Political Blogs:
A Dynamic Text Network

David Banks
Duke University

1. Introduction

Dynamic text networks arise in many situations related to national security:

- text and voice transmission via telephone and email;
- Internet content linked by browsing patterns;
- national political blogs.

One would like to use information in the text to improve network models of connectivity, or growth and change, and use network information to improve topic discovery.

We focus on U.S. political blogs, and seek to discover cliques that focus on topics relevant to security. Sentiment analysis may enable identification of groups that are becoming extreme.

To study this, MaxPoint Interactive and one of my graduate students scraped the text from the 1,509 top U.S. political blogs (as determined by `Technorati`) between Jan. 1 2012 and Dec. 31 2012. We are currently scraping the text from the 2016 blog posts (another election year).

The importance of a political blog is estimated by `Technorati` based on the the site's "standing and relevance in the blogosphere" as determined by the relevance of its content and link behavior.

Samples of the text were reviewed by poltical science majors at the University of North Carolina Chapel Hill. The review found some problems with inconsistent dating of blog posts, and instances in which the scrape captured not just the blog text but also comments posted by readers.

There were 114,611 posts on the 1,509 domains, generating 324,658 unique tokens (50 copies of War and Peace).

The first step in text mining is to remove stop words (and, of, the) and words that convey little topic distinctivity (therefore, interesting, whenever).

The next step is to tokenize the vocabulary, so that words with the same root are combined (fight, fights, fought).

The third step is to create n-grams, which are sets of words that co-occur improbably often, and thus denote a single idea (white house, Trayvon Martin, Supreme Court). This recovers some of the semantic information lost in bag-of-words models. And co-occurence can take place within a window, so that President Barack Obama, President Obama, President of the United States, and Barack Obama can all be equated.

Negation is a difficult problem.

Latent Semantic Indexing and **Word2vec** are procedures that address synonomy and polysemy by interpreting the meaning of words in the context of other words in the same document.

Synomyms are an issue for n-grams. Probability for the same "meaning" gets allocated across multiple sequences. But LSI can recognize synonyms:

- reduce the deficit by taxing job creators
- reduce the national debt by taxing fat cats

lead to the n-grams "job creators" and "fat cats" being nearby in term space.

Polysemy is harder; it requires disambiguations, and one wants to use only cues in the text, not domain knowledge, to do this.

For example, "Grateful Dead" can refer to a rock band or to a genre of German folktale. If the document includes the words "music" or "drugs" or "Haight-Ashbury" then the context suggests the former meaning. But if the document contains "woodcutter" or "coffin" or "magic goose" then the latter sense is implied.

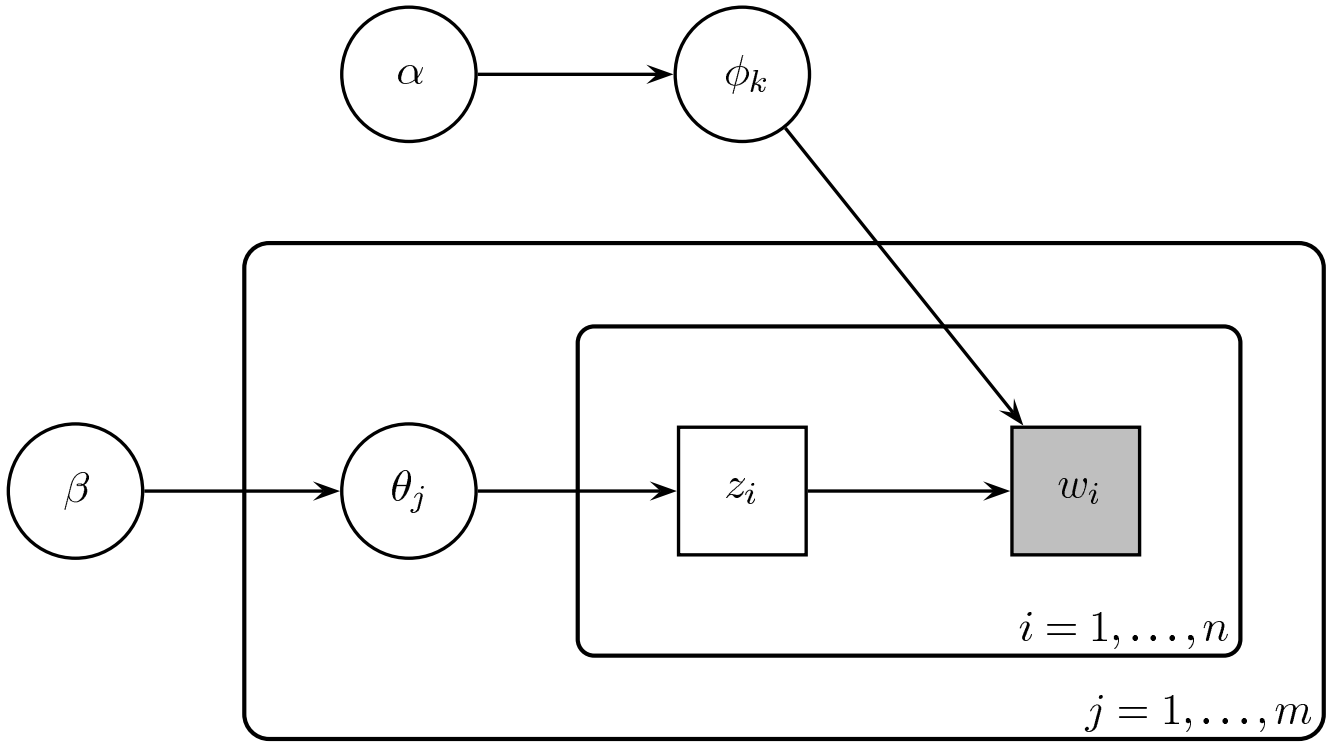
2. The Model

The most popular topic models use Latent Dirichlet Allocation (LDA), by Blei, Ng, and Jordan (2002). The generative model for assigning words to documents is as follows:

- For each topic, draw $\phi_k \in \mathbb{R}^v$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$; this determines the distribution over the vocabulary for topic k.
- For document D_j, draw $\boldsymbol{\theta}_j \in \mathbb{R}^K$ from a Dirichlet distribution with parameter $\boldsymbol{\beta}$, which determines the extent to which document D_j participates in each of the K topics.
- Independently, for each word in document D_j, first draw a single topic z_i from the one-trial multinomial with parameter $\boldsymbol{\theta}_j$. If one draws the ith topic, then the word is chosen as a single draw from the one-trial multinomial with parameter $\boldsymbol{\phi}_i$.

LDA chooses topics for each document according to one Dirichlet distribution, and then, conditional on a topic, the vocabulary is chosen according to its corresponding Dirichlet distribution.

This generative model is often described through a plate diagram, which represents the relationships between the mechanisms for composing documents as random mixtures of topics, with vocabulary drawn independently, with probabilities depending upon the topic.



To handle topic dynamics, we need a mechanism for change the vocabulary weights.

At each time step, the current distribution on the vocabulary can change in four ways:

- the weight on a word can be unchanged,
- the weight on a word can decrease by 10% (e.g., to allow Newt Gingrich's candidacy to slowly fade from blog discussion over the course of the year),
- the weight on a word can increase by 10% (e.g., Mitt Romney became a more popular topic as the campaign progressed)
- a new word can be added, and take large probability (e.g., Benghazi first appeared on Sept. 11, 2012, and had relatively high probability).

Every word has equal and independent probability of undergoing one of these changes. Then the distribution is renormalized.

To model the network structure, we suppose that each blogsite is assigned to a single block, and that there are a Poisson number of blocks. Members of a block are interested in the same set of topics, and with one exception, no block is interested in more than three topics.

A block is interested in at most three topics, with one exception. We force one block to be interested in all topics, to account for such things as The Huffington Post and the New York Times blogsite.

A mixed-membership stochastic blockmodel controls how likely members of the same block are to link a post, how likely members in different blocks that share one or more topics are to link to each other, and how likely members in blocks that share no topics are to link.

Additionally, there is a logistic regression model that controls the probability that one blogger links to another. The regression uses the following covariates:

- same block?
- same topic?
- ever linked to this blogger before?
- prestige of linkee
- base rate of linking
- excitation

Many of the weights derive from a random effects model.

The excitation refers to whether or not the topic is hot. We use a Hawkes process model to describe this.

As usual, MCMC reverse engineers this generative model to make inferences. To illustrate some of the results, we first look at the Trayvon Martin corpus.

Of the 114,611 blog entries that were scraped, 1,103 from 145 domains mention "Trayvon" one or more times. This was the corpus for analysis, together with links to made by those posts to any other of the 145 domains.

Recall that Trayvon Martin was an unarmed 17-year-old African-American teenager who was shot by George Zimmerman in a gated community in Sanford, FL. Zimmerman was the neighborhood watch coordinator for the community. He thought Martin was acting suspiciously, and reported him to the police. The police instructed Zimmerman not to approach the suspect, but were disregarded. Zimmerman claims that Martin attacked him, and he used his gun in self-defense; initially, he was not charged with any crime.

Sentiment analysis is a critical aspect of mining text networks for national security. Sentiment analysis involves trying to determine whether a document has a positive, negative, or neutral tone. Irony and sarcasm make this hard—humans agree only about 80% of the time.

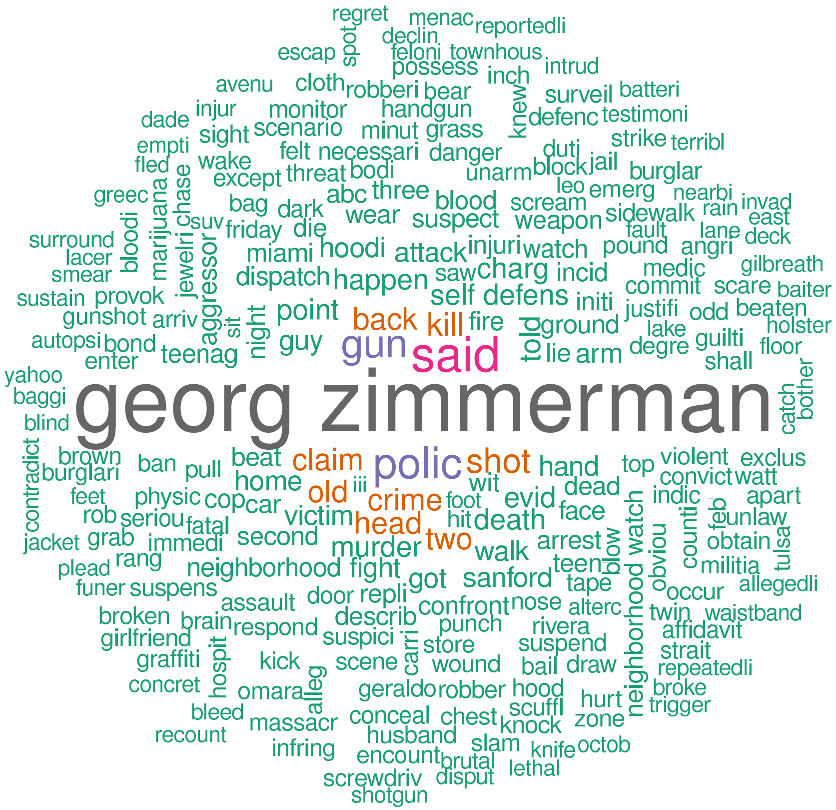
AFINN is a list of 2,477 English words scored from -5 to 5 in terms of positive/negative connotation (and negation reverses the sign). To reduce situational bias, before measuring the tones of the blog posts we removed terms that had negative association but which were necessarily relevant to the event; e.g., kill, gun, crime.

To measure the polarity of a document, let d_w be the number of occurrences of word w and let s_w be the AFINN score of that word. Then the polarity of a document is

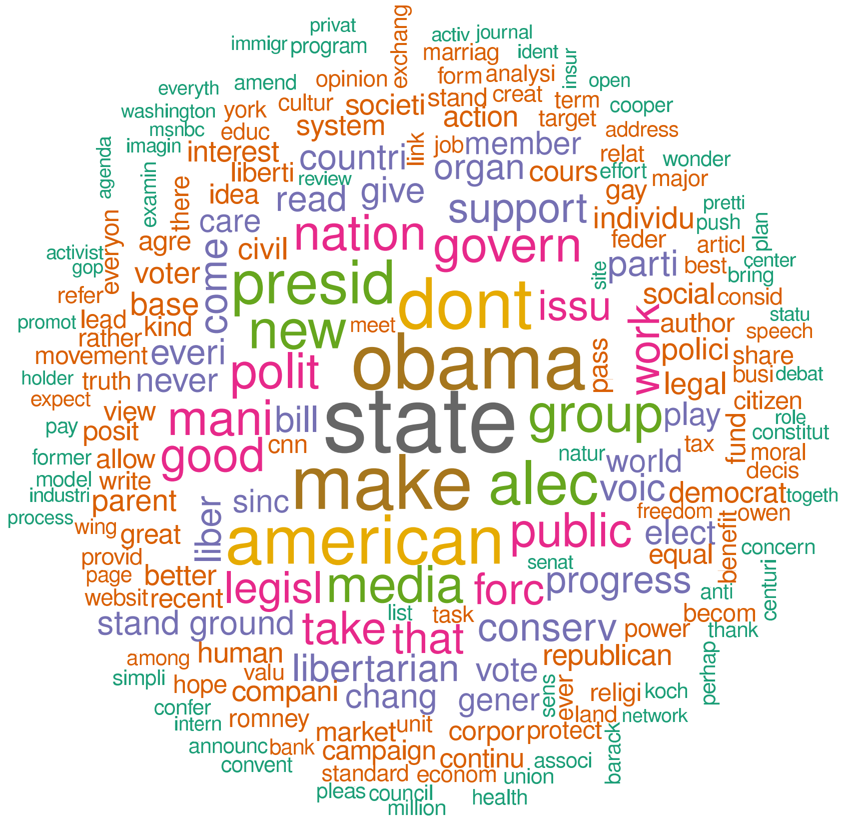
$$\textbf{polarity} = \frac{\sum d_w s_w}{\sum d_w |s_w|} \in [-1, 1]$$

We now identify the words in the blog posts that were differentially used between posts that had positive tone (i.e., polarity greater than 0.2) and those that had negative tone (polarity less than -0.2).

Negative Documents: Significant Words



Positive Documents: Significant Words



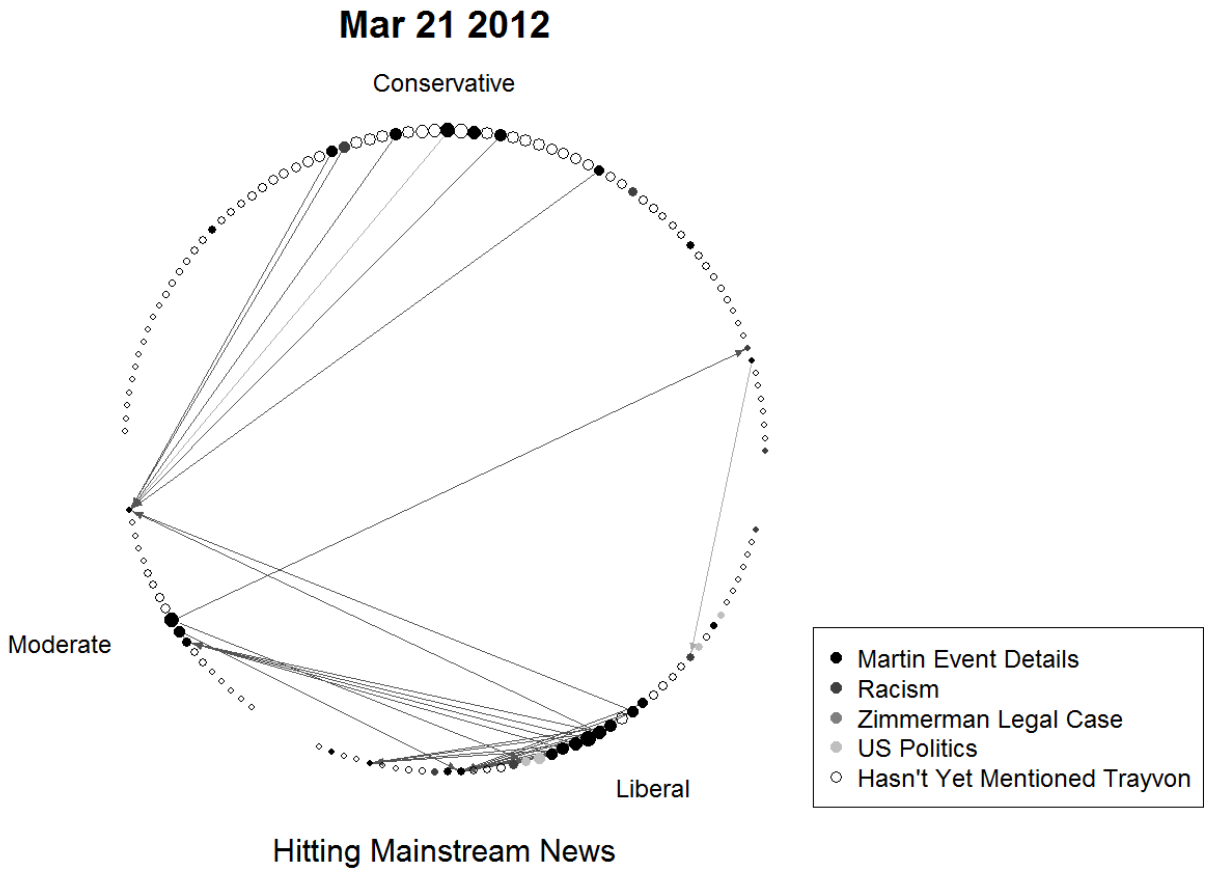
Next we used LDA, which found five topics in the blog posts. To characterize those topics, these are the n-gram tokens that loaded most heavily on each of the topics:

[1]	[2]	[3]	[4]	[5]
money	obama	think	georg zimmerman	trayvon martin
log	presid	dont	trayvon martin	black
scream	state	comment	polic	white
donat	law	would	law	georg zimmerman
dave	american	even	gun	media
voic	year	peopl	case	news
perjuri	alec	like	would	sharpton
expert	govern	one	self defens	racial
bond	gun	make	said	year
paypal	group	get	charg	hoodi
owen	republican	thing	prosecutor	look
bail	democrat	know	evid	said
forens	war	use	shot	fox
omara	nation	point	state	death
websit	legisl	say	shoot	race

LDA topics are interpreted impressionistically:

- Topic 1 seems to focus on the court case (O'Mara is Zimmerman's attorney, Owen is the forensic audiologist, and Zimmerman's disclosure of money raised from supporters on PayPal became a legal issue).
- Topic 2 relates to political aspects of the story.
- Topic 3 reflects the social function of blogging, and absorbs many unspecific words.
- Topic 4 is about the facts in the evolving story.
- Topic 5 is about racism.

The topic weights can be viewed as covariates that enable one to fit a dynamic network model to predict where links will form and how memes get passed among the blogs.



The blog network for the Trayvon Martin discussion when the event was noticed by national media.

4. The Full Corpus

It is helpful to look at an example of the text, before and after processing.

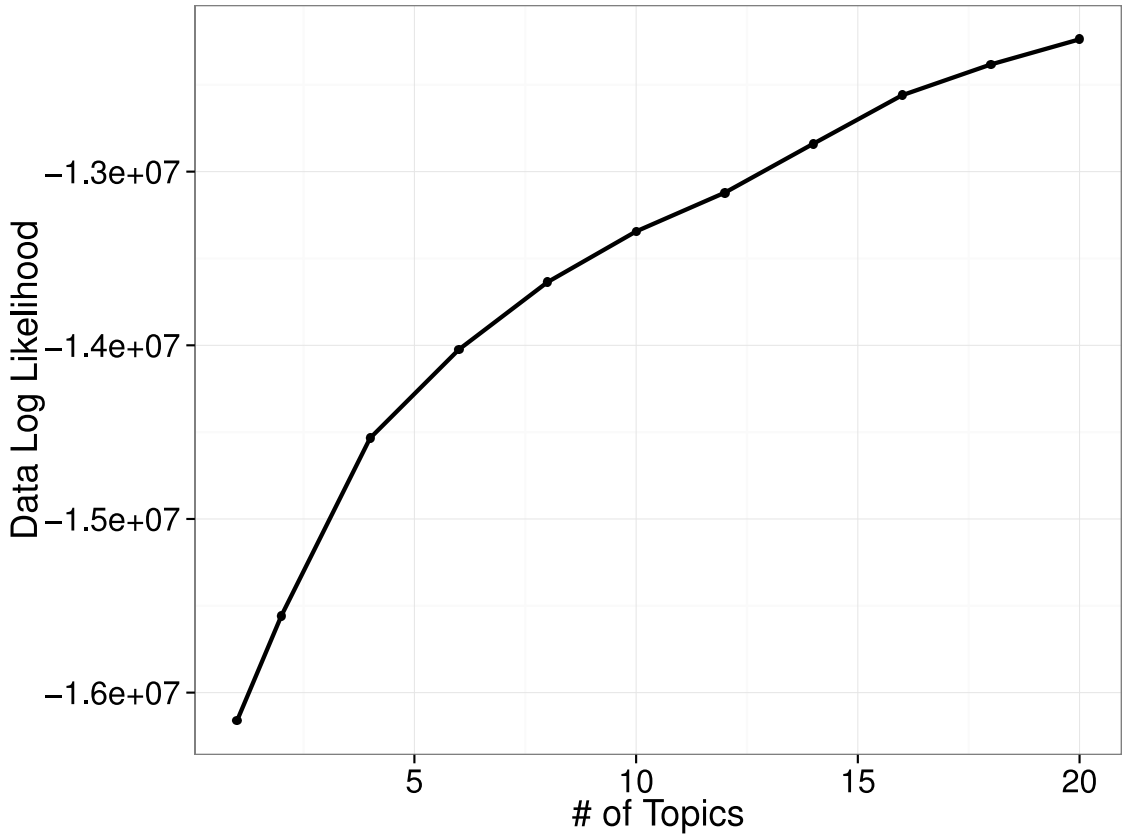
Domain: blog.heritage.org
Date: September 28, 2012
Text: "Recent news about the Obama Administration divestitures from AIG and GM in some cases at a loss of billions of dollars stands as a reminder of the privilege and cronyism . . . Thus we need policies that help all Americans, not just the ones who can afford lobbyists and large campaign donations"

Retained Tokens

ryan, olson, septemb, govern, less, econom, recent, news, obama, administr, divestitur, . . . , polici, help, american, afford, lobbyist, larg, campaign, donat

We applied LDA, as previously described, to the set of retained tokens. However, we allow the topics to evolve over time; i.e., the distributions change so that new words, such as Benghazi, can appear.

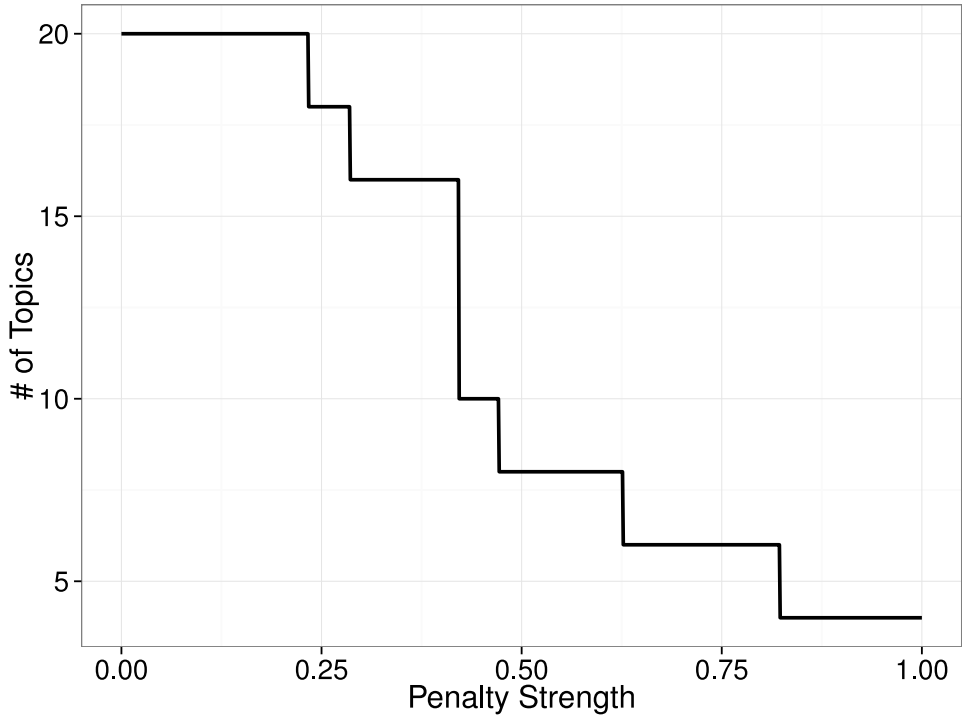
The next issue is to determine K, the number of topics. Since there is topic structure at multiple levels of resolution, this is not a well-posed question. Adding topics will always improve the likelihood.



For interpretability, we want a small number of topics. Using a penalty on the log-likelihood function

$$\mathbf{P} = \mathbf{K} \times (\#\text{ Time Periods}) \times (\#\text{ Words}) + \mathbf{K} \times (\#\text{ Documents}) + \mathbf{C}$$

gives the model score $S(s) = 2P \times s - 2L$. Since there is a jump from 10 to 16, we fit 16 topics.



As an example of the result of the fitted model, consider the tokens with high probability for Topic 10, for different weeks as it evolves through the year.

Weeks	Tokens
2012-02-19	polic, said, offic, kill, crime, arrest, year, case, death, charg, shoot, murder, drug, peopl, report, shot, crimin, home, told, call
2012-04-15	polic, said, zimmerman, case, martin, report, offic, crime, kill, arrest, charg, year, trayvon, crimin, shoot, death, murder, victim, peopl, call
2012-06-10	polic, said, year, crime, offic, case, report, charg, arrest, kill, crimin, kimberlin, death, peopl, call, drug, told, home, murder, victim
2012-08-05	polic, said, shoot, offic, kill, peopl, crime, year, right, arrest, death, murder, shot, report, crimin, home, fire, violenc, weapon, victim
2012-09-30	polic, said, arrest, kill, peopl, offic, crime, home, murder, year, right, death, free, told, crimin, charg, shoot, state, fire, victim
2012-11-25	kill, polic, peopl, said, shoot, murder, crime, right, year, home, weapon, death, free, victim, violenc, children, dead, control, state, happen

It is clear that Topic 10 is about crime. But it would be nice to have a less subjective and more automatic way to identify the topic. One strategy is to match the word distributions in the topics to word distributions in Wikipedia articles. To that end we scraped the Wikipedia categories "2012 in the United States", "2012 in American politics", "2012 in international relations", "United States federal policy", "Political controversies in the United States", "Media-related controversies in the United States", and "Politics of the United States by issue". The provided a total of 4,101 articles.

To quantify the distance between topics and labels we use the Total Variation Distance

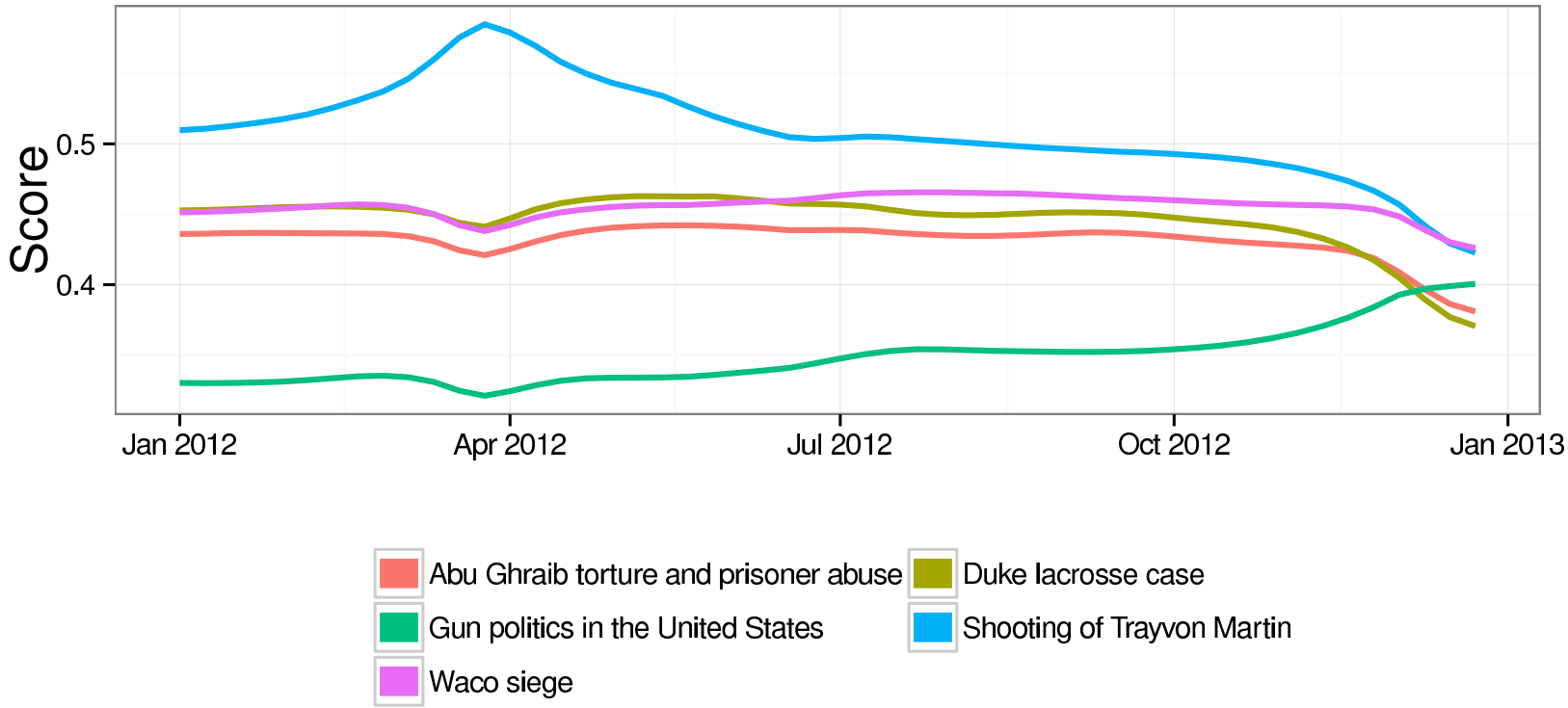
$$\delta(P, Q) = \sup_{\sigma \subset \Sigma} |P(\sigma) - Q(\sigma)|$$

which in this case is just

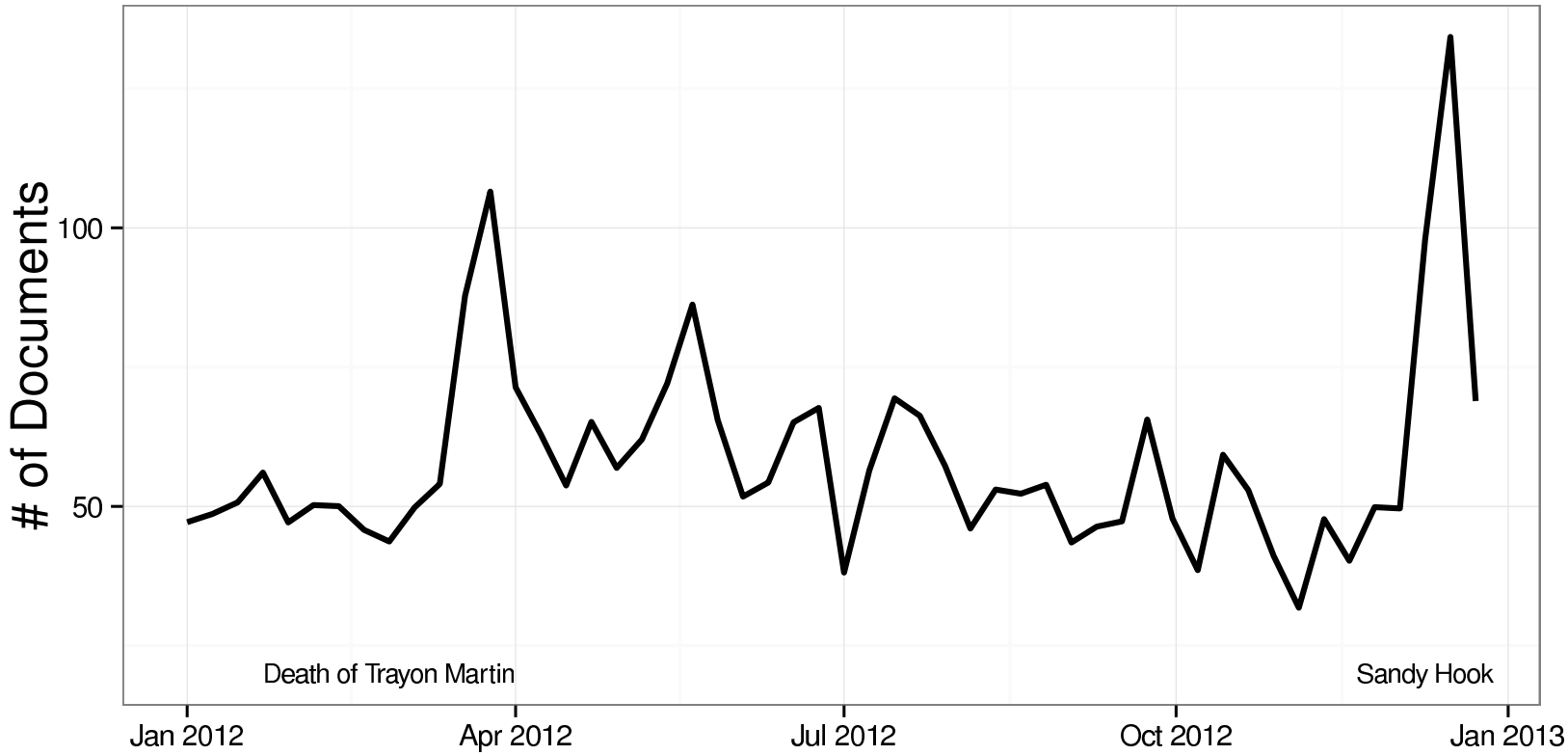
$$\delta_{TVD}(P, Q) = \frac{1}{2}||P - Q||_1$$

Going through our topics we score the fit of the Wikipedia articles by $1 - \delta_{TVD}(P, Q)$.

The following figure shows how weights on Wikipedia labels change over time for Topic 10. It is not perfect—the time line is over-smoothed, but it correctly picks up the Trayvon Martin and Sandy Hook shootings. The Duke lacrosse case label refers to several cases in which college athletes were charged with rape.



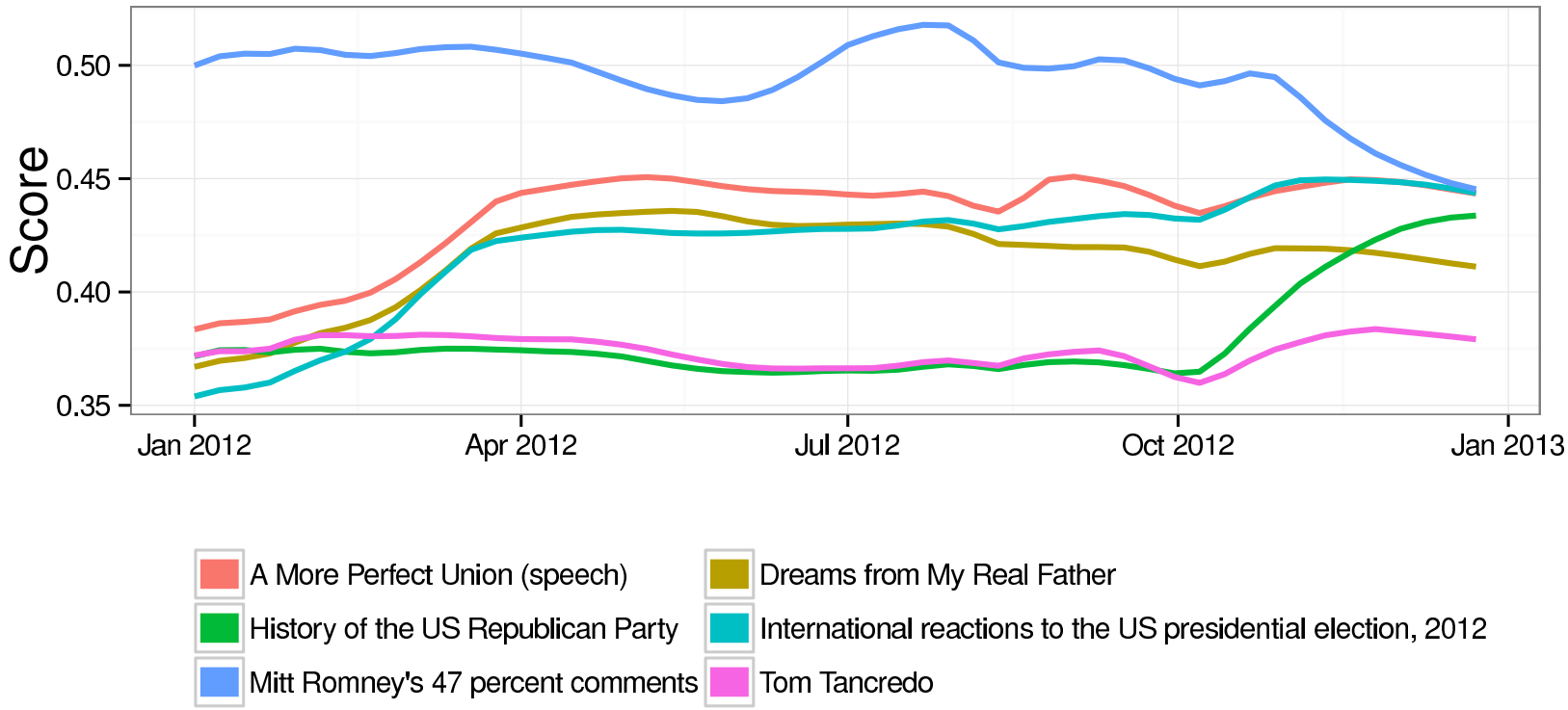
To measure the extent to which people are talking about Topic 10, we can use the document-specific weights on differennt topics. Summing the weights for Topic 10 over all posts in a given period of time shows the saliency of that topic at that time.



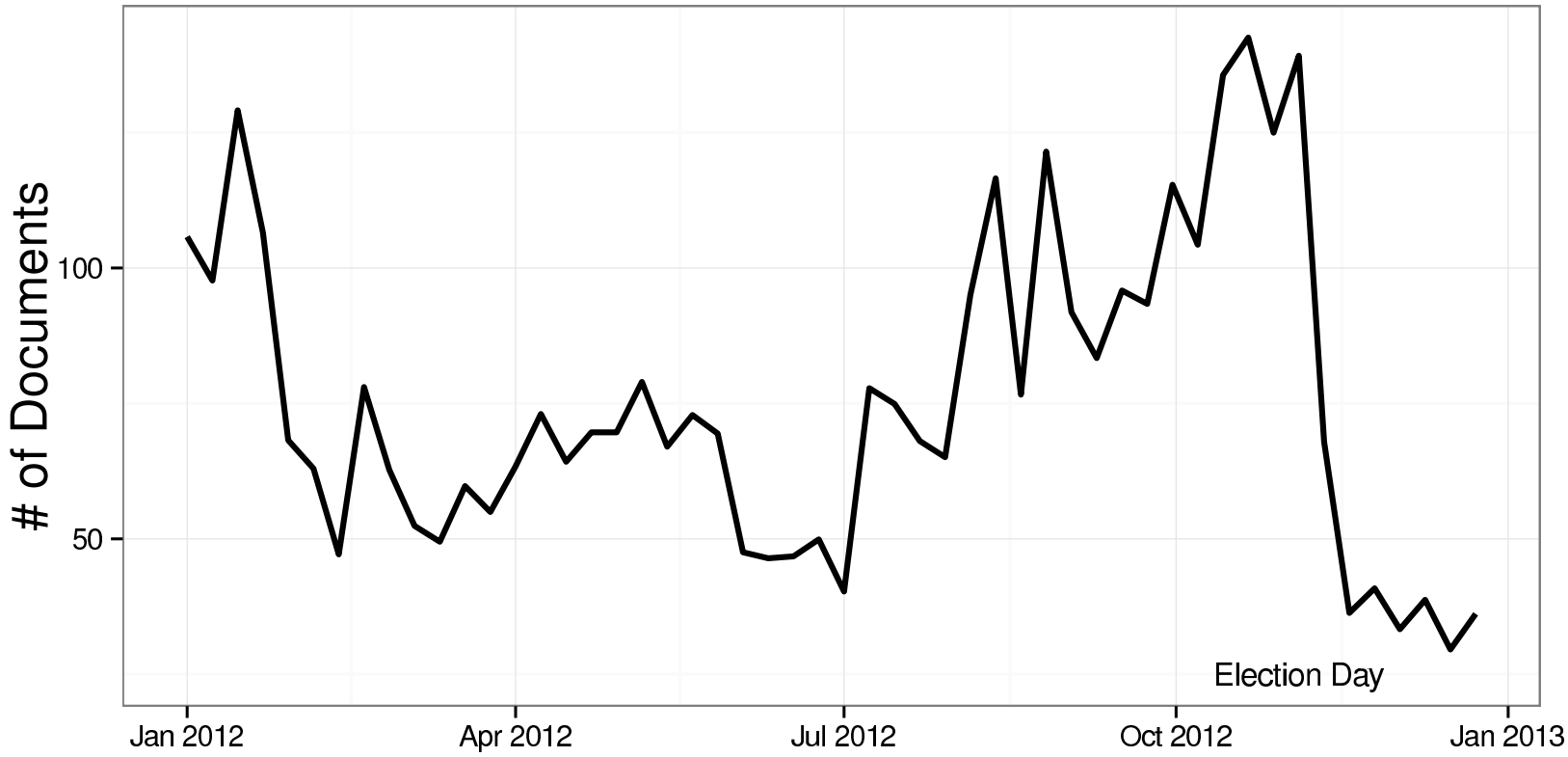
Topic 7 is clearly about the U.S. presidential election. Here are the high-probability tokens for that topic over different weeks in 2012.

Weeks	Tokens
2012-02-19	romney, santorum, obama, mitt, presid, rick, republican, gingrich, campaign, conserv, paul, newt, support, elect, debat, polit
2012-04-15	obama, romney, presid, campaign, mitt, republican, barack, elect, polit, bush, would, candid, democrat, conserv, santorum
2012-06-10	obama, presid, romney, campaign, mitt, barack, bush, said, american, republican, elect, clinton, polit, peopl, bain, polici, democrat
2012-08-05	romney, obama, ryan, presid, mitt, campaign, paul, republican, barack, elect, polit, would, america, busi, candid, bush, polici
2012-09-30	obama, romney, presid, debat, mitt, ryan, campaign, republican, barack, said, biden, elect, polici, polit, time, candid, last
2012-11-25	obama, presid, romney, elect, republican, barack, campaign, mitt, bush, democrat, year, polit, presidenti, polici, time, right, nation, first

The following figure shows how weights on Wikipedia labels change over time for Topic 7. It is still too smooth. The Wikipedia article on Mitt Romney's 47% comments seems to have captured many of the terms that were in play. The Tancredo article has an extensive discussion of his positions on many issues, and that seems to have matched the blogging vocabulary.



To measure the extent to which people are talking about Topic 7, we proceed as before and sum the document-specific weights on Topic 7.



5. Improving the Model

To handle topic dynamics, we need a mechanism for change the vocabulary weights.

At each time step, the current distribution on the vocabulary can change in four ways:

- the weight on a word can be unchanged,
- the weight on a word can decrease by 10% (e.g., to allow Newt Gingrich's candidacy to slowly fade from blog discussion over the course of the year),
- the weight on a word can increase by 10% (e.g., Mitt Romney became a more popular topic as the campaign progressed)
- a new word can be added, and take large probability (e.g., Benghazi first appeared on Sept. 11, 2012, and had relatively high probability).

Every word has equal and independent probability of undergoing one of these changes. Then the distribution is renormalized.

To model the network structure, we suppose that each blogsite is assigned to a single block, and that there are a Poisson number of blocks. Members of a block are interested in the same set of topics, and with one exception, no block is interested in more than three topics.

A block is interested in at most three topics, with one exception. We force one block to be interested in all topics, to account for such things as The Huffington Post and the New York Times blogsite.

A mixed-membership stochastic blockmodel controls how likely members of the same block are to link a post, how likely members in different blocks that share one or more topics are to link to each other, and how likely members in blocks that share no topics are to link.

Additionally, there is a logistic regression model that controls the probability that one blogger links to another. The regression uses the following covariates:

- same block?
- same topic?
- ever linked to this blogger before?
- prestige of linkee
- base rate of linking
- excitation

Many of the weights derive from a random effects model.

The excitation refers to whether or not the topic is hot. We use a Hawkes process model to describe this.

6. Conclusions

- Many alternate models are possible, and we continue to improve and refine our preprocessing and analysis.
- Community detection, the block structure, is of critical interest to identification of echo chambers of participants focused upon topics relevant to national security.
- Sentiment analysis is important to identify radicalization. And sentiment analysis is much more complex than the treatment described in this talk.
- For many applications, this approach needs to scale and to be fast. One wants the capacity to handle streaming data.
- Word2vec or LSI offer strategies for handling codewords and disguised discussion.