Hillary Clinton Emails

Leland Bybee, Roger Fan, Ryan Vaughn

April 14, 2016

Data

- From 2008 to 2013, Hillary Clinton and some of her staff used a private email server for much of their communication
- In response to FOIA requests, the State Department has released pdfs of roughly 30,000 of these emails, with roughly 8,000 sent by Clinton
- We mostly want to condense and explore the data
  - Break down a huge interpretation problem to a human-sized one
  - Find specific and relatable observations, not just statistical ones

Data

- We have a dataset of all these released emails, including sender and receiver information[1]
- Bag-of-words model
  - Stemming: combining words with the same "root"
  - Remove extremely rare words and extremely common (stop) words
- After cleaning, roughly 28,000 emails and a total vocabulary of over 3,000 words

[1] Thanks to Ben Hamner and the WSJ for making their code and data available, respectively.

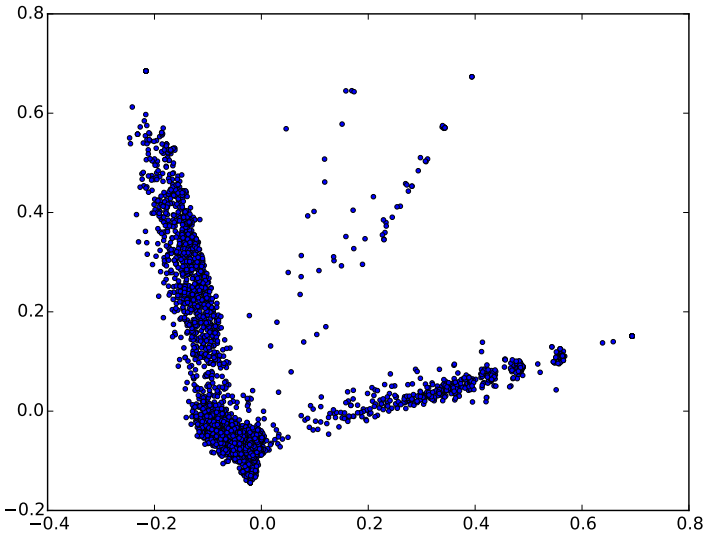
Visualization

- In order to visualize these emails, we first need a measure of distance
- Cosine similarity:

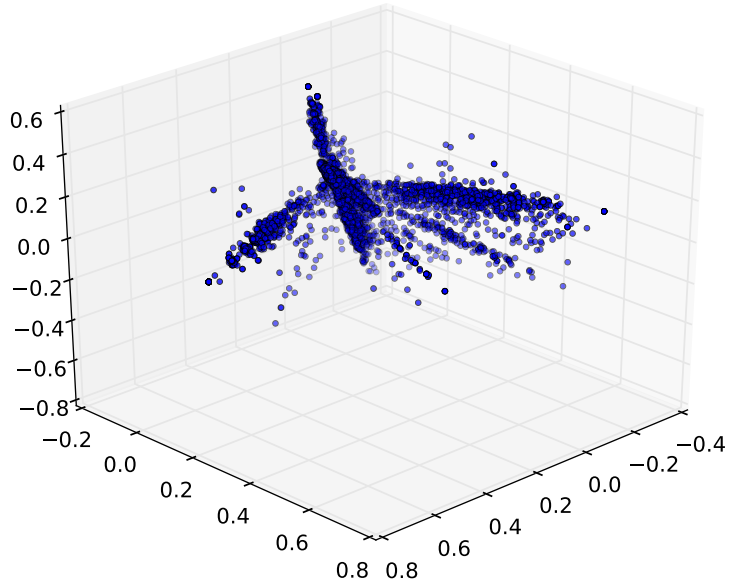
$$\cos(\phi) = \frac{\langle V, W \rangle}{\|V\|\|W\|}$$

- We can then use classical multi-dimensional scaling to visualize this data in lower dimensions

Visualization



(a) 2D



(b) 3D

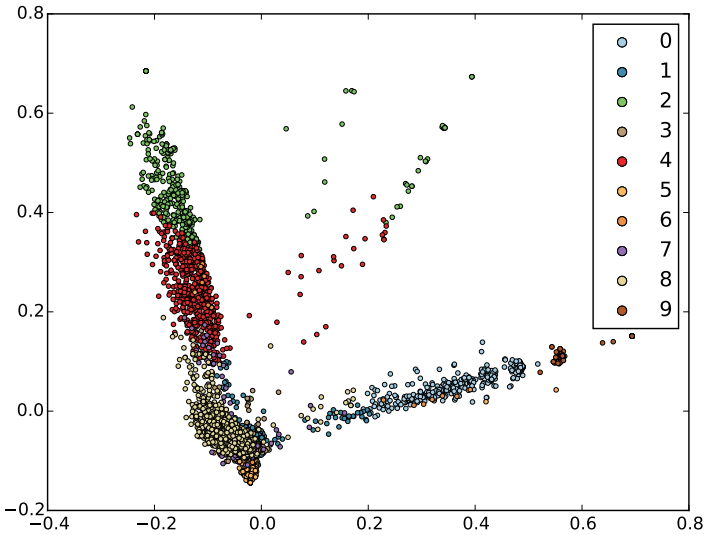
Spectral Clustering

- Cluster points using their pairwise similarity
- Computer the first K eigenvectors of the normalized Laplacian

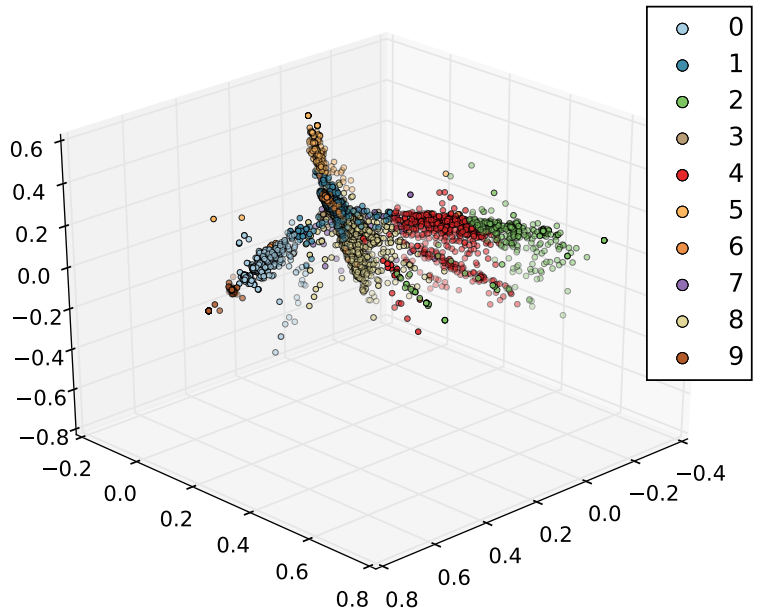
$$L = I - D^{-1/2} S D^{-1/2}$$

- Perform k-means clustering on the resulting eigenvectors

Spectral Clustering

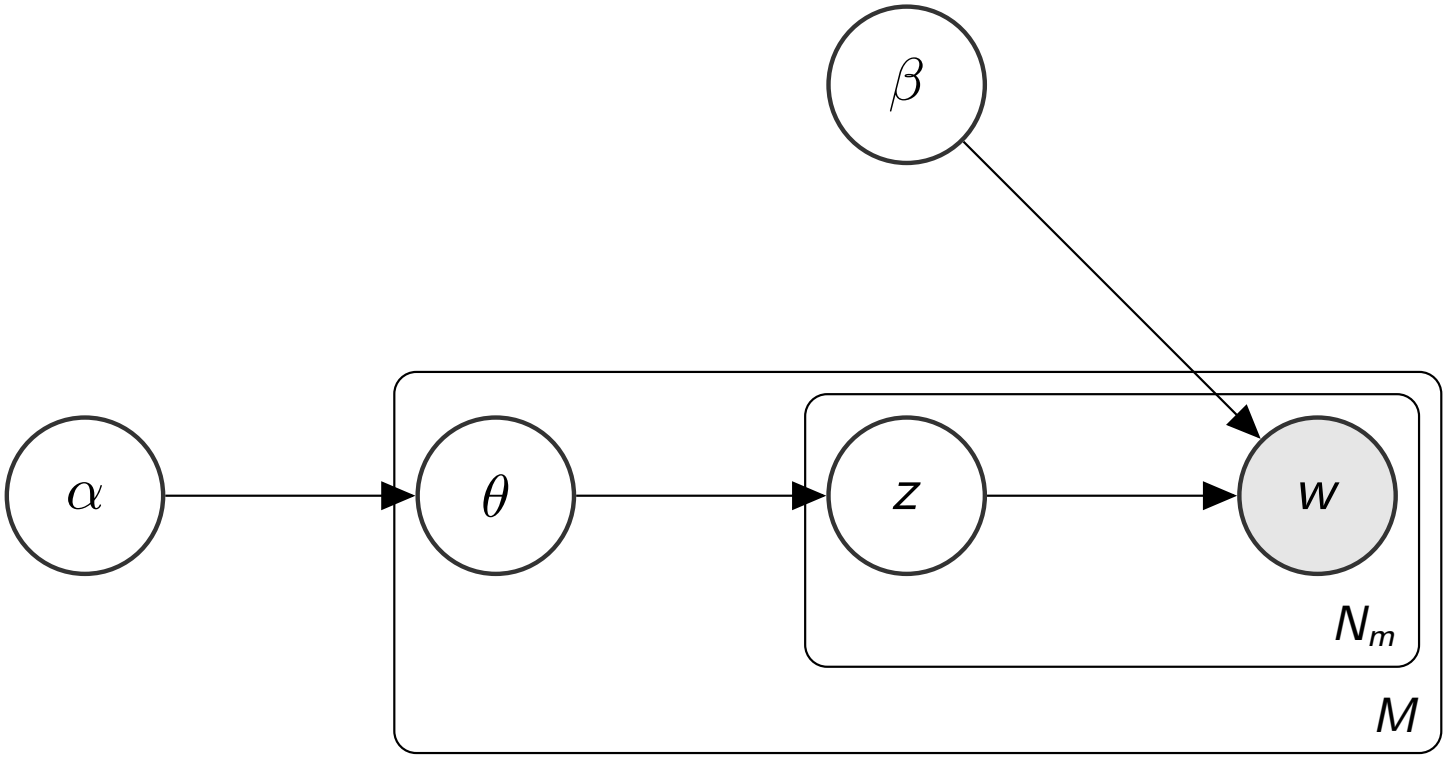


(c) 2D



(d) 3D

Latent Dirichlet Allocation



Latent Dirichlet Allocation: Some Details

- We use $K = 30$ topics for our resulting model.
- What we ultimately care about are θ and β.
- θ corresponds to how likely a topic is to appear in a document.
- β corresponds to how likely each word is to be associated with each topic.

Latent Dirichlet Allocation: Some Details

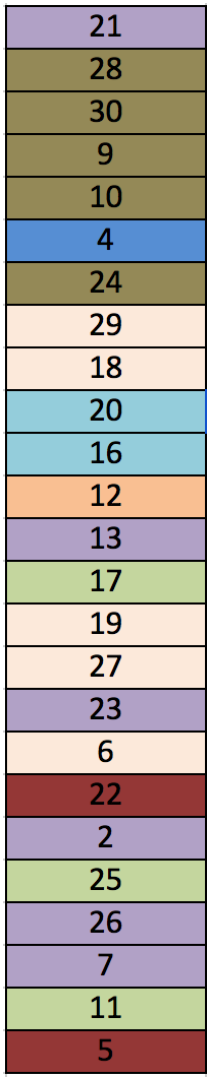
- Some questions we want to answer with LDA
  - Can we see sensible topics in the LDA output?
  - What are the most common topics of the emails?
  - Do the topics for each email line up with real world events?
  - Can we associate email senders with topics?

Exploring Topics

Common	UK	Press	Israel	XXXXXXX XXXXXXX XXXXXXX	Meetings (General)	US Politics (Elections)	XXXXXXX XXXXXXX XXXXXXX	Hillary and Staff	US Politics (Governing)	Syrian Civil War/ Conflict With Russia	Obama Administration	International Economic Policy	Middle East (Afghanistan Pakistan) Military
Military/War													
Middle East													
Africa/Middle East	minist	see	israel	think	meet	democrat	make	secretari	senat	iran	presid	china	afghanistan
Isreal/Palestine	last	report	isra	like	reuters	republican	well	depart	bill	syria	obama	econom	pakistan
Foreign Affrs (General)	said	haiti	peac	good	follow	american	take	office	vote	nuclear	said	year	afghan
Terrorism	parti	news	palestinian	say	update	polit	way	room	health	syrian	hous	chines	militari
UK/EU	david	great	netanyahu	dont	list	parti	point	state	care	turkey	white	job	general
Russia	prime	read	east	look	happi	elect	hope	arriv	pass	russia	administr	million	karzai
Asia	deal	media	negoti	got	monday	obama	time	rout	hous	regim	polici	countri	offici
North/South America	leader	event	arab	tell	schedule	percent	move	privat	john	iranian	aid	bank	war
United States	govern	post	state	thing	set	candid	possibl	offic	committe	intern	advis	polici	forc
Sec of State Affairs	ireland	connect	middl	idea	mtg	poll	one	departmen	republican	opposit	former	economi	nato
Obama Admin	brown	piec	jewish	anyth	friday	new	that	meet	said	arm	chief	state	diplomat
Economy	two	articl	settlement	much	tom	campaign	end	white	sen	arab	first	fund	american
Diplomacy	gordon	interview	land	love	followup	public	help	confer	reform	forc	washington	problem	taliban
Women's Issues	cameron	stori	jerusalem	didnt	weekend	voter	might	residence	kerri	weapon	clinton	financi	troop
Press	northern	africa	gaza	cant	add	group	know	meeting	boehner	assad	approv	market	strategi
Politics/Politicians	per	help	process	realli	thursday	support	happen	daili	hold	militari	offici	billion	defens
Government	support	januari	bibi	still	tuesday	vote	alreadi	time	legisl	support	havent	also	oper
Meetings/Scheduling	labour	suggest	west	hes	wrote	presid	made	staff	congress	unit	nation	trade	petraeus
(National) Security	polit	june	minist	your	return	nation	better	house	amend	resolut	enough	money	command
Intelligence	vote	page	parti	someth	afternoon	conserv	even	outer	democrat	war	senior	govern	civilian
Aid/Cooperation	time	hit	envoy	lot	lunch	like	matter	hous	member	leagu	biden	cut	pakistani
Domestic Issues	tori	gave	bank	day	wednesday	run	clear	photo	reid	council	barack	growth	iraq
Islam	hagu	cover	american	ive	bob	major	case	floor	leader	sanction	favor	deal	gate
	justic	cnn	prime	friend	saturday	among	long	mini	hill	russian	staff	need	secur
	day	offer	unit	heard	wed	win	believ	airport	mccain	zone	jone	aid	kabul
	first	ope	abba	what	cell	year	sinc	press	debat	lebanon	campaign	mean	said
	british	highlight	secur	big	fine	tea	rais	senior	floor	govern	state	unit	mission
	say	article	quartet	feel	item	romney	need	presidenti	nomin	intervent	team	energi	armi

Primary Topic Frequency

Meetings	9,10,24,28,30
Middle East	6,18,19,27,29
Staff	16,20
Domestic Politics	11,17,25
Terrorism	5,22
Foreign Policy	2,7,13,21,23,26
Press	4
Women's Issues	12



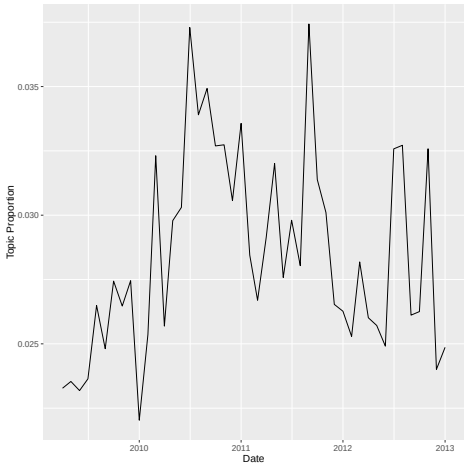
Some "Not" Interesting Topics

Common 1	Common 2	Common 3	Logistics	Meeting	Communication
know	think	will	secretari	meet	call
just	like	work	depart	reuters	huma
want	good	week	office	follow	abedin
let	say	need	room	update	schedul
ask	dont	next	state	list	sheet
come	look	also	arriv	happi	request
tri	got	plan	rout	monday	speak
sure	tell	start	privat	schedule	readout
thx	thing	issu	office	set	calls

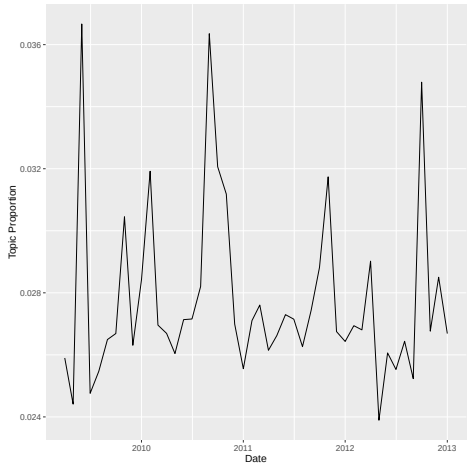
Some Interesting Topics

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	polici
arab	percent	kill	war	work	aid
state	candid	march	forc	includ	advis

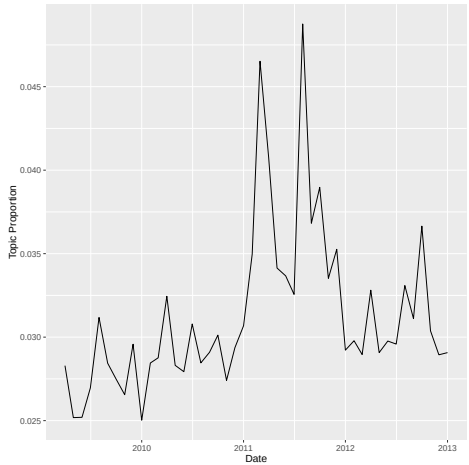
Topic Importance Over Time



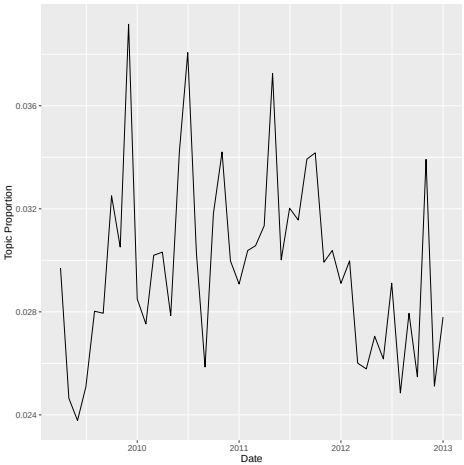
(e) Israel



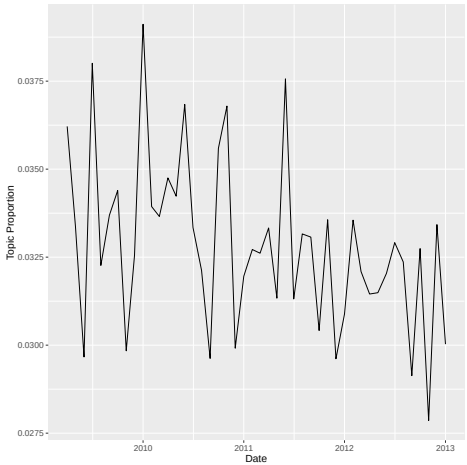
(f) Elections



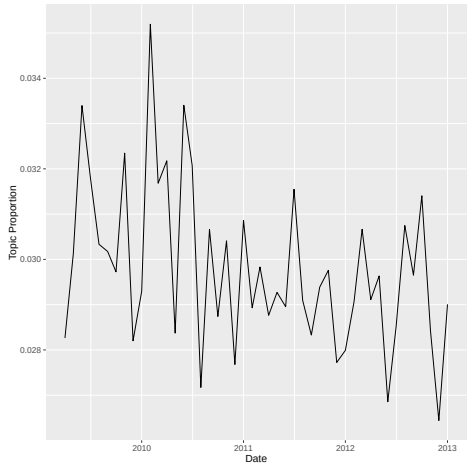
(g) Libya



(h) Afghanistan



(i) Int. Dev.



(j) Obama

Who Says What?

Israel	Elections	Libya	Afghanistan	Int. Dev.	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

Multinomial Logistic Regression

- Can we do a better job associating email senders with topics?
- To answer this question we employ multinomial regression.
- We take a subset of people who sent emails, only those who have more that 100 emails sent and try to predict the sender using the topics.

Multinomial Logistic Regression Results

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

Conclusion

- We've broken the corpus down to a level where we can effectively analyze it and draw conclusions using both human and statistical analysis
- This has only been a sample of the many questions we could address using the LDA output
- Possible future directions
  - Some emails are redacted, how do email topics relate to this
  - Correlation between topics, what topics appear together
  - Incorporating information on receivers
- If this data intrigues you, all the code we used to download, process, and analyze the data is (will be) available at `https://github.com/lbybee/601_Final_Project`.