Hillary Clinton Emails

Leland Bybee, Roger Fan, Ryan Vaughn

April 11, 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 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.

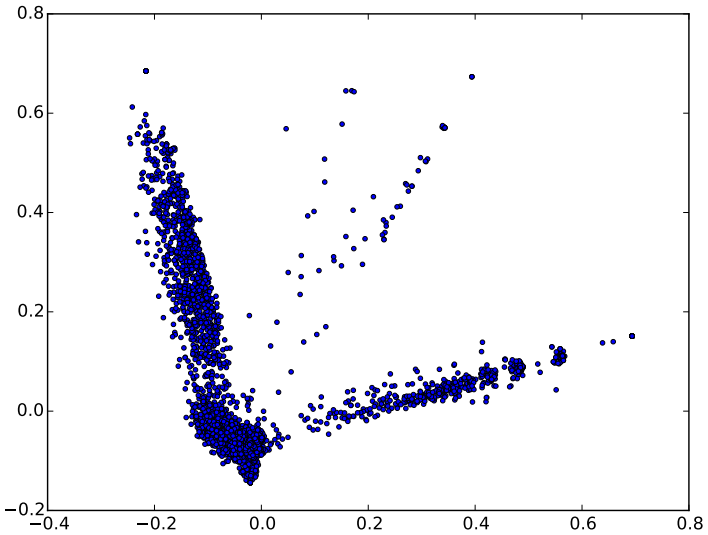
Visualization

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

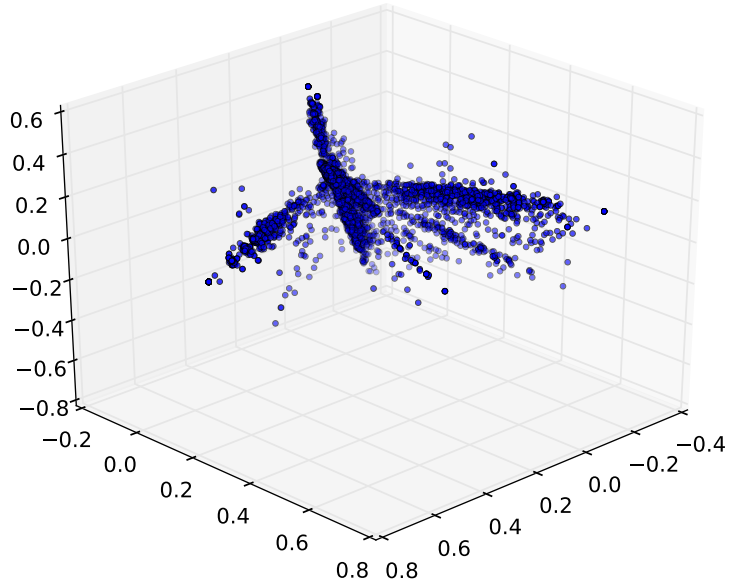
$$\cos(\phi) = \frac{< V, W >}{\|V\|\|W\|}$$

- We can then use classical Multi-dimensional Scaling to visualize this data in lower dimensions

Visualization



(a) 2D



(b) 3D

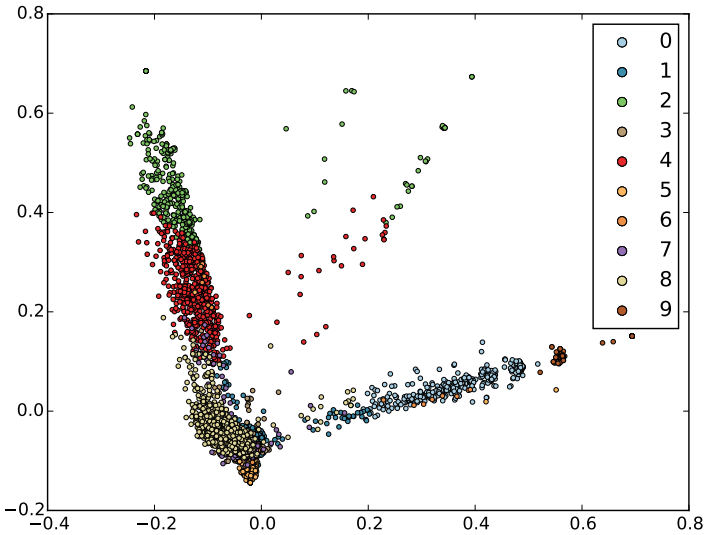
Spectral Clustering

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

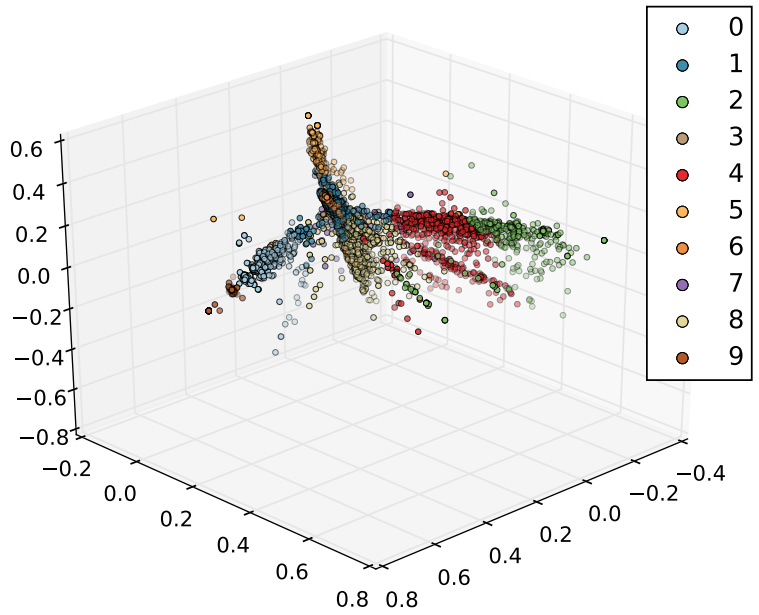
$$L = I - D^{-1/2} S D^{-1/2}$$

- Perform k-means clustering on the resulting eigenvectors

Spectral Clustering

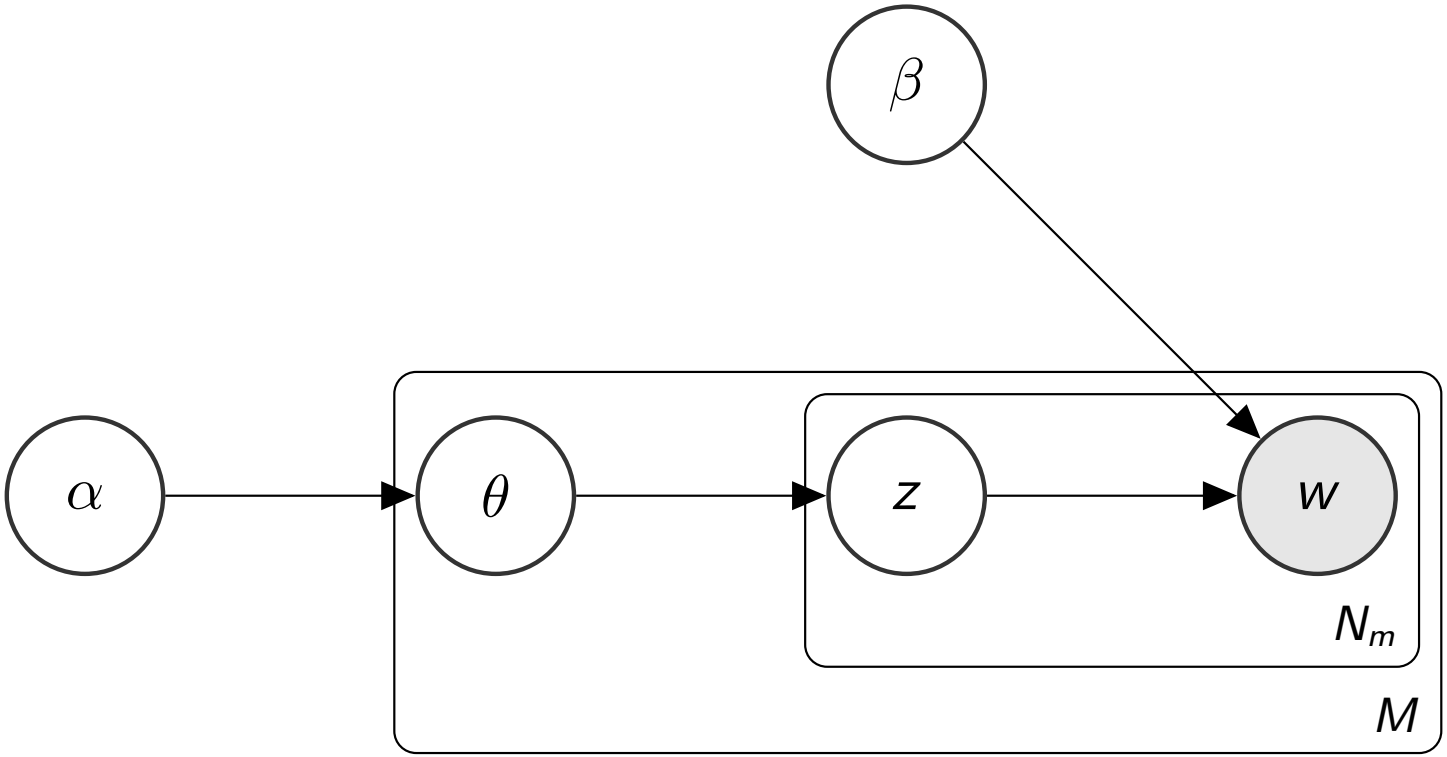


(c) 2D



(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.

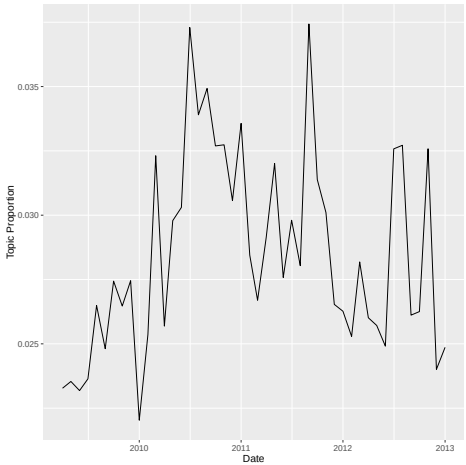
Latent Dirichlet Allocation: Some Details

- Some questions we want to answer with LDA
  - Can we see sensible topics in the LDA output?
  - Who is associated most with each issue?
  - Do the topics proportions for each email line up with real world events?
  - Can we predict the source of a given email?

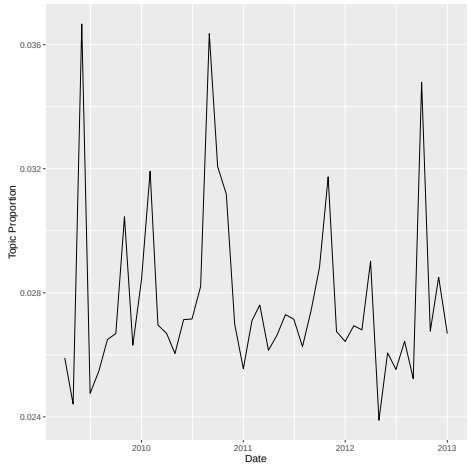
Some Interesting Topics

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

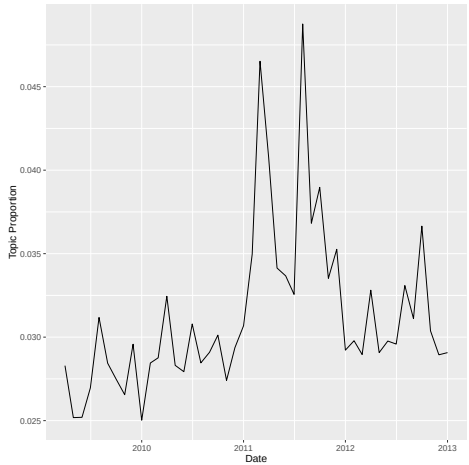
Topic Importance Over Time



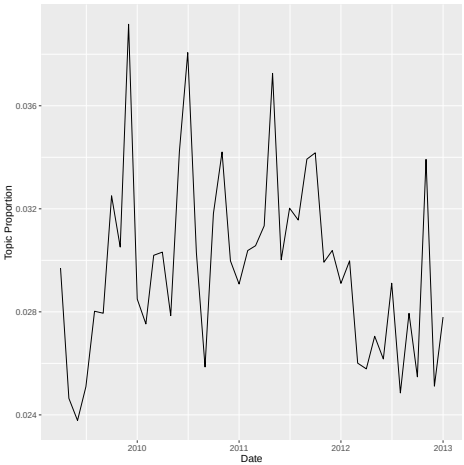
(e) Israel



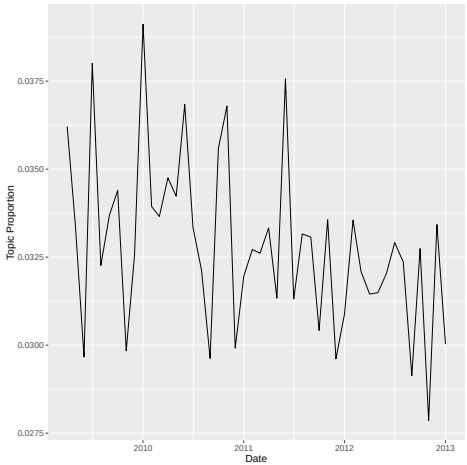
(f) Elections



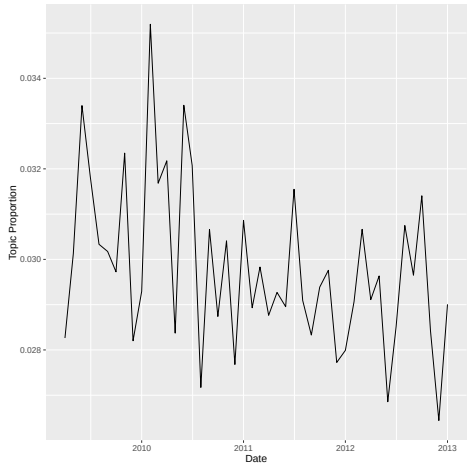
(g) Libya



(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

- We next want to ask can we predict who said what?
- 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 prediction for these.

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