

Exploring Hillary Clinton’s Emails*

Leland Bybee, Roger Fan, Ryan Vaughn

April 22, 2016

1 Introduction

From roughly 2008 to 2013, Hillary Clinton and some of her staff used a private family email server for most of their communication. This time period includes her tenure as U.S. Secretary of State, and has therefore become a controversial topic due to the possible security concerns of not using official government servers. Various investigations into the legality and appropriateness of the email server have been launched, and the issue has had and will potentially have a significant effect on Clinton’s ongoing presidential campaign.

Due to the Freedom of Information Act (FOIA), the State Department has released the vast majority of the roughly 55,000 pages of emails to the public, leading to an unprecedented opportunity to explore the email usage of a major public figure.

Due to the size of the corpus, reading individual emails in the hope of discovering interesting characteristics is unrealistic. We therefore exploit statistical methods to break this vast analysis problem to a more manageable one, decreasing the scale so that we can more easily apply human and statistical analysis.

In order to tackle this problem, we use classical multi-dimensional scaling and spectral clustering in order to visualize and explore the corpus. We also use a statistical topic model, latent Dirichlet allocation (LDA), in order to identify and analyze the meanings and semantics. Finally, we use the LDA results and other metadata to examine how other covariates, such as email sender and time period sent, are related to the topic of an email.

*STATS 601 final project. All of the code used to download, process, and analyze the data for this project is available at https://github.com/lbybee/601_Final_Project and has been open-sourced under the MIT license.

2 Data

The State Department released the raw data in the form of roughly 30,000 pdfs of individual emails from the Clinton server.¹ We first process the pdfs to extract the raw text.² From the raw email text, email headers and forward/response information are stripped, leaving only the subject line and body text.

All of our methods use a bag-of-words model, so from this raw data we construct a document term matrix containing vectors of word counts. Some standard data cleaning for text data is also performed, including the removal of punctuation and email addresses, word stemming (combining words with common roots, e.g. calling and called), and removing extremely common words (i.e. stop words) and extremely rare words (words that appear in less than 0.1% of emails).

After this processing, we are left with a dataset of roughly 28,000 emails and a total vocabulary of over 3,000 words. Note that each word count vector is extremely sparse, so we need statistical methods that can apply in a sparse coordinate setting. Also, methods that can efficiently exploit this sparseness to reduce computation are preferred.

3 Visualization

Due to the large size and high dimensionality of the data, we must apply some dimensionality reduction in order to effectively visualize it. We therefore apply classical multi-dimensional scaling, which attempts to place each point in a much smaller dimension such that pairwise distances are preserved.

However, we expect standard Euclidean distance to poorly capture the actual relationships between vectors of word counts. In particular, Euclidean distance is affected by the total number of words, two emails with the exact same word distribution that differ in length could have a very large distance between them. In order to better capture pairwise relationships, we will use a distance metric based on the angles between vectors.

We define the cosine similarity between word frequency vectors V and W to be the cosine of the angle between the two vectors, calculated as

$$\cos(\phi) = \frac{\langle V, W \rangle}{\|V\| \|W\|} \quad (1)$$

¹Thanks to Ben Hamner and the Wall Street Journal for making the data processing code (<https://github.com/benhamner/hillary-clinton-emails>) and email pdfs (<http://graphics.wsj.com/hillary-clinton-email-documents/>) available, respectively.

²We used `pdftotext`, an open-source program used to extract text from pdf files.

Since counts must be nonnegative, note that $\phi \leq \pi/2$. This implies that the cosine similarity is always between 0 and 1, with 0 corresponding to vectors that are multiples of each other and 1 corresponding to orthogonal vectors (i.e. no shared words). So we can define the cosine distance as

$$d(V, W) = 1 - \cos(\phi) = 1 - \frac{\langle V, W \rangle}{\|V\| \|W\|} \quad (2)$$

Using this measure of distance, we can construct a distance matrix D and apply MDS. The results for an 8,000 email subset are shown in Figure 1, plotted in both two and three dimensions. We can see several interesting patterns. The vast majority of emails (roughly 5,000) are contained in the main central grouping, with two large groupings that extend away. Besides these two “arms,” however, it is difficult to visually identify other significant structures.

3.1 Spectral Clustering

In order to better identify structure that may exist in the data, we attempt to apply clustering methods. However, it is important to choose an appropriate clustering method due to the high dimensionality and sparse nature of the data. In particular, we expect centroid-based methods such as Gaussian mixture models and k -means clustering to perform poorly in this setting.

Spectral clustering, however, exploits the information contained in the pairwise distances (or, equivalently, the similarities) between points. By using pairwise distances, spectral clustering is particularly well-suited to finding clusters that may not necessarily be spherical or ellipsoidal and that may lie along lower-dimensional manifolds.

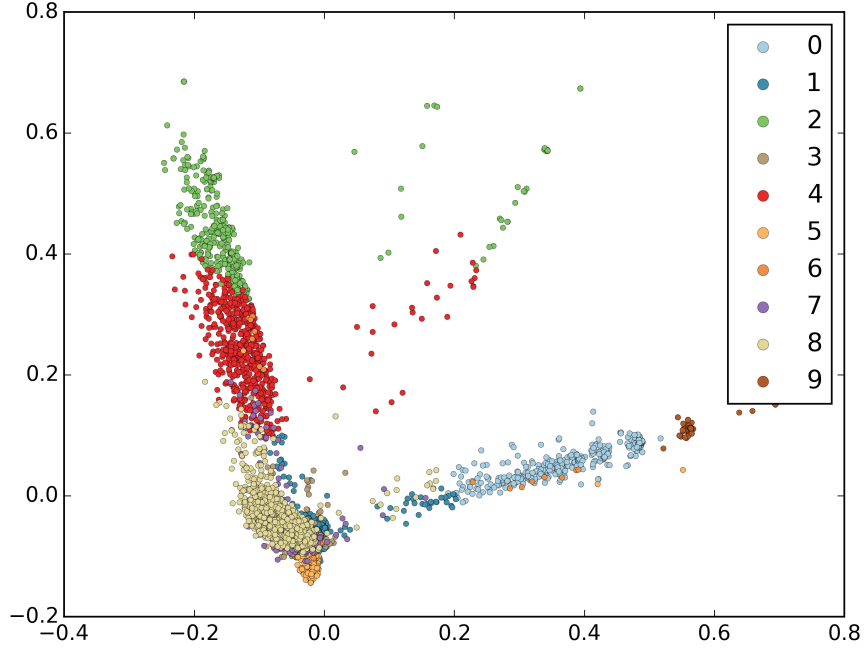
To actually apply spectral clustering, we first construct the pairwise similarity matrix S using our cosine similarity as defined in Equation 1. We define the degree matrix D to be a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and construct the symmetrix normalized Laplacian matrix as

$$L = I - D^{-1/2} S D^{-1/2} \quad (3)$$

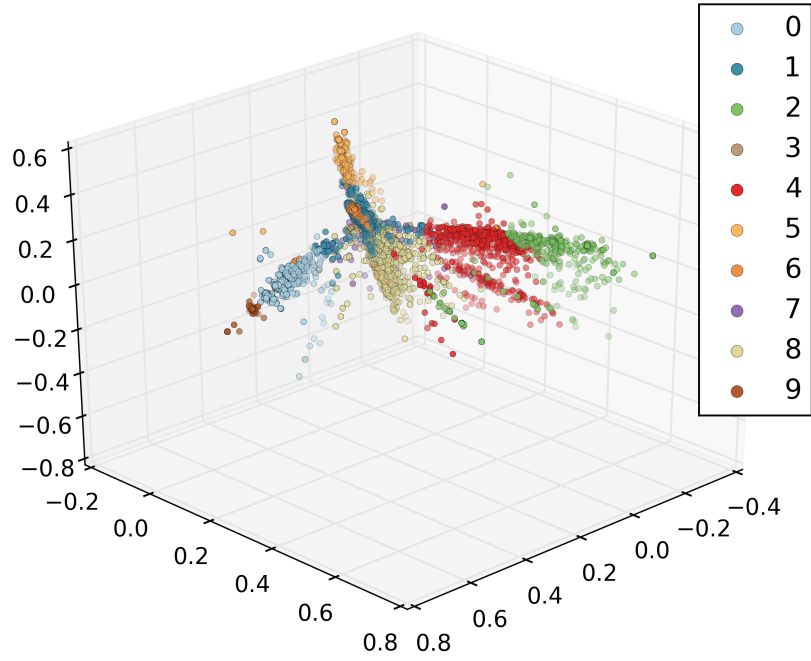
Then the eigenvectors corresponding to the smallest k eigenvalues of L are computed and k -means clustering is performed on these eigenvectors.

Figure 1 classifies each point according to the results of spectral clustering with k clusters. We can see that many of the cluster correspond to the visually identifiable structures, including clusters that correspond to each of the “arms.”

By investigating the individual emails, we can also get a sense of what each cluster consists of. For instance, Clusters 2 and 4 consist of emails related to calls and communication. In



(a) 2 Dimensions



(b) 3 Dimensions

Figure 1: Visualization of a 8,000 email subset using classical multi-dimensional scaling and with coloring according to the results from spectral clustering. The similarity and distance metrics used for spectral clustering and MDS, respectively, are cosine similarity and distance as defined in Equations 1 and 2.

this arm, those closer to the main central clump contain more conversation, while those farther out are much briefer and directly about the call (e.g. “I’d like to call tomorrow.”). Cluster 3 consists entirely of “minis,” summaries of the day’s schedule sent to Clinton each morning. Emails from Cluster 5 are generally requests to print materials, while those from Cluster 6 are forwarded news articles.

In general, it seems that spectral clustering does a good job at identifying emails by their structure, but does not tell us much about the meaning of an email. We can tell that an email is regarding a call or is a news headline by what cluster it is in, but not what that call or headline is about. On a related note, it does not do a good job separating the “regular” emails. Over 5,000 of the 8,000 emails depicted fall into Clusters 1 or 8, which essentially consist of all the emails without specific identifiable structure. Ideally we would be able to further classify these emails according to subject matter, but clustering does not seem suited to this task.

4 Topic Modeling

To move beyond the limitations of clustering we use the latent Dirichlet allocation (LDA) model to hopefully capture both the content and structure emails. LDA is a popular form of topic modeling that has historically been effective for identifying meaningful “topics” in text corpuses. See Figure 2 for a graphical representation of the model. The model represents each document as a mixture of topics, each topic as a probability distribution over words in the vocabulary, and therefore draws each word based on the topic proportions in the document it comes from. θ represents the topic proportion for each document and β represents the topic weights for each word. The model is estimated using variational Bayesian methods (for more details, see Blei and Lafferty, 2009).

The primary parameter that we need to specify for LDA is the number of topics. Since

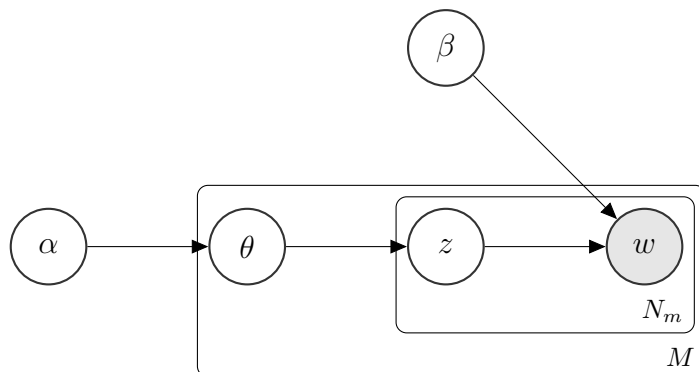


Figure 2: LDA Graphical Representation

our ultimate goal is to find sensible topics that can be used to understand the email corpus, we initially turned to coherence measures, which attempt to quantify whether estimated topics would make sense to a human observer. In particular, we considered two measures, proposed by Mimno et al. (2011) and Newman et al. (2010). However, the coherence of our topics did not appear to vary much across a wide range of reasonable topic numbers, anywhere from 15 to 50 topics used. As a result we manually examine the results from models using a range of topics and attempt to select the most parsimonious one, resulting in a 30 topic model.

4.1 Interpreting Topics

After choosing and estimating our topic model, the first task is to check whether we can actually find meaning in the resulting topics. If topics correspond to meaningful issues or structures then we can use these results to make useful observations about the data.

This ability to then perform “human” analysis on the results is one of the key features of LDA. Bringing human intuition to bear on a 28,000 email corpus is nearly impossible, but by condensing the data into a handful of topics LDA allows us to leverage our human understanding of semantics and meaning to analyze the data.

Interpretation is most easily done by examining the top words from each topics to find patterns and themes. Table 1 shows the top 10 words for each of the 30 estimated topics. The LDA algorithm makes no attempt to interpret these meanings, but by viewing them there are several easily distinguishable patterns. For example, many of the top words in Topic 6 relate to Israel and Palestine. Not every topic is as well-identified as this, and some of the estimated topics are thematically weaker than others, but overall the topics seems to correspond very well to interpretable subjects.

Another useful example is Topic 8, which does not seem to have any semantically meaningful words but primarily consists of “common” words like think, good, and tell. These “structural” topics illustrate one of the key advantages of the LDA model over spectral clustering. As we discussed in Section 3.1, spectral clustering is only able to identify clusters based on the structure of emails, not their meaning. This is presumably because it primarily relies on the distribution of these “common” or “structural” words. By separating these types of words into separate topics, LDA is therefore able to better identify meaning in emails independent of their structure.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
call	mistr	know	see	state	israel	peopl	think	meet	discuss
huma	last	just	report	benghazi	isra	world	like	reuters	letter
abedin	said	want	haiti	inform	peac	power	good	follow	draft
schedul	parti	let	news	date	palestinian	polit	say	update	cdm
sheet	david	ask	great	subject	netanyahu	mani	dont	list	note
request	prime	come	read	agreement	east	right	look	happi	eam
speak	deal	tri	media	depart	negoti	can	got	monday	secretar
readout	leader	sure	event	doc	arab	govern	tell	scschedule	question
calls	govern	thx	post	hous	state	one	thing	set	final
phone	ireland	best	connect	produc	middl	american	idea	mtg	brief

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
democrat	women	new	will	make	secretari	senat	libya	iran	time
republican	clinton	foreign	work	well	depart	bill	secur	syria	press
american	right	trip	week	take	office	vote	travel	nuclear	hrc
polit	hillari	told	need	way	room	health	libyan	syrian	staff
parti	secretari	import	next	point	state	care	iraq	turkey	dinner
elect	year	agre	also	hope	arriv	pass	embassi	ruussia	lona
obama	state	high	plan	time	rout	hous	attack	regim	ambassador
percent	human	head	start	move	privat	john	kill	iranian	close
candid	day	later	issu	possibl	offic	committe	march	intern	monica
poll	mani	spoke	keep	one	department	republican	august	opposit	remark

Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
fyi	said	develop	email	presid	china	afghanistan	can	statement	tomorrow
print	book	state	offic	obama	econom	pakistan	talk	egypt	today
sid	one	support	sent	said	year	afghan	get	egyptian	cheryl
thank	year	global	messag	hous	chines	militari	speech	respond	sullivan
memo	two	program	pleas	white	job	general	now	announc	jacob
pls	day	effort	may	administr	million	karzai	back	releas	wil
latest	case	intern	check	polic	countri	offici	updat	elect	morn
info	famili	work	copi	aid	bank	war	send	saudi	jake
intel	school	includ	forward	advis	polic	forc	tonight	reuter	mill
urgent	court	diplomaci	blair	former	economi	nato	soon	presid	confirm

Table 1: Topic Top Words

4.2 Topic Categorization

Looking through the word lists, we can roughly categorize the topics into overarching themes. This is obviously a subjective exercise, but is an example of how LDA can enable human, subjective analysis of the data. Table 2 contains an example categorization.

In the LDA model, a document is a probability distribution over the available topics. We also estimate these topic proportions for each email. To get an idea of the most common topics, we identify the “primary” topic in each email.³ Table 3 ranks the topics in terms of how frequently they are the primary topic.

There are several interesting observations we can make from Tabela 2 and 3. An easy one is that many of these emails are about meetings and similar issues. This not only makes

³To identify primary topics, we ignore “common” topics as identified in Table 2. Then for each email we identify the highest percentage topic as the “primary” topic.

■	Terrorism	5, 22
■	Middle East	6, 18, 19, 27, 29
■	Foreign Policy	2, 7, 13, 21, 23, 26
■	Politics	11, 17, 25
■	Staff	16, 20
■	Press	4
■	Hillary	12
■	Meetings	1, 9, 10, 24, 28, 30
■	Common	3, 8, 14, 15

Table 2: Topic Categorization






																									
21	1	28	30	10	9	4	24	18	29	16	12	20	17	13	19	23	27	22	6	2	25	26	7	11	5
Most Frequent												Least Frequent													

Table 3: Primary Topic Frequency

sense because of Clinton’s position as Secretary of State, where she presumably spends a large amount of time in meetings or briefings, but also because these shorter, administrative emails create high volume and are constantly being sent.

We can also observe that relatively few of the topics are about domestic issues such as health care, the economy, or education. Again, this matches what we would expect given that Clinton’s position of Secretary of State means that she primarily deals with foreign affairs, which many of the topics are related to.

4.3 Topic Analysis

We can also look into specific topics in more detail to see what they consist of and how they might correspond to real-world events. Table 4 shows the top 10 words for six topics that are all examples of content topics. In particular, these topics correspond to specific politically significant events or issues. For instance, the Libya topic seems to capture discussion about the Benghazi scandal as well the collapse of Muammar Gaddafi’s administration.

The model seems to effectively identify words that human observers would associate with these events. Topic 11 picks up on both of the primary political parties in the United States as well as other words associated with elections, such as politics, percent, and candidate. Similarly, the Topic 23 identifies words that associated with international development, including develop, support, and global.

Given that we have time stamps for each email, we can also examine how prevalent each

Topic 6	Topic 11	Topic 18	Topic 27	Topic 23	Topic 25
Israel	Elections	Libya	Afghanistan	Int. Dev.	Obama
israel	democrat	libya	afghanistan	develop	presid
isra	republican	secur	pakistan	state	obama
peace	american	travel	afghan	support	said
palestinian	polit	libyan	militari	global	hous
netanyahu	parti	iraq	general	program	white
east	elect	embassi	karzai	effort	administr
negoti	obama	attack	offici	intern	polic
arab	percent	kill	war	work	aid
state	candid	march	forc	includ	advis

Table 4: Some example content topics

topic is over time. Figure 3 shows the mean monthly topic proportion for each of the topics in Table 4.

We can see that our topics seem to be correlated with the expected real-world events. The Israel topic’s first peak corresponds to the July 26th, 2010 joint Israeli and Romanian helicopter disaster, while the second peak corresponds to the September 9th, 2011 attack on the Israel embassy in Egypt. Similarly, for the election topic each of the peaks corresponds to a U.S. election. Looking at the Libya topic, the first peak corresponds to the fall of Gaddafi’s administration in October of 2011, while the second peak corresponds to the Benghazi attack on September 11th, 2012. These kinds of patterns over time are interesting to verify, and could also be used to identify the prevalent topics of discussion for times when there are not obvious world events driving discussion.⁴

⁴These monthly topic proportion plots for all our estimated topics are available at https://github.com/lbybee/601_Final_Project/tree/master/images.

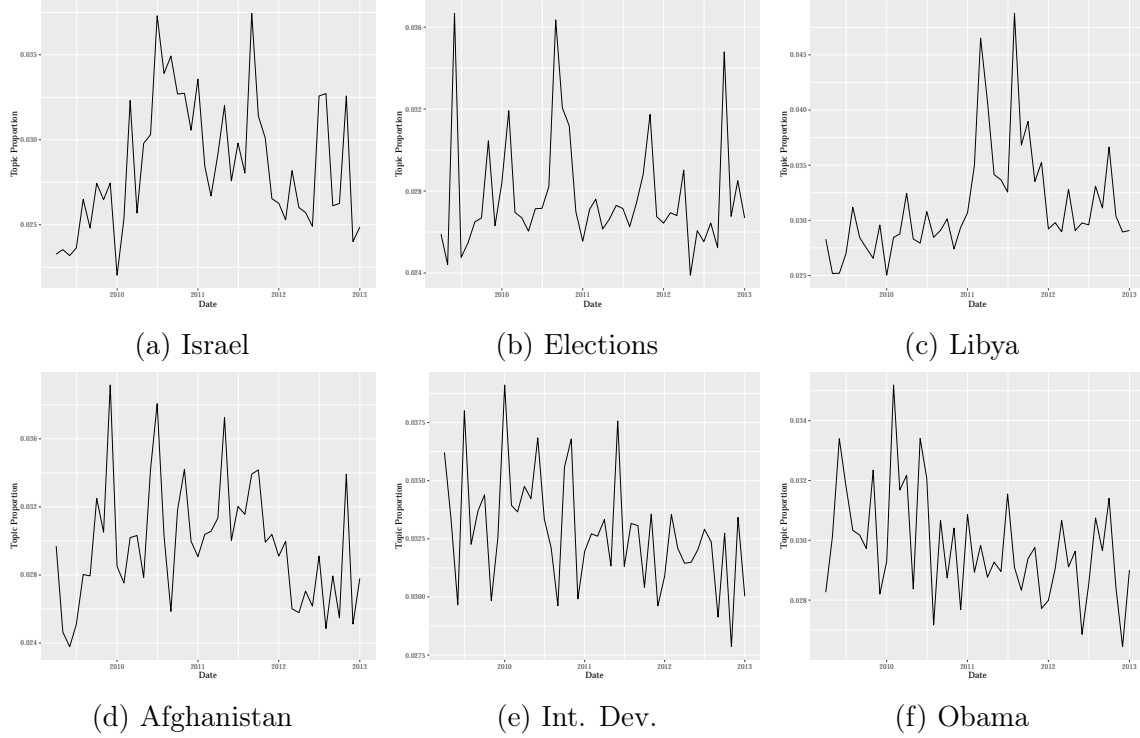


Figure 3: Average monthly topic proportions.

5 Email Sender Prediction

Finally, we want to consider how we might incorporate sender information with our topic model. There are a total of 395 unique senders in our email corpus, however, only 14 senders sent more than 100 emails. We focus on this subset of 14 senders for the remainder of our analysis. This leaves us with a subset of 25,809 emails. We use the topic proportions from the model estimated on the full corpus for this subset.

Table 5 presents the top five senders for each of several example topics based on the average topic proportions for each sender. The sorting that we get back seems to fit our understanding of the senders. For example, Sidney Blumenthal, who appears to be most associated with the Israel, Election and Obama topics is a long standing advisor to the Clinton family who has written extensively on U.S. politics as well as foreign policy. Similarly, Huma Abedin, who is most associated with Libya, was a key figure during the Benghazi hearings and was generally strongly associated with foreign policy (especially regarding the Middle East) while serving as Hillary Clinton’s Deputy Chief of Staff. Judith McHale, who is most associated with the Afghanistan and international development topics, was Under Secretary of State for Public Diplomacy and Public Affairs during the early portion of of the 2008-2013 time period. She has also been strongly associated with a number of philanthropic

Topic 6 Israel	Topic 11 Elections	Topic 18 Libya	Topic 27 Afghanistan	Topic 23 Int. Dev.	Topic 25 Obama
Sidney Blumenthal (SB)	SB	Huma Abedin (HA)	Judith McHale (JM)	JM	SB
HA	Philippe Reines (PR)	Wendy Sherman (WS)	HA	Melanne Verveer (MV)	PR
Jake Sullivan (JS)	JM	JS	MV	Anne-Marie Slaughter (AMS)	Cherly Mills (CM)
AMS	CM	Monica Hanley (MH)	JS	CM	HA
Hillary Clinton (HC)	HA	SB	Richard Verma (RV)	JS	HC

Table 5: Primary senders for some example topics

efforts.

It is interesting to note that Clinton does not appear to be strongly associated with any of these content topics. This pattern persists through the remainder of the content topics; in general, Clinton does not appear to send nearly as many emails strongly associated with content topics as she receives.

5.1 Multinomial Logistic Regression

While the above results do give some sense of which senders are most associated with which topics, we next want to formalize this approach and get more general results that can more easily show us which senders are most related to each topic. To approach this problem we use multinomial logistic regression. Multinomial logistic regression generalizes the logistic regression approach to cases with more than two outcome labels. The probability that the j th label is assigned to the i th email is

$$p(y_i = j \mid x_i, \beta) = \frac{\exp(\beta_j^T x_i)}{\sum_{k=1}^K \exp(\beta_k^T x_i)} \quad (4)$$

This produces a $P \times K$ matrix of coefficients, β , corresponding to how strongly related each of the K variables of interest are related to each of the P labels.

For our particular case we estimate a model using the 30 topics as our explanatory variables and the 14 senders as our labels. We use a training subset of 23,229 emails and their corresponding topic proportions to train our model and a subset of 2,580 emails for testing. See Table 6 for the success rate for our testing data. Overall the model appears to perform well, correctly identifying roughly 30% of the senders correctly. For test emails sent

by Clinton we correctly identify the sender 74% of the time.

Source	Success Rate	Testing Observations
Hillary Clinton	0.74	849
Philippe Reines	0.06	34
Claire Coleman	0.52	27
Lauren Jiloty	0.42	71
Huma Abedin	0.36	376
Jake Sullivan	0.23	410
Sidney Blumenthal	0.29	87
Cherly Mills	0.31	491
Anne-Marie Slaughter	0.36	42
Monica Hanley	0.09	53
Judith McHale	0.25	28
Robert Russo	0.00	10
Richard Verma	0.26	19
Wendy Sherman	0.08	12
Melanne Verveer	0.30	30
Lona Valmoro	0.34	41

Table 6: Multinomial logistic regression test results.

The estimated coefficients from the multinomial logistic regression model give us a sense of how strongly each sender is associated with the different topics. See Figure 4 for a visualization of the resulting coefficient matrix.⁵

The results we see here all us to tell a story about the email senders that corresponds to our intuition based on each of the senders’ backgrounds. Sidney Blumenthal is strongly associated with the Middle East, foreign policy and general politics topics. He does not have much to say about Clinton, her significance to the press, or administrative topics such as meetings. Huma Abedin and Jake Sullivan provide advice on the Middle East and foreign policy. Judith McHale focuses her emails on foreign policy and handling press issues. And, as discussed before, Clinton herself primarily sends emails about meetings.

The pattern that seems to emerge from these explorations is that Clinton herself does not discuss world events. Instead, she has surrounded herself with a set of advisors who constantly inform her about issues while she mostly uses email to coordinate meetings and calls where (presumably) she does her real work.

⁵Note that the coefficients are standardized and then thresholded at ± 2 for illustrative purposes.

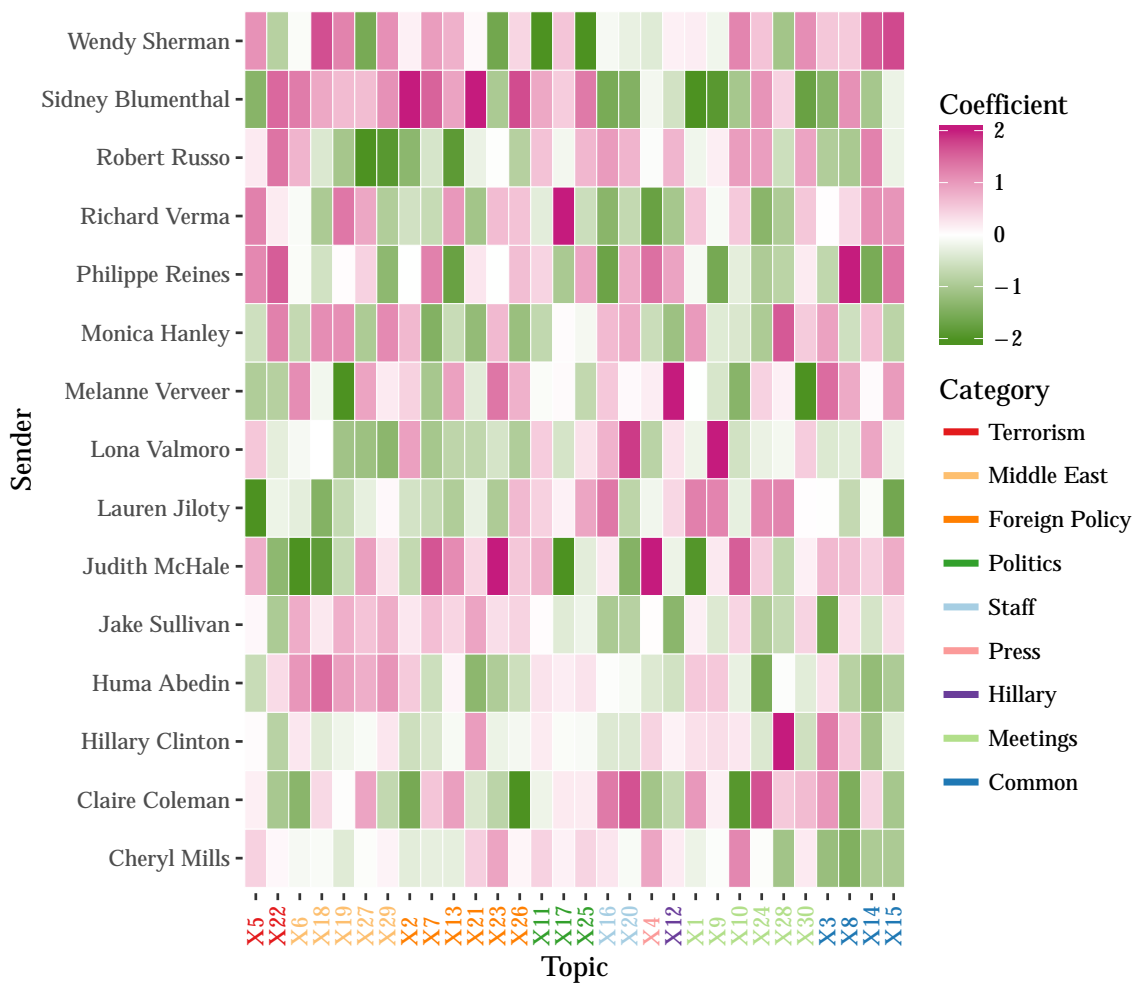


Figure 4: Standardized coefficients from a multinomial logistic regression of email sender on LDA topic proportions.

6 Conclusion

The release of the emails from the Clinton family server has provided us with an unprecedented and incredibly interesting and relevant dataset to explore. We have used statistical methods to condense this large corpus of emails so that analysis is approachable to the human user. We have been able to identify general structures in the email corpus using both spectral clustering and LDA. For LDA we find that we can separate content topics from structural ones and that the topics we get back are sensible and align with what we would expect. Additionally, we have applied simpler statistical methods to the results from LDA to make conclusions about email senders in our corpus.

There are many ways that this work could be expanded, and our work should be considered exploratory and a starting place. The amount of additional questions and topics

that could be explored in this dataset is essentially endless, and the following are only a couple more potential starting points. Information on receivers, while not as clean as sender information, could also be incorporated and explored. This could help build a better model of interactions within the email corpus, possibly even incorporating network information. Additionally, LDA itself can be extended in a variety of ways. Some examples include time-varying topics, time-varying topic frequencies, or even modeling the topic or word usage of individual senders.

Overall, we have started some explorations but have only scratched the surface of possible analysis that could be performed on the Hillary Clinton email corpus. We have demonstrated some of the important tools that can be used for future summary and analysis.

References

- Blei, D. M. and Lafferty, J. D. (2009). Topic models. In Srivastava, A. N. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*, chapter 4, pages 71–94. Chapman and Hall/CRC.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.
- Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.