look at with an extensive textual component. The study into the emails started with being able to identify important nodes. For example, are there information brokers? Are there certain individuals within the network that carry more influence over others? And how is information disseminated throughout the network?

Eventually, it became clear that that there were certain individuals with more ties to other individuals. Ultimately, the authors were curious if there were defined communities within the network. Furthermore, are there certain topics that come up within these communities that are unique to one community over another? For example, is there one community within the email network that is pro-Bernie Sanders and another pro-Hilary Clinton?

We pre-processed and analyzed the DNC email leak in detail. A network composed of different nodes and edges with specific weights. Through the centralized matrix, betweenness, closeness, centrality and eigenvalue properties. Twenty-seven different communities were detected. Of these twenty-seven, three were chosen for analysis that had appropriate size and email population. For these three local communities, we cleaned and concatenated the body's of the emails to prepare them for LDA. We applied the LDA model to extract three topics, each with ten words. Tracking backwards a bit, the keywords from all the topics were used to find the thirty most relevant emails.

*All three authors contributed equally to this research.

Authors' address: Magnus Fyhr, magnus.fyhr@marquette.edu; Xinming Wei, xinming.wei@marquette.edu; Tom Kissane, thomas.kissane@marquette.edu, Marquette University, P.O. Box 1881, Milwaukee, Wisconsin, 53201-1881.

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

Using these “most relevant” emails, allowed us to better understand the topics uncovered from the LDA. From this we were able to determine if the communities significantly differed in terms of their overarching subject and purpose.

2 LITERATURE REVIEW

In order to do an effective network analysis on our dataset, a solid understanding of how intra-politics work, specifically as it relates to the Democratic National Convention (DNC), needed to be obtained. One paper by Masket et. al., pointed out several nuances to the party that the authors were unaware of. For example, the Democratic Party has typically been seen as less organized that that of the Republican Party in the way that decisions for selecting a candidate are carried out, and suggests that the network structure may be more broad than that of a Republican network [9].

In recent years there has been a widening acceptance of using network analysis within the context of political science. Ward et al. provides a solid introduction into the use of network analysis within the sphere of political science, as well as a rich list of literature to direct our research [13]. One of the first widely lauded applications of network analysis within the context of political science was put forth by Keck and Sikkink in their book, Activists Beyond Borders [7]. Although our study is not about international activism, much of their ideas about being bounded together by common goals, rather than a top-down or bottom-up approach, still very much apply [13].

The movement towards network thinking marks a progression of thought as it relates to understanding intra-party interactions over the years. For years, much emphasis has been placed upon teaching party interactions with what is commonly referred to as the “tripartite model”, first introduced by V.O. Key [8]. The model is known for looking at the party in three separate, but interrelated ways: “party-in-government”, “party-in-electorate”, and “party-as-organization”. Masket et al. argues that with the advent of social network analysis, such a view is incomplete and necessitates a move to look from a “parties-as-networks” perspective [9].

Another cornerstone of political science thought that could prove relevant to this study is the law of “curvilinear disparity” [10]. This law argues that party members that are labeled as sub-leaders tend to be more extreme in their views than those of leaders or non-leaders. This is important to acknowledge in our study because we are working directly with the data produced by the “sub-leaders” of the party, thus we expect the views expressed in the emails to be a bit more extreme, and could potentially show up in our topic modeling.

One of our goals for the project is to be able to effectively identify communities within the email network. The literature in the area of community detection is quite extensive. Probably the most commonly well-known method for community detection is the Girvan-Newman algorithm [6]. For this study, we elected to use a state-the-art algorithm known as the "Louvain Method" [4]. Both methods are known as "network optimization strategies", however the former is based upon local information and generally used for larger networks [5]. Though it was developed with very large networks in mind, it can also be applied to networks of any size [4]. Our choice to use the Louvain Method will be discussed further in the methods section.

The topic modeling represented by LDA (latent Dirichlet allocation) model is a hot research topic in the field of text mining in recent years [3]. The topic model has excellent dimensionality reduction ability, modeling ability for complex system and good expansibility. Text mining by subject modeling can help people understand the hidden semantics behind massive texts and can also be used as input to other text analysis methods to complete text mining tasks including text classification, topic detection, automatic text summarization and relevance judgement. LDA topic model has been widely used in text mining and related fields and has been very successful in traditional web text mining based on news data.

At present, social network is a representative form of social media. Social networks are based on the "The small world problem" and the well-known "six degrees of segmentation theory" [11]. The social network forms the social network organization in the Internet application based on the realization of the network of social relationship and influences the information transmission. Among them, E-mail system is the most basic communication method, which can realize the functions of instant messaging and text sharing.

Information users rely heavily on E-mail systems as one of the main sources of communication. Despite the growth of mobile applications, social networks, and more, their importance and use continue to grow. E-mail is used at both the personal and professional levels. They can be regarded as official documents for communication between users. E-mail data mining can also be analyzed for a variety of purposes, such as subject classification, user clustering, etc. [1]. E-mail systems are very complex text data. The text is mostly fragments, the data sparsity problem is serious; The writing is relatively casual, the grammar is not standard, the network language, symbolic language and new words appear a lot, the data noise is big; E-mail supports users to release instant information through mobile phone and network, which has the characteristics of fast update speed and large document data scale. Therefore, in order to process and analyze the email text, it is necessary to preprocess the text in the email to get pure text, so as to distinguish different communities more accurately.

With the development of the Internet and the rapid spread, facing the network of exploding and random data, text mining work becomes increasingly important. People want to be able to get the accurate information from vast amounts of information in the text [12].So, how to effectively obtain valuable information, how to automatic classification, organization and management the voluminous text data becomes increasingly . Therefore, in the face of these problems and needs, as a research hotspot in the field of natural language processing, automatic text classification technology has been developed rapidly and widely used.

Blei et al. [3] first proposed the LDA (Latent Dirichlet Allocation) topic model, and the traditional LDA model followed the word bag hypothesis, which was a three-layer Bayesian structure. The emergence of this model completed the expansion of the topic model at the Bayesian layer and obtained a wide range of applications. At the same time, the original author and many scholars have improved and expanded LDA model and applied it in different fields. Topic relevance, this paper puts forward the parameter distribution by the authorship of the traditional LDA Model Dirichlet instead of Logistic, the Topic Model (Correlated Topic Model, CTM) [2], in order to solve the challenges in traditional Model of word bag. Thus, it can be seen that modifying parameter distribution is one way to solve the problem, while using word embedding form to solve the problem of topic relevance is another feasible way. In addition, in previous theoretical and empirical studies, the validity and reliability of the LDA model have been fully proved, but the problem of choosing the number of topics in the LDA theme model has not been effectively solved. The selection of the number of topics directly affects the LDA model's interpretation of text data and the effect of topic recognition, so it is necessary to solve this problem.

3 DATA METHOD

3.1 Community Detection

We originally wanted to use the Girvan-Newman algorithm within the Python library NetworkX, but because there was a struggle to get the algorithm working on one of the team members' machine, we ended up using Louvain Method found in the python-louvain library. It was originally chosen for its simplicity to run, but we quickly found that it was simple, very fast, and accurate (see Literature Review).

The Louvain Method is designed to find high modularity partitions of large networks in a short period of time [4]. The first step of the algorithm is assigning each node to its own community. For node *i*, each of its neighbors, *j* are looked at to determine if there is a gain of modularity if *i* were placed in the same community as *j*. Node *i* is then placed within the community where the outcome of the gain is maximum.

The actual process of community creation was very simple and straightforward– we created the network via NetworkX and ran the *best-partition()* function within python-louvain library. This resulted in a partitioning of 27 communities, most of which were very small and inconsequential to the analysis. Of these 27, we selected three communities that we believed to be relevant to our analysis; henceforth Communities A, B, and C.

3.2 Topic Modeling

Latent Dirichlet Allocation (LDA) was proposed by [3] to predict the topic distribution of documents. LDA is a collection of discrete data, such as a text corpus, that generates a probabilistic model. LDA is a three-layer hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Simply put, it will document the theme of the concentration of each document in the form of probability distribution, and through the analysis of some documents to extract their subject distribution, can according to the subject of clustering or text classification.

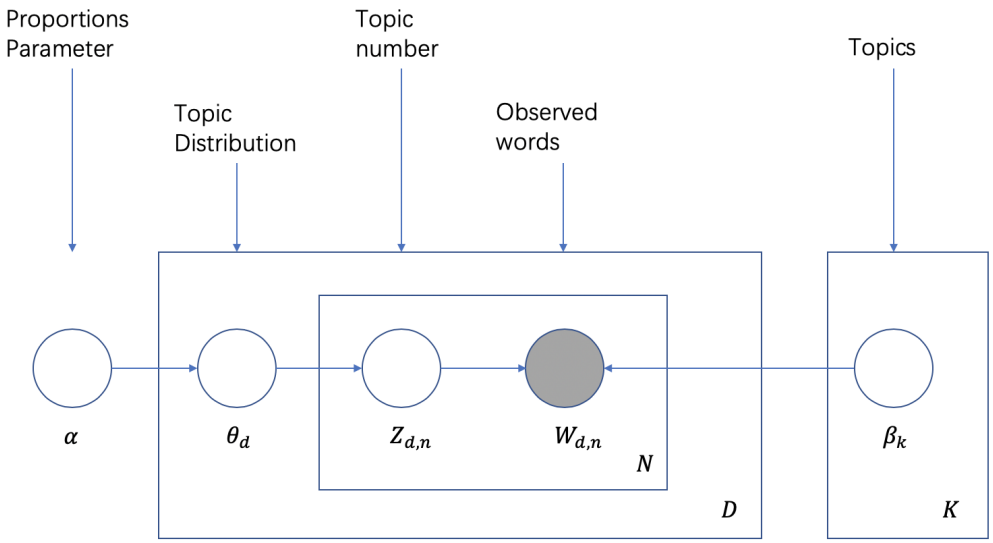

Fig. 1. Graphical model of LDA

LDA adopts the word bag model. The so-called word bag model refers to a document in which we only consider whether or not a word appears, regardless of the order in which it appears. Dirichlet distribution is the conjugate prior probability distribution of polynomial distribution. Therefore, as

described in the LDA Bayesian network structure, a document is generated in the LDA model as follows:

We have **D** corpus, so we have n words in the w document. Our goal is to find the topic distribution for each document and the word distribution for each topic. In the LDA model, we need to assume a number of topics **K**, that all are based on the distribution of **K** topics. LDA is assuming that the prior distribution of the document topic Dirichlet distribution, in any corpus **D**, its topic distribution θ_d as follows:

$$\theta \sim Dir\,(\alpha)$$

Parameter α for distribution to them, that is a **K** dimensional vector. For each of the **N** words w_n, the topic z_n in the topic distribution θ as follows:

$$z_n \sim Multinomial\,(\theta)$$

The Probability distribution of the word w_n on the topic z_n as follows:

$$w_n \sim Multinomial\,(\beta \times z_n)$$

where the word probabilities are parameterized by a $\mathbf{K} \times \mathbf{V}$ matrix β.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of **N** topics z, and a set of **N** words w is given by:

$$p\,(\theta, \mathbf{z}, \mathbf{w}, |\alpha, \beta) = p\,(\theta|\alpha) \prod_{n=1}^{\mathbf{N}} p\,(z_n|\theta)\, p\,(w_n|z_n, \beta) \tag{1}$$

Finally, the edge probability of a single document is multiplied to obtain the probability distribution of a corpus:

$$p\,(\mathcal{D}|\alpha, \beta) = \prod_{d=1}^{\mathbf{M}} \int p\,(\theta_d|\alpha)\, (\prod_{n=1}^{\mathbf{N_d}} \sum_{z_{dn}} p\,(z_{dn}|\theta_d)\, p\,(w_{dn}|z_{dn}, \beta))\, d\theta_d \tag{2}$$

Based on the maximum likelihood estimation of $p(w_i|\alpha, \beta)$, we want to solve the distribution of words in each topic based on this LDA model. Generally, there are two methods: the first method is based on Gibbs sampling algorithm, and the other method is based on EM algorithm to estimate the parameters in the model.

3.3 Data Collection

The data for this project was scraped Wikileaks using python's 'request' and 'beautiful soup' packages, where each email could be pulled and parsed via changing the the integer value at the end of the URL; "https://wikileaks.org/dnc-emails/emailid/EMAIL-ID" where "EMAIL-ID" represents the id of the email to be requested. Using a loop all 44000+ emails could be pulled, parsed and formatted into a csv file. In this file, for each email there was an id, sender, recipients, subject, and content. This allowed for future processes to run smoothly, such as the generation of nodes and edges csv files. Both nodes and edges csv files were generated using a python script to parse the original data set. The nodes csv contains the email address of every address involved in the DNC email leak. Each row represents a different address that contains user-id, emails-sent, emails-received, and community-id. The edges csv file contains all unique communications that occurred between two any email addresses along with a weight which represents how many times this communications occurred; in this context an edge is made anytime two emails are involved in the same email, it does not matter whether it is a sender and a receiver or if they both receive the same emails they are all treated equally.

3.4 Potential Biases

We realize that there are always going to be bias in what we are trying to achieve when we do a network analysis. One of our original desires was to see if we could identify, so we obviously bring this to the table when reading and interpreting the text. Its very possible that the there might be something more nuanced within the text, but we missed it because our focus has been on identifying various factions within the party.

As stated previously, we ran the Louvain Method algorithm to give us clusters of nodes. From these clusters, we kept the top 3 clusters by size as we felt they would provide the most value to our analysis.This, of course, means that we eliminated 23 other clusters. Most of these were very small; however, it could have skewed our results to eliminate these nodes from our analysis.

LDA also tends to be biased towards words with higher frequency as the words with higher frequency usually have a higher probability of being within a given document, thus showing up more frequently in the result. It is reasonable to say that an important word that could change the determination of a topic could be eliminated from the result list simply because it was a less frequent word compared to others.

4 RESULTS

4.1 Full Network

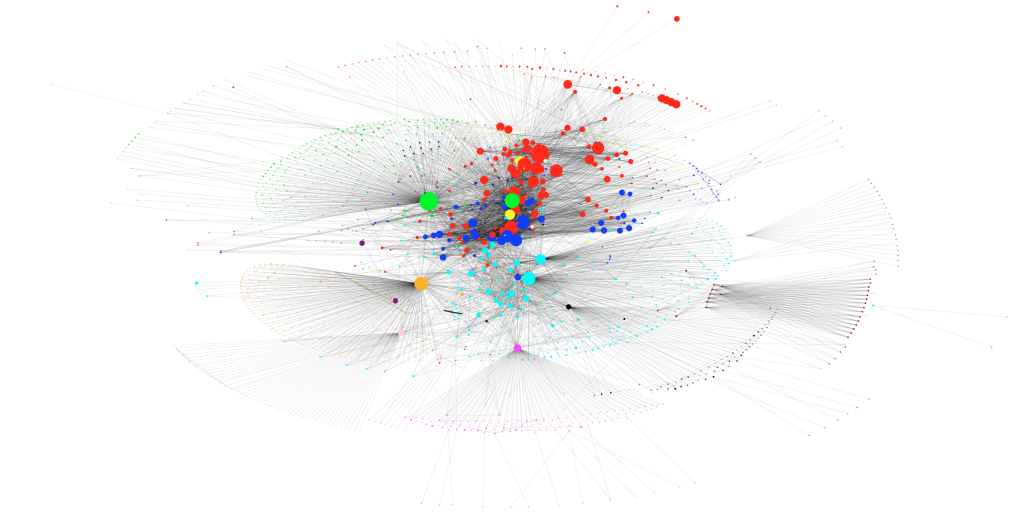

Fig. 2. Network Graph Of Full Dataset; Colorized By Community

4.2 Community A

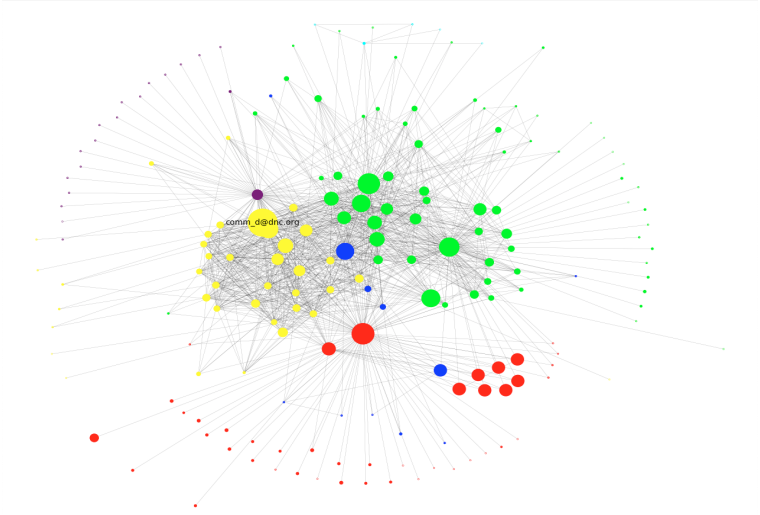

Fig. 3. Network Graph Of Community 0; Colorized By Sub-Community

Metrics	Degree	Eigenvector	Closeness	Betweenness
Results	0.06985	0.04671	0.44806	0.00699

Table 1. Community A Centrality Metrics

4.2.1 LDA.

Topic	1	2	3
Words	['"trump" ', '"republican" ', '"would" ', '"president" ', '"state" ', '"donald" ', '"democrat" ', '"american" ', ' "voter" ', '"party"']	['"clinton" ', '"trump" ', '"state" ', '"campaign" ', '"party" ', '"democratic" ', '"cruz" ', '"hillary" ', '"donald" ', '"sander"']	['"trump" ', '"said" ', '"would" ', '"people" ', '"donald" ', '"going" ', '"think" ', '"like" ', '"know" ', '"want"']
Interpretations	Trump As Pres. Party	Trump Interview	Trump Giving What people Want

Table 2. Topic Modelling Of Community A

4.3 Community B

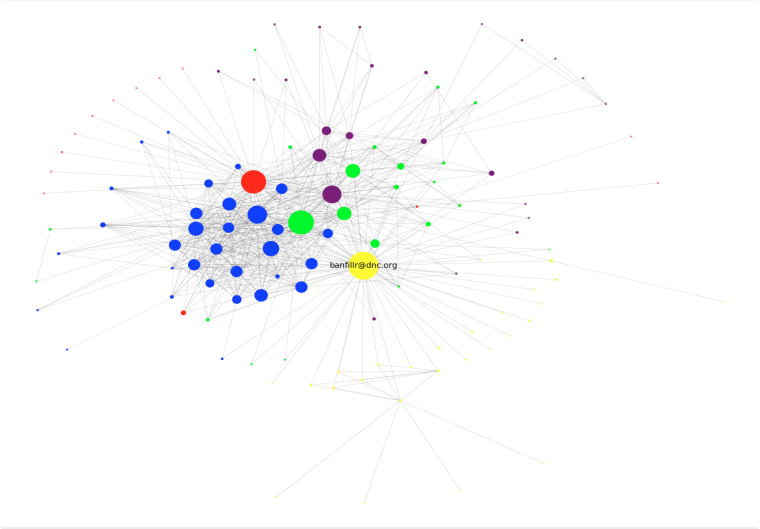

Fig. 4. Network Graph Of Community 2; Colorized By Sub-Community

Metrics	Degree	Eigenvector	Closeness	Betweenness
Results	0.09358	0.05993	0.44986	0.01110

Table 3. Community B Centrality Metrics

4.3.1 LDA.

Topic	1	2	3
Words	['"drive" ', '"call" ', '"liana" ', '"office" ', '"chair" ', '"meeting" ', '"committee" ', '"democratic" ', '"national" ', '"time"']	['"drive" ', '"time" ', '"call" ', '"minute" ', '"courtney" ', '"meeting" ', '"liana" ', '"angeles" ', '"democratic" ', '"comms"']	['"party" ', '"democratic" ', '"schultz" ', '"state" ', '"trump" ', '"wasserman" ', '"chair" ', '"convention" ', '"said" ', '"committee"']
Interpretations	Schedules, and upcoming meetings		

Table 4. Topic Modelling Of Community B

4.4 Community C

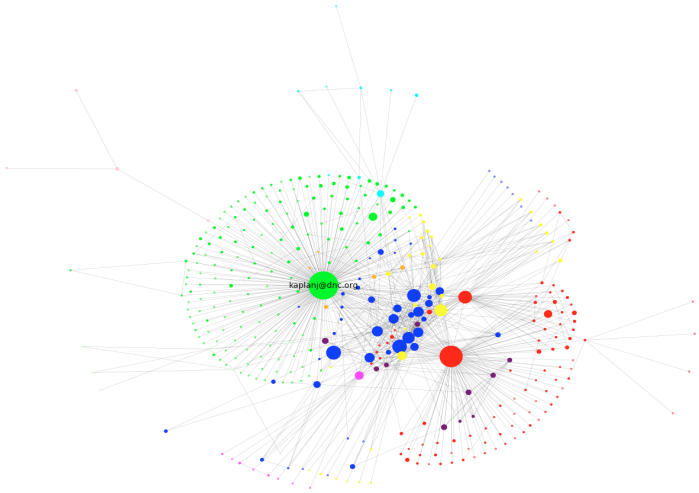

Fig. 5. Network Graph Of Community 5; Colorized By Sub-Community

Metrics	Degree	Eigenvector	Closeness	Betweenness
Results	0.01135	0.03120	0.39961	0.00407

Table 5. Community C Centrality Metrics

4.4.1 LDA.

Topic	1	2	3
Words	['"republican" ', '"trump" ', '"campaign" ', '"court" ', '"said" ', ' "state" ', '"president" ', '"year" ', '"would" ', '"senate"']	['"user" ', '"deleted" ', '"washington" ', '"said" ', '"york" ', '"please" ', '"trump" ', '"state" ', '"campaign" ', '"attachment"']	['"email" ', '"would" ', '"know" ', '"best" ', '"please" ', '"time" ', '"service" ', '"work" ', '"event" ', '"office"']
Interpretations	Trump Campaign	Trump delegation New York	Scheduling Events

Table 6. Topic Modelling Of Community C

5 DISCUSSION

We want to explore the existence of different communities through the analysis of DNC emails. To understand what people are talking about in different communities, we need to preprocess the text in an email. Eliminate time, address, symbols, emojis, strings, informal expressions, etc. We found that this was tough to do because there are so many nuances and abbreviations contained within email text. Often times, we filtered out one thing only to realize the side effect it had on another important bit of information contained within the emails.

We were hoping to be able to clearly identify divisions within the DNC by identifying communities and see what they were discussing and possibly derive some different factions of the party. We did this By extracting keywords from different topics within each community, analyze how they are working on the topic and get the most accurate description of the topic. We also gathered a centrality table, obtaining the four main metrics including betweenness, closeness, centrality and eigenvalue. We determined the different community and its nodes. The characteristic structure of each community is also analyzed to determine the topics discussed in the community.

The LDA model provides a good solution to the topic model. Through the hierarchical Bayesian model, the topic distribution in the text and the probability distribution of keywords in the topic are extracted. Through these keywords, we give our understanding and summary of the topics in different communities. We analyzed the topic models of the global network and two local communities, and identified the different topics of e-mail discussion.

6 LIMITATIONS AND FUTURE WORK

Since we built the dataset directly from our source, there were many limitations removed from our research. However, not being able to identify spam or repetitive emails, such as daily updates, donation updates or schedules was definitely the largest limitation to our research question. These emails 'distracted' or topic modeling as their sheer frequency, repetitiveness and length allowed the words within these emails to dominate the probability spreads of words within a community. This made it difficult to narrow down the content matter of emails that occurred between two or multiple persons. The topics of these emails would have more consistently reflected that of the community rather than a specific subset that dominated the proportion purely based on repetition, identical phraseology, etc. This leads into many of our ideas for future improvement. Given more time there are many more steps of 'cleaning' we would have taken to remove these "daily update" emails that corrupted our communities and topic modeling. We believe focusing on person-to-person conversations would have been better at uncovering the "under-the-table" discussions that crippled the DNC and exposed the bias and collusion that manifested against Bernie Sanders leading to the resign of DNC chairman Debbie Wasserman Schultz.

7 CONCLUSION

We used DNC emails leak in 2016 as data for social network analysis. Through extracting and analyzing betweenness, closeness, centrality and eigenvalue four feature matrix, we detected 27 different communities. According to the different structure and characteristics of the community, we selected six communities with typical structure. Through the analysis of three of the most representative communities, the text of email was purified eliminating the text noise. We applied the LDA model to analyze the subject of the email text. The top three topics and the top ten keywords in each community were extracted. And get our summary and analysis of the topics. As the results show, although there were communities within the DNC, their emails did not indicate strong differences from one another. One of our primary research questions was to determine there was a group that supported Bernie Sanders and another that supported Hilary Clinton. We were

ultimately unable to draw this conclusion. Interestingly, our LDA revealed much more talk in the emails about Trump than we expected.

REFERENCES

[1] Izzat Alsmadi and Ikdam Alhami. 2015. Clustering and Classification of Email Contents. *Journal of King Saud University-Computer and Information Sciences* 27, 1 (2015), 11.

[2] David Blei and John Lafferty. 2006. Correlated Topic Models. *Advances in Neural Information Processing Systems* 18 (2006), 7.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (Jan 2003), 29.

[4] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (October 2008), 7.

[5] Generalized Lovain Method for Community Detection in Large Networks. 2011. Networking the Parties : A Comparative Study of Democratic and Republican National Convention Delegates. In *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, IEEE, Cordoba, Spain, 88–93.

[6] Michelle Girvan and Mark EJ Newman. 2002. Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 6.

[7] Margaret E Keck and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press, Ithaca, NY.

[8] Valdimer Orlando Key. 1952. *Politics, Parties, and Pressure Groups*. Thomas Y. Crowell Company, New York, NY.

[9] Seth E Masket, Michael T. Heaney, Joanne M Miller, and Dara Z Strolovitch. 2009. Networking the Parties : A Comparative Study of Democratic and Republican National Convention Delegates. In *APSA 2009 Toronto Meeting Paper*. Toronto, Ontario, Canada.

[10] John D. May. 1973. Opinion Structure of Political Parties: The Special Law of Curvilinear Disparity. *Political Studies* 21, 2 (1973), 16.

[11] Stanley Milgram. 1967. The Small World Problem. *Psychology Today* 2, 1 (1967), 7.

[12] Tingting Wang, Man Han, and Yu Wang. 2018. Optimixing LDA Model with Various Topic Numbers: Case Study of Scientific Literature. *Data Analysis and Knowledge Discovery* 2, 1 (2018), 11.

[13] Michael D Ward, Katherine Stoval, and Audry Sacks. 2011. Network Analysis and Political Science. *Annual Review of Political Science* 14 (2011), 19.