CrossMark

Modeling Community Structure and Topics in Dynamic Text Networks

Teague R. Henry[1] · David Banks[2] · Derek Owens-Oas[2] · Christine Chai[2]

Abstract

The last decade has seen great progress in both dynamic network modeling and topic modeling. This paper draws upon both areas to create a bespoke Bayesian model applied to a dataset consisting of the top 467 US political blogs in 2012, their posts over the year, and their links to one another. Our model allows dynamic topic discovery to inform the latent network model and the network structure to facilitate topic identification. Our results find complex community structure within this set of blogs, where community membership depends strongly upon the set of topics in which the blogger is interested. We examine the time varying nature of the Sensational Crime topic, as well as the network properties of the Election News topic, as notable and easily interpretable empirical examples.

Keywords Networks · Natural language processing · Topic modeling · Political blogs · Community detection

1 Introduction

Dynamic text networks have been widely studied in recent years, primarily because the Internet stores textual data in a way that allows links between different documents. Articles on the Wikipedia (Hoffman et al. 2010), citation networks in journal articles (Moody 2004), and linked blog posts (Latouche et al. 2011) are examples of dynamic text networks, or networks of documents that are generated over time. But each application has idiosyncratic features, such as the structure of the links and the nature of the time varying documents, so analysis typically requires bespoke models that directly address those aspects.

This article studies dynamic topic structure and the network properties of the top 467 US political blogs in 2012. Some key features of this data set are (1) topics, such as presidential election news, that evolve over time and (2) community structure among bloggers with similar interests. We develop a bespoke Bayesian model for the dynamic interaction

✉ Teague R. Henry
  trhenry@email.unc.edu

[1] University of North Carolina, Chapel Hill, NC 27599, USA

[2] Duke University, Durham, NC, USA

between text and network structure, and examine the dynamics of both the discourse and the community structure among the bloggers.

Our approach combines a *topic model* and a *network model*. A topic model infers the unobserved topic assignments of a set of documents (in this case, blog posts) from the text. And a network model infers communities among the nodes (in this case, blogs that tend to link to one another). In combination, we find blocks of blogs that tend to post on the same topics and which link with one another. These blocks, which we call *topic interest blocks*, allow one to examine sets of similar blogs, such as those that post only on the 2012 election or those that are only interested in both the Middle East and foreign policy. Topic interest blocks allow text content to guide community discovery and link patterns to guide topic learning.

We begin with a review of terminology in topic modeling. A *corpus* is a collection of *documents*. A document, in our case a post, is a collection of *tokens*, which consist of words and *n-grams*, which are sets of words that commonly appear together ("President of the United States" is a common 5-gram). In our application, a *blog* produces posts. A *topic* is a distribution over the tokens in the corpus. Typically, a post concerns a single topic. One such topic might be described as "the 2012 election," but this labeling is usually done subjectively, after estimation of the topic distributions, based on the high-probability tokens. For example, the 2012 election topic might put high probability on "Gingrich," "Santorum," "Cain," and "primaries."[1]

An early and influential topic model is Latent Dirichlet Allocation (LDA), proposed in Blei et al. (2003). It is a bag-of-words model, since the order of the tokens is ignored. LDA assumes that the tokens in a document are drawn at random from a topic. If a document is about more than one topic, then the tokens are drawn from multiple topics with topic proportions that must be estimated. The LDA generative model can produce a document that is 70% about the 2012 election topic and 30% about a Supreme Court topic by repeatedly tossing a coin with probability 0.7 of coming up heads. When it is heads, LDA draws a word from the 2012 election distribution; otherwise, it draws from the Supreme Court distribution. Markov chain Monte Carlo allows one to reverse the generative model, so that given a corpus of documents, one can estimate the distribution corresponding to each topic, and, for each document, the proportion of that document that is drawn from each topic.

In our application, topic specific word probabilities evolve over time—the token "Gingrich" is more probable early in 2012 than later, when he dropped out. Blei and Lafferty (2006) develops a method that allows for topic drift, so the probability of a token in a topic can change (slowly) through an auto-regressive process. But blog data requires the possibility of rapid change; "Benghazi" did not occur in the corpus before September 11, but thereafter was a high-probability token. We develop a dynamic version of a topic model described in Yin and Wang (2014). The way we infer topics allows for both slow drift and the sudden appearance of new words, and even new topics, over the course of the year.

There is a second source of information in the blog data that previous dynamic topic models cannot utilize. It is the links between blogs, which prompt a network model. Here a blog is a *node*, and a hyperlink between blogs is an *edge*. We use an exponential random graph model (Holland and Leinhardt 1981; Wasserman and Pattison 1996) to estimate the probability of an edge through a logistic regression on predictors that include node characteristics and other explanatory variables. This framework can be combined with clustering

---

[1]Gingrich, Santorum, and Cain all refer to candidates in the 2012 Republican presidential primary.

methods to perform *community detection*, where a community is a set of nodes that are more likely to create edges among themselves than with other nodes.

There are a number of recent methods for community detection. One is a family of algorithms that use modularity optimization (Newman and Girvan 2004). But the corresponding models are not parametric and do not support Bayesian inference. A popular alternative is the latent space model of Hoff et al. (2002). It estimates nodel locations in an unobserved space which then defines the community structure; but it is too computationally demanding for the large blog posts data set.

We prefer the *stochastic block model* of Snijders and Nowicki (1997). Stochastic block models place nodes into latent communities based on the observed pattern of links between nodes, which are modeled using independent Bernoulli random variables. It has been extended as the mixed membership block model (Airoldi et al. 2008), which allows nodes to be members of more than one community. In that spirit, the model developed in this paper keeps the stochastic block modeling framework, but permits nodes to have idiosyncrasies in their connection patterns that are not solely due to community membership, but also reflect node covariates (in this application, the degree of the blogs' interests in specific topics). Shared community membership increases edge formation probability, and nodes in different communities that have shared topic interests also have elevated probabilities of linking. A stochastic block model can be easily expressed within an exponential random graph modeling framework.

Combining topic information and linkage information through the topic interest blocks is our key methodological contribution in this article. A secondary contribution is extending the topic model of Yin and Wang (2014) into a dynamic topic model. Researchers have started to develop models that combine network analysis and topic analysis, mostly in the context of static networks. Chang and Blei (2009) describes a relational topic model in which the probability of links between documents depends upon their topics and applies it to two datasets of abstracts and a set of webpages from computer science departments. Ho et al. (2012) applies such methods to linked hypertext and citation networks. Wang et al. (2011) develops a model for the case in which there are noisy links between nodes, in the sense that there are links between documents whose topics are not related. Yin and Wang (2014) does related work on clustering documents through use of a topic model. However, none of these methods allow for the simultaneous modeling of dynamic topics with a community structure on the nodes.

Our model uses text and covariate information on each node to group blogs into blocks more likely to post on the same topics and link to one another. This approach expands upon community detection, but also fundamentally alters how communities are defined. We assume that if two blogs are interested in the same topics, then they are more likely to link to each other and form a community. Estimating the extent to which blogs post about the same topics helps explain community structure, above and beyond the community structure described by linkage pattern. Furthermore, integrating community detection into topic models allows the linkages to inform the allocation of topics, connecting network structure to topic structure. So inference on topic distributions is supplemented by non-text information. This results in communities that are defined both on the pattern of links (traditional community detection), as well as textual data. One consequence of this approach is that communities are more grounded in the substantive reason for any community structure, shared interest in various topics.

In particular, for the 2012 blog application, we wanted a bespoke Bayesian model that (1) allows topic distributions to change over time, both slowly and quickly, (2) classifies blogs into blocks that share topic interests and have elevated internal linkage probabilities, and

(3) theoretically enable use of covariate information on blogs. This includes prestige, sociability, whether recently linked, and more. Here, we extend covariates from static stochastic block models, as in Faust and Wasserman (1992), to dynamic networks. Some covariates are fixed (e.g., topic interests) whereas others are time-varying (e.g., whether recently linked).

Section 2 describes our dataset and its preparation. Section 3 gives a generative dynamic model for both the text and the network. Section 4 specifies the Bayesian prior and posterior inference algorithm used to estimate model parameters. Finally, in Section 5, we present several findings from the political blog data, and Section 6 finishes with a discussion of possible generalizations.

2 Political Blogs of 2012

Our data consists of the blog posts from the top 467 US political blogs for the year 2012, as ranked by Technorati (2002). This dataset has a dynamic network structure since blog posts often link to each other, responding to each other's content. Additionally, the topic structure of the blog posts reflect different interests, such as the presidential campaign or sensational crime. The token usage in each topic changes over time, sometimes quite suddenly, as with the appearance of the tokens "Trayvon" and "Zimmerman"[2] in March, 2012, and sometimes more gradually, as with the slow fade of the token "Gingrich"[3] during the spring. Over the 366 days in 2012, a leap year, the political blogs accumulated 109,055 posts.

2.1 Data Preparation

Our data were obtained through a collaboration with MaxPoint Interactive, now Valassis Digital, a company headquartered in the Research Triangle that specializes in computational advertising. Using the list of 467 U.S. political blog sites curated by `Technorati`, computer scientists at MaxPoint scraped all the text and links at those sites (after declaring robot status and following all robot protocols).

The scraped text was stemmed, using a modified version of Snowball (McNamee and Mayfield 2003) developed in-house at MaxPoint Interactive. The initial application removed all three-letter words, which was undesirable, since such acronyms as DOT, EPA, and NSA are important. That problem was fixed and the data were restemmed.

The second step was filtering. This filtering was based on the variance of the unweighted term-frequency, inverse document frequency (TF-IDF) scores (Ramos 2003). The TF-IDF score for token w in blog post d is

$$\text{TF-IDF}_{wd} = f_{wd}/n_w \tag{1}$$

where f_{wd} is the number of times that token w occurs in blog post d, and n_w is the number of posts in the corpus that use token w. Words that have low variance TF-IDF scores are such words as "therefore" and "because," which are common in all posts. High-variance scores are informative words that are used often in a small number of posts, but rarely in other posts, such as "homosexual" or "Zimmerman." Interestingly, "Obama" is a low-variance TF-IDF token, since it arises in nearly all political blog posts.

---

[2]George Zimmerman shot and killed Trayvon Martin in March of 2012.

[3]Newt Gingrich gradually faded to political irrelevance after a failed presidential primary run.

Next, we removed tokens that were mentioned in less than 0.02% of the posts. This reduced the number of unique tokens that appeared in the corpus, as these were unlikely to be helpful in determining the topic token distribution across all posts. Many of these were misspellings, e.g., "Merkle" for "Merkel", the Chancellor of Germany. Overall, these misspellings were either rare (as in the case of "Merkle-Merkel"), or incomprehensible.

After all tokens were filtered, we computed the n-grams, starting with bigrams. A bigram is a pair of words that appear together more often than chance, and thus correspond to a meaningful phrase. For example, the words "white" and "house" appear in the blog corpus often, in many different contexts (e.g., race relations and the House of Representatives). But the phrase "White House" refers to the official residence of the president and appears more often than one would predict under an independence model for which the expected number of phrase occurrences is $Np_{\text{white}}p_{\text{house}}$, where N is the total amount of text in the corpus and p_{white} and p_{house} are the proportions of the text that are stemmed to "white" and "house." Bigrams were rejected if their significance probability was greater than 0.05. In examining the bigram set generated from this procedure, it appeared to be too liberal; English usage includes many phrases, and about 70% of tested bigrams were retained. Therefore, we excluded all bigrams occurring less than 500 times corpus-wide. This significantly reduced the set of bigrams.

After the bigrams were computed and the text reformatted to combine them, the bigramming procedure was repeated. This produced a set of candidate trigrams (consisting of a previously identified bigram and a unigram), as well as a set of candidate quadrigrams (made up of two previously accepted bigrams). These candidates were retained only if they had a frequency greater than 100. This cutoff removed the majority of the candidate trigrams and quadrigrams. The final vocabulary consisted of 7987 tokens.

It is possible to go further, finding longer n-grams, but we did not. However, we identified and removed some long n-gram pathologies, such as the one created by a blogger who finished every post by quoting the Second Amendment. There is a large literature on various n-gramming strategies (Brown et al. 1992). Our work did not employ sophisticated methods, such as those that use information about parts of speech. After this preprocessing complete, we had the following kinds of information:

- Stemmed, tokenized, reduced text for each post, the date on which the post was published, the blog the post it was published on, and links to other blogs in the network.
- Blog information, including the web domain, an estimate of its prestige from `Technorati`, and sometimes information on political affiliation.

From this information, we want to estimate the following:

- Time evolving distributions over the tokens, where the time evolution on a token may be abrupt or gradual.
- The topic of each post—our model assumes that a post is about a single topic, which is usually but not always the case (based upon preliminary work with a more complicated model).
- The topic interest blocks, which are sets of blogs that tend to link among themselves and which tend to discuss the same topic(s).
- The specific topics of interest to each of the topic interest blocks.
- The linking probabilities for each pair of blogs, as a function of topic interest block membership and other covariates.
- Posting rates, as a function of blog covariates and external news events that drive discussion.

We now describe the generative model that connects dynamic topic models with network models in a way that accounts for the unique features of this data set.

3 Model

The generative model can be described in two main phases: initialization of static quantities, such as blogs' topic interest block membership, and generation of dynamic quantities, specifically posts and links. First, the model creates k topic distributions that are allowed to change over time. Next, it generates time-stamped news events for each topic. Each blog is randomly assigned to a topic interest block. With these elements in place, post and link generation proceeds. For each blog, on each day, the number of posts from each topic is generated, in accordance to the topic interest block of that blog. The content of the post is generated from the day-specific topic distribution, and links are generated so as to take account the blog's topic interest block. We now describe each step in the generative model in more detail.

3.1 Topic and Token Generation

We begin with the topic distributions, which must allow dynamic change. For the kth topic, on a specified day t, we assume the token probabilities $\mathbf{V}_{kt}$ are drawn from a Dirichlet distribution prior. This set of topic-specific token probabilities is the *topic distribution* on day t. To encourage continuity across days, we calculate the average of topic k's topic distribution $\mathbf{V}_{k(t-1):(t-\ell)}$ from the previous ℓ days and use it as the concentration parameter for the Dirichlet distribution from which the present day's topic $\mathbf{V}_{kt}$ is drawn. The sampling proceeds in sequence, first calculating each topic's concentration parameter as in Eq. 2 and then sampling each topic as in Eq. 3 and then moving to the next day. This procedure repeats for times $t = 1 : T$. The topics are then distributed:

$$\boldsymbol{a}_{kt} = \frac{1}{\ell} \sum_{t'=1}^{\ell} \mathbf{V}_{k(t-t')}, \tag{2}$$

$$\mathbf{V}_{kt} \sim Dir_{|W|}(\mathbf{a}_{kt}). \tag{3}$$

3.1.1 Topic Event Generation

To capture the event-driven aspect of blog posting, we generate events which then boost the post rate on the corresponding topic. For each topic k, at each time t, there is some probability η_k of an event occurring. One can choose $\eta_k = .01$ for all k, which suggests each topic has on average 1 event every 100 days. Alternatively, different topics can be given different daily event probabilities or one can put a prior on η_k. Given η_k, the daily, topic-specific event indicators are sampled as

$$E_{kt} \sim \text{Bern}(\eta_k). \tag{4}$$

When an event happens on topic k, blogs with interest in topic k have their posting rates increase by a factor determined by ψ_k. Speculating that some topics have events which are much more influential than others, we let this multiplier be topic specific:

$$\psi_k \sim \text{Gam}(a_\psi, b_\psi). \tag{5}$$

3.1.2 Block- and Blog-Specific Topic Interest Specification

With our topic distributions and topic specific events generated, we can now assign blogs to topic interest blocks. We begin by defining the block-specific topic-interests matrix $\mathbf{I}$, where each column b indicates which of the k topics are of interest to block b. The first $\binom{K}{1}$ columns correspond to the singleton blocks, which are interested only in topic 1, topic 2, up through topic K, respectively. The next $\binom{K}{2}$ columns define doublet blocks, which have interest in all of the possible topic pairs. The next $\binom{K}{3}$ columns correspond to blocks which have interest in exactly three topics, and the final column is for the block which has interest in all K topics:

$$
I_{kb} = \begin{bmatrix}
1 & 0 & 0 & \dots & 0 & 1 & 1 & 1 & \dots & 0 & 1 & 1 & 1 & \dots & 0 & 1 \\
0 & 1 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 & 1 & 1 & 1 & \dots & 0 & 1 \\
0 & 0 & 1 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 & 1 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 1 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 1 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 \\
\vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 1 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 1 & 1 \\
0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 1 & 1
\end{bmatrix}
\tag{6}
$$

To assign blogs to blocks, we sample their membership with a single draw from a multinomial distribution. This means each blog is a member of only a single block, characterized by the topic interests in the above matrix. Each block assignment is then drawn from a multinomial distribution:

$$
b_i \sim \text{Mult}(1, \mathbf{p}_B).
\tag{7}
$$

One can choose the probabilities of belonging to each block uniformly, by setting each element of $p_b = 1/B$ where $B = \binom{K}{1} + \binom{K}{2} + \binom{K}{3} + 1$ gives the total number of blocks. Another approach is to partition the probabilities vector into the singlets, doublets, triplets, and all-topics blocks, and allocate probability uniformly to each of these categories, and then uniformly divide up the probability among blocks within each category:

$$
\mathbf{p}_B = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ p_K \end{pmatrix}, \text{ with } \mathbf{p}_1 = \begin{pmatrix} p_{1,1} \\ p_{1,2} \\ \vdots \\ p_{1,\binom{K}{1}} \end{pmatrix}, \mathbf{p}_2 = \begin{pmatrix} p_{2,1} \\ p_{2,2} \\ \vdots \\ p_{2,\binom{K}{2}} \end{pmatrix}, \mathbf{p}_3 = \begin{pmatrix} p_{3,1} \\ p_{3,2} \\ \vdots \\ p_{3,\binom{K}{3}} \end{pmatrix}, p_K = p_{K,\binom{K}{K}}.
\tag{8}
$$

For notational convenience throughout the rest of the paper, we define $\mathbf{B}_i$ to be the set of topics which are of interest to blog i:

$$
\mathbf{B}_i = \{k : I_{kb_i} = 1\}.
\tag{9}
$$

With each blog's topic interest indicators known, we can generate blog-specific topic-interest proportions. For example, two blogs may be in the block with interest in topic 1 and

topic 2, but one may have interest proportions (.9, .1) while the other has (.5, .5). As is conventional in topic modeling, topic (interest) proportions are drawn from a Dirichlet distribution, though we make the distinction that each blog has a specific set of hyper-parameters $\boldsymbol{\alpha}_i$. An individual topic interest vector $\boldsymbol{\pi}_i$ is then a draw from a Dirichlet distribution:

$$\boldsymbol{\pi}_i \sim Dir_K(\boldsymbol{\alpha}_i). \tag{10}$$

The hyperparameters are chosen such that a blog with interest in topics 1 and 2 is likely to have most of its interest in those topics, though it allows for interest in other topics to occur with small probabilities:

$$\boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{iK} \end{pmatrix}, \text{ with } \alpha_{ik} = P1(k \in \mathbf{B}_i) + 1(k \notin \mathbf{B}_i). \tag{11}$$

3.1.3 Post Generation

Given the blogs' topic interest and block membership, along with the event distribution, we can now generate the number of posts a blog produces on a particular topic. Each blog may post on multiple topics, but each post is associated with a single topic. Every blog has a baseline posting rate which characterizes how active it generally is on days without events. For blog i, the baseline post rate ρ_i is sampled from the following distribution:

$$\rho_i \sim \text{Gam}(a_\rho, b_\rho). \tag{12}$$

With the blog-specific baseline post rate ρ_i, the blog-specific topic interest proportions π_{ik}, the topic-specific daily event indicators E_{kt}, and topic-specific post rate multipliers ψ_k accounted for, we construct the expected post rate for each topic, on each blog, each day:

$$\lambda_{tki} = \rho_i \pi_{ik} + \rho_i E_{tk} \psi_k. \tag{13}$$

Given this post rate, the count D_{tki} of posts about topic k, on blog i, on day t are generated:

$$D_{tki} \sim \text{Pois}(\lambda_{tki}). \tag{14}$$

In the observed data, we do not know the post counts D_{tki} on each topic, but instead we know the marginal counts D_{ti}. These are referenced throughout the inference procedure described in Section 4 and are calculated:

$$D_{ti} = \sum_{k=1}^{K} D_{tki}. \tag{15}$$

With daily topic-specific post counts and token probabilities available, the posts can be populated with tokens. We first sample a total number of tokens for each post. In particular, on day t, the token count W_{tkid} for post d about topic k on blog i is sampled:

$$W_{tkid} \sim \text{Pois}(\lambda_D). \tag{16}$$

where λ_D is the average number of tokens over all posts. The W_{tkid} tokens can then be sampled from the appropriate day and topic specific multinomial distribution with probability vector $\mathbf{V}_{kt}$. This is done for all of the posts in the corpus like so:

$$N_{tkid}^w \sim Mult(W_{tkid}, \mathbf{V}_{kt}). \tag{17}$$

3.1.4 Network Generation

Finally, we generate the network of links between blogs. Rather than modeling link generation at a post level, we model it at a daily blog to blog level. Specifically, we model a directed adjacency matrix A_t of links, with entry $a_{ii't}$ indicating whether any posts from blog i have links to blog i' on day t. The binary logistic regression is suitable for this scenario. We assume the link probability $p_{ii't} = p(A_{ii't} = 1)$ depends on the following factors:

- $B(i, i') = 1(b_i = b_{i'}) + \boldsymbol{\pi}_i^T \boldsymbol{\pi}_{i'} 1(b_i \neq b_{i'})$ is the similarity (in topic interests) for nodes i and i', and is in the interval [0,1], taking value 1 if and only if blogs i and i' are in the same block.
- $L_{i'it} = 1((\sum_{t'=t-7}^{t-1} a_{i'it'}) > 0)$ indicates if blog i' has linked to blog i within the last week (previous to the current time t).
- $I_{i't} = \frac{1}{t-1} \sum_{t'=1}^{t-1} \sum_i a_{ii't'}$ is the average indegree (through time $t-1$) of the receiving node i'.
- $O_{it} = \frac{1}{t-1} \sum_{t'=1}^{t-1} \sum_{i'} a_{ii't'}$ is the average outdegree (through time $t-1$) of the sending node i.

The first covariate is sampled and constructed in Eqs. 6, 7, 8, 9, 10, and 11, and the other three covariates are defined and calculated as statistics of the past data $\{A_{t'}\}_{t'=1}^{t-1}$. Together with an intercept, they comprise the regressors in a logistic regression for links, which can be written as in Eq. 18,

$$\log\left(\frac{p_{ii't}}{1 - p_{ii't}}\right) = \theta_0 + \theta_1 B(i, i') + \theta_2 L_{i'it} + \theta_3 I_{i't} + \theta_4 O_{it}. \tag{18}$$

We specify a normal prior for the intercept and the regression coefficients:

$$\theta_p \sim \text{Norm}(\mu_\theta, \sigma_\theta^2). \tag{19}$$

We can use the logistic function to write the probability of a link as

$$p_{ii't} = p(A_{ii't} = 1) = \frac{\exp(\boldsymbol{\theta}^T \mathbf{S}_{ii't})}{1 + \exp(\boldsymbol{\theta}^T \mathbf{S}_{ii't})}, \tag{20}$$

with coefficients and covariates written as

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)^T \text{ and } \mathbf{S}_{ii't} = (1, B(i, i'), L_{i'it}, I_{i't}, O_{it})^T. \tag{21}$$

This form makes it more clear how the model can be cast within the ERGM framework (Holland and Leinhardt 1981). Covariates are time dependent as in TERGM literature (Krivitsky and Handcock 2014). An important note is that covariates depend only on past linking data, which makes this a predictive model of links. Finally, we sample each link as a single Bernoulli trial with the appropriate probability as defined in Eq. 20:

$$A_{ii't} \sim \text{Bern}(p_{ii't}). \tag{22}$$

This generative model for the links can be thought of as a variant of stochastic block modeling (Snijders and Nowicki 1997), where block membership is "fuzzy." In our model, while members of the same block will have the highest probability of linking with other members of the same block, individuals who share similar topic interests, but do are not in the same block are more likely to link than individuals who share no topic interests. This allows for a pattern of linkages that more accurately reflect the empirical phenomena of topic based blog hyperlinks.

At the end of data generation, we have $\{\mathbf{B}_i\}_{i=1}^{I}$ and $\boldsymbol{\pi}_i$ giving the topic interest set and topic interest proportions, respectively, for blog i; $K \times T$ matrix $\mathbf{E}$ with daily topic-specific event indicators; $K \times I \times T$ array $\mathbf{D}$ with entry D_{kit} giving the number of posts about topic k on blog i at time t; $|W| \times K \times T$ array $\mathbf{V}$ of daily topic specific token probabilities; and multidimensional object $\mathbf{N}$ containing the count N_{tkid}^{w} for each token w in the dth post about topic k on blog i at time t.

With a theoretically justified data generating mechanism in place, we proceed to Section 4 to "invert the generative model" and derive posterior inference for the parameters of interest.

4 Estimation

As our dataset of blog posts consists of posts, time stamps, which blog posted each post, and each post's links to other blogs, our inferential model needs to estimate a number of quantities. This section gives the details of how we estimate quantities of interest from the data and how we specified our priors. The notation is dense, so the following guide is helpful.

- *Each post's topic assignment*. We observe the content of each post, but do not know the topic assignment of each post. This must be inferred. We denote this estimate as z_d for post d.
- *Topic distributions*. We do not know what the content of each topic is, or how each topic changes over time. We use $\mathbf{V}_t$ for the topic token-distributions matrix, and for specific topics, we denote this as $\mathbf{V}_{kt}$.
- *Events and Post Rate Boosts*. Events are not observed and must be inferred. This $T \times K$ matrix is $\mathbf{E}$. The event-caused, topic-specific boosts in post rate ψ_k are also inferred.
- *Blog-specific parameters*. A blog's average post rate and topic interests must be inferred. The blog average post rate is denoted ρ_i, and the topic interest proportions is a vector of length K, denoted $\boldsymbol{\pi}_i$.
- *Blogs' block membership*. A Blog's block membership is inferred using the linkage pattern and topic assignments of each of the blog's posts. The ith blog's block membership is denoted as b_i and its corresponding topic interests indicator vector is $\mathbf{B}_i$.
- *Network parameters*. The network parameters govern the probability of linkage. These are depend upon block membership, lagged reciprocity, indegree and outdegree through a logistic regression whose coefficients must be estimated. These five network parameters (including intercept) are denoted θ_0, θ_1, θ_2, θ_3, and θ_4, respectively.

4.1 Hyper Parameter Specification

The model requires that several parameters be specified a priori. In this subsection, we describe these hyper parameters in general terms, while in Section 5.1, we show which specific values we used to analyze the political blog data. The first hyper parameter is K, the total number of topics. In principle, one could place an informative prior on the number of topics and use the posterior mean determined by the data. This is, however, computationally cumbersome and so we make the decision to specify the number of topics in advance. This approach is used in Blei and Lafferty (2006) and Blei et al. (2003) for the Latent Dirichlet Allocation models . One can use penalized likelihood as a selection criterion, as described

in Yin and Wang (2014), or an entropy-based criterion, such as the one described in Arun et al. (2010). We chose the number of topics by running models with different values of K and selecting the number of topics using the entropy-based criterion of Arun et al. (2010).

The time lag ℓ for topic dependency needs to be specified. This time lag determines the scale of the topics and has units in number of days. This determines how long tokens remain in a topic, and can be conceptualized as a smoother over the time changing vocabulary. Smaller values of ℓ will produce more varied topic distributions over time, while larger values will reflect slower shifts in the topic content. For the node-specific parameters, only P, the Dirichlet concentration parameter on the topics which are of interests to a block, is needed. This parameter governs how often blogs are allowed to post outside of topics they are interested in, with lower values allowing for more out of interest posting, and higher values corresponding to restricted topic interests. Finally, for any reasonable number of topics, a restriction on the block structure is required to ensure computational feasibility. For an unrestricted block structure with K topics, the total number of possible blocks that must be evaluated is $\sum_{i=1}^{K} \binom{K}{i}$, which is computationally intractable for moderate K. In this paper, we restrict blocks to have 1, 2, or 3 topic interests, and allow one block to have interest in all topics. Finally, we specify the expected number of non-zero topic interest blocks using the prior λ_B.

4.2 A Simple Data Augmentation

While the generative model assumes Poisson distribution on post counts D_{kit}, we rely on a data augmentation for the inference procedure. Because counts D_{it} of posts on each blog each day are already known, we augment the generative model with latent variables $\{z_{d_{it}}\}_{d_{it}=1}^{D_{it}}$ which instead tell the latent topic assignment of post d_{it}. We can then re-write the Poisson likelihood $\prod_{k=1}^{K} \text{Pois}(D_{kit}|\lambda_{kit})$ as a multinomial likelihood $\prod_{d_{it}=1}^{D_{it}} \text{Mult}(z_{d_{it}}|1, \boldsymbol{\xi}_{it})$ with $\xi_{kit} = \frac{\lambda_{kit}}{\sum_{k=1}^{K} \lambda_{kit}}$. This reformulation enables use of the topic assignment inference algorithm from GSDMM.

4.3 Metropolis Within Gibbs Sampling

We use a Metropolis within Gibbs sampling algorithm (Gilks et al. 1995) to obtain posterior distributions for the parameters defined in the generative model. This approach consists of four stages:

1. Each day t, for each blog i, sample a topic assignment $z_{d_{it}}$ for each post d_{it} and update the matrix of daily topic specific token-distributions $\mathbf{V}_t$.
2. For blogs, update topic interest proportions (π_{ik}), and base rate for posting (ρ_i). For events, update the event matrix $\mathbf{E}$, and activation level parameters (ψ_k).
3. Update the network parameters, i.e., θ_0, θ_1, θ_2, θ_3, and θ_4.
4. Update each blog's block assignment b_i and corresponding topic interest indicators $\mathbf{B}_i$.

4.4 Topic Modeling and Post Assignment

Both posts' topic assignments as well as the topic distributions themselves are unobserved and must be inferred. A preferred algorithm for inferring post topic assignment and topic token-distributions would first assign each post a single topic, and second have some flexibility in collapsing topic distributions together.

To those ends, we adapt the Gibbs Sampler for the Dirichlet Mixture Model (GSDMM) of Yin and Wang (2014). As originally proposed, the GSDMM classifies a set of documents into specific topics. The tokens of a post are assumed to be generated from the topic-specific multinomial which that post was assigned, and many tokens may be instantiated in multiple topics (e.g., common words such as "therefore"). The assignment of each post to a *single topic* differs from the Latent Dirichlet Allocation model, which models documents as mixtures across a number of topics. GSDMM estimates the probability that a document d is about topic k, given the current topic vocabulary distribution, as

$$P(z_d = k | \mathbf{V}_k, d) = \frac{m_{k,-d} + \alpha}{|D| - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (N_{k,-d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{k,-d} + |W|\beta + i - 1)}, \quad (23)$$

where $m_{k,-d}$ is the number of posts currently assigned to topic k (not including the topic assignment of post d), N_d^w is the number of occurrences in post d of token w, and $N_{k,-d}^w$ is the number of occurrences of token w in topic k (not including the content of post d). The α controls the prior probability that a post is assigned to a topic; increasing α implies that all topics grow equally likely. The β relates to the prior probability that a token will have relevance to any specific topic; increasing β results in fewer topics being found by the sampler. Finally, $|D|$ is the number of posts in total, and $|W|$ is the size of the vocabulary.

As originally proposed by Yin and Wang (2014), GSDMM is a static model. We modify it by allowing $\mathbf{V}$ to vary over time. For readability, we suppress the subscripts and denote the specific post d_{it} by d. We define

$$m_{k,t,-d}^* = \left(\sum_{t'=t-\ell}^{t} D_{t'k} \right) - 1 \text{ with } D_{tk} = \sum_i D_{tki}, \quad (24)$$

to be the number of posts assigned to topic k in the interval from $t - \ell$ to t, not including post d by blog i at time t. Also we let

$$N_{k,t,-d}^{*w} = \sum_{t'=t-\ell}^{t} N_{k,t'}^w, \quad (25)$$

be the number of times that token w occurs in topic k in the interval from $t - \ell$ to t, not including post d. This defines a sliding window that allows the sampler to use information from the recent past to infer the topic to which a post belongs, while allowing new tokens to influence the assignment of the post at the current time point. The probability of assigning post d to topic k is then:

$$\mathbb{P}[z_d = k | \mathbf{V}_{kt}, d] = \frac{m_{k,t,-d}^* + \alpha}{|D_{t-\ell:t}| - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{s=1}^{N_d^w} (N_{k,t,-d}^{*w} + \beta + s - 1)}{\prod_{i=1}^{N_d} (N_{k,t,-d}^* + |W|\beta + i - 1)}, \quad (26)$$

where $|D_{t-\ell:t}|$ is the number of posts within the lag window. Note that Eq. 26 does not use information about the blog that generates the post. So the final step is to incorporate the tendency of blog i to post on topic k at time t, using the Poisson rate parameter in Eq. 13. Using the normalized point-wise product of conditional probabilities, the final expression for the probability that post d (i.e., d_{it}) belongs to topic k is

$$\mathbb{P}[Z_d = k | \mathbf{V}_{k,t}, d, \lambda_{ikt}] = \frac{\mathbb{P}[Z_d = k | \mathbf{V}_{k,t}, d]\mathbb{P}[Z_d = k | \lambda_{ikt}]}{\sum_{q=1}^{K} \mathbb{P}[Z_d = q | \mathbf{V}_{q,t}, d]\mathbb{P}[Z_d = q | \lambda_{iqt}]}. \quad (27)$$

To reduce computation, we approximate $\mathbb{P}[Z_d = k \mid \lambda_{ikt}]$ with $\lambda_{ikt} / \sum_{j=1}^{K} \lambda_{ijt}$, as clarified in the data augmentation above.

The topic assignment of a post can now be Gibbs sampled using Eq. 27. The sampler assigns the first post to a single topic, updates the topic-token distributions with the content of that post, then continues to the next post, and repeats. At each time point, the sampler sweeps through the set of posts several times so that the topic assignments can stabilize. The preferred number of sweeps depends on the complexity of the posts on that day, but it need not be large. After some exploration, this study used 10 sweeps at each time point per iteration.

In summary, after topic assignment has been completed for a given day t, all posts within that day will have a single topic assignment. Moving to day $t+1$, the topic assignment for all posts that day will utilize the information from day t through that day's topic-specific, token-distribution estimator specified in Eq. 25. Once the topic assignment estimator reaches the final day T, all posts will have assigned topics, and all topics will have a time varying token-distribution. The post specific topic assignments are then used in the next step of estimating blogs' topic interest vectors.

4.5 Node Specific Parameters and Event Parameters

Once posts are assigned to topics, the next step is to update the node specific parameters, specifically the blog topic interest vector π_i and the blog posting rate ρ_i.

The topic interest vector is updated in a Metropolis-Hastings step. As is standard in M-H, we specify a proposal distribution, a likelihood, and a prior. The proposal π_i^* is a draw from a Dirichlet distribution with $\alpha_i = \pi_i D_i$, where D_i is the total number of posts generated by node i. The likelihood is

$$\prod_{t=1}^{T} \prod_{k=1}^{K} \mathbb{P}(D_{kit} | \lambda_{kit}), \tag{28}$$

where $\mathbb{P}(D_{kit} | \lambda_{kit})$ is the Poisson likelihood from Eq. 14 representing the day-specific number of posts on blog i assigned to each topic. Note the dependence of λ_{kit} on π_i comes through (13). A hierarchical prior is used, which is Dirichlet(α_{B_i}), where the parameters are defined by the current block assignment of node i, as in Eq. 11. This step requires estimates of each blog's block assignment, which will be described later.

The ith blog's posting rate ρ_i is also updated using a Metropolis-Hastings step, where the proposal distribution is a Normal truncated at 0, with mean equal to ρ_i and standard deviation equal to σ_ρ^2. The likelihood evaluated is the same as in Eq. 28. The prior is a univariate normal truncated at 0 and with mean ρ and variance σ_ρ^2. Truncated normal distributions are used to uncouple the mean and the variance.

Next to update are the event matrix $\mathbf{E}$ and activation boost parameters ψ_k. The event matrix is updated with a series of Metropolis steps for each time point and topic. The proposal is simply 1 if $E_{k,t} = 1$ and 0 if $E_{k,t} = 0$. The likelihood is the same as Eq. 28, except, for each topic k at each time t, the product of Poisson densities is over the blogs $i = 1 : I$. The prior is simply a Bernoulli with parameter E_π.

Each activation parameter ψ_k is updated with a Metropolis-Hastings step, where the proposal ψ_k^* is a truncated normal at 0 with mean ψ_k and standard deviation σ_ψ. Again, the likelihood is similar to Eq. 28, in that it is the product indexed over time of the Poisson densities for every blogs' number of posts in topic k. Unlike the original likelihood, however, the second product is over blogs $i = 1 : I$. The prior distribution on ψ_k is a normal truncated at 0 with mean ψ and standard deviation σ_ψ^*.

4.6 Network Parameters

The network parameter set consists of the vector $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_4,)$ as defined in Eq. 18. Each network parameter can be sampled using a Metropolis within Gibbs step. Specifically, the Bernoulli likelihood portion (equivalently a logistic regression) can be expressed as

$$\prod_{i \neq j} \prod_{t \in \{1,\ldots,T\}} \frac{\exp((\theta_0 + \theta_1 B(i, j) + \theta_2 L_{jt} + \theta_3 I_j + \theta_4 O_i))^{A_{ijt}}}{1 + \exp(\theta_0 + \theta_1 B(i, j) + \theta_2 L_{jt} + \theta_3 I_j + \theta_4 O_i)}. \tag{29}$$

To update each parameter, one conditions on all other pieces of information in the model. Proposals are normal with mean set to the current value of the parameter, and a standard deviation specific to the parameter, while the priors are normal with a given mean and standard deviation. Note here that this sampling relies on estimates of the block membership of each blog.

4.7 Block Assignment

Previously, we described the sampling routine for post-topic assignment, topic token-distributions, node-specific parameters, event matrix and boosts, and network parameters. The blog-specific block memberships remain to be estimated. Due to the complexity of their dependence on many other components of the model, we describe block estimation last.

Recall that a blog's block describes two things. The first is the set of topics the blog will be most likely to post on. The second is a blog will tend to link more to other blogs that are within the same block as it is. Therefore, block assignment for a given node i is informed by several pieces of data. These are its topic interests, the network parameters, other blog's block memberships, and the observed links. One assumption of our model is that a node's network position and topic interest are conditionally independent given block assignment, which in turn makes the sampling of a block assignment considerably simpler. After simplification, a node's potential block assignment is informed by the number of nodes already assigned to each block. Ultimately, the probability that a node i will be assigned to the bth block is proportional to

$$\mathbb{P}[b_i = b \,|\, \mathbf{A}, \boldsymbol{\theta}, \boldsymbol{\pi}_i, \mathbf{B}_{-i}]$$
$$\propto \frac{N_{b,-i} + \alpha_B}{\alpha_B |B| + N - 1} P(A \,|\, \boldsymbol{\theta}, B_i = b) P(\pi_i \,|\, B_i = b) P(|B| \,|\, \lambda_B), \tag{30}$$

where $N_{b,-i}$ is the number of nodes assigned to block b, not including node i, α_B is related to the prior probability of being assigned to any block (analogously to α in the topic model), θ is the complete set of network parameters, $|B|$ is the number of blocks with non-zero membership while node i is being considered for potential assignment to block b, λ_B is the prior number of blocks expected to exist, and $\mathbf{B}_{-i}$ is the set of block assignments with the ith blog's block assignment removed. As such, the first term acts as a penalty term on the number of blogs in any given block, the second term is the Bernoulli likelihood (logistic regression) of observed links given other block assignments and network covariates, and the third term is the probability of blog i's topic interest vector given the considered block assignment. Finally, $P(|B| \,|\, \lambda_B)$ is the Poisson probability of $|B|$ given λ_B and acts as a penalty term for the number of blocks with non-zero membership.

To elaborate briefly, the first term in Eq. 30 and the final term in Eq. 30 together act as tunable priors on the distribution of sizes of blocks, as well as the number of non-empty blocks. In the generative model, we specify individual priors on the probability of

membership in each block. However, for any reasonable K, the total number of blocks $B = 1 + \sum_{j=1}^{3} \binom{K}{j}$ exceeds the number of blogs to assign to blocks. So the block assignment sampler accounts for blocks that have no members. This is a desirable feature since the analyst need not specify exactly how many blocks are instantiated in the model. As compared to the number of possible blocks, the number of non-empty blocks is a rare event. This enables use of the Poisson distribution as a reasonable (and computationally feasible) approximation for the number of non-empty blocks.

Due to the computational intensity of computing the block specific probabilities, we restrict the number of topics in which a block can have interest. In our work, blocks may be interested in at most three topics, except for one block that is interested in all topics (to account for such blogs as *The Huffington Post* or *The New York Times*'s blog). Furthermore, during sampling, we restrict the blocks considered for a given node i by only considering blocks that have topics for which node i generated at least one post. These restrictions change the normalizing constant, though the relative probabilities of the blocks considered remains the same.

4.8 Summary of Estimation

The primary goals of the estimation routine are to obtain post-topic assignments, and then topic token-distributions, as well as blog-specific block assignments. These are the main parameters of interest in our application, as they describe the dynamic nature of each content topic and the range of interests and communications each blog has respectively. Alongside this information, we also estimate several other parameters such as those which govern linkage formation, which topics are active when (via events), topic post-rate boost parameters, and blogs' topic interests. While each of these are informative in their own right, in our application below, we chose to focus on the topic distributions and the block assignments of each blog.

5 Results

5.1 Prior Choices

Changes in the topics for this dataset are expected to be slow, aside from sudden events that abruptly add new tokens (e.g., "Benghazi" or "Sandy Hook"). Therefore, we used a lag ℓ parameter of 62 days to capture gradual drift in the topics over time. Specifically, the distribution over tokens for each topic was estimated based upon a sliding window for the preceding 62 days. Within that window, all posts had equal weight.

To determine the number of topics, we used the criterion developed by Arun et al. (2010). The goal of this criterion is to determine the number of topics that lead to the smallest value of the criteria. It is important to note that this criteria is only based off of the topic distribution over the posts, and does not take into account the network structure. This is, of course, a limitation of this criteria and a suggestion for further research. Figure 1 shows the criterion curve, which considers fitting anywhere between 1 and 30 topics. The curve has its minimum at 22, and thus our study fixes the number of topics to be 22.

Once the number of topics is established, the restrictions on the blocks and the parameter P, as introduced in Eq. 1, can be set. Recall that each block may only be interested in 1, 2, 3, or all topics. Finally, P, the out of block interest parameter which governs the blogs ability to post on out of interest topics, was set to 50, to allow some freedom for blogs to
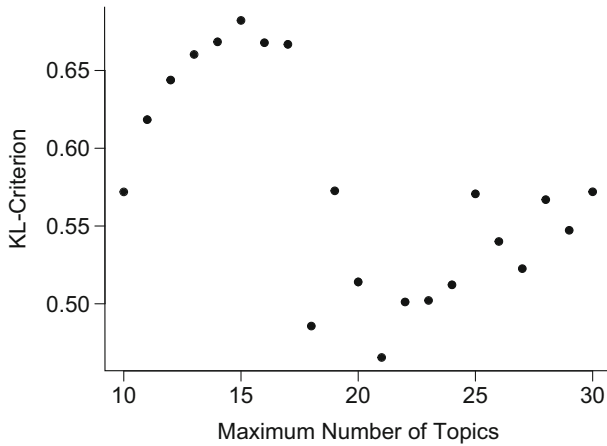
Fig. 1 The criterion curve, as in Arun et al. (2010), for determining the number of topics

post on topics outside of their block's interests, but nonetheless mostly focus on the block's interests.

The network model specified an edge parameter, a mean in-degree and out-degree parameter, a 7-day lag parameter, and a block membership parameter. The edge parameter acts as the intercept for the network model. Mean in-degree and out-degree are nodal covariates consisting of the average daily in-degree and out-degree for each node. This allows modeling of differentially popular blogs. Finally, to add in temporal dependency, the 7-day lag is an indicator function that takes the value 1 if and only if the pair of blogs has been linked within the previous 7 days, and is otherwise 0 (this captures the fact that bloggers sometimes have debates, which produce a series of links over a relatively short period of time). Vague priors were set for each of the network model parameters; all were normals with mean 0 and standard deviation 1000. The proposal standard deviation was set to 1 for the edge parameter, and to 0.25 for each of the other parameters in the network model.

For the topic model the α and β parameters were both set to 0.1. The prior for the average post rates ρ_i in Eq. 2 was a truncated normal at 0, with mean 4 and standard deviation 1000. The prior for topic activation parameters ψ_k in Eq. 2 was set as a truncated normal at 0 with mean 0 and standard deviation 1000, and a proposal standard deviation of 0.5.

Additionally, 25 was set as the prior mean number of blocks (λ_B), and the prior tendency for block membership α_B was set to 1. The prior probability of topic activation was set to 0.2.

The sampler ran for 1000 iterations. To ensure mixing for the network parameters, at each iteration the network parameters were updated 10 times. During each iteration, there were 10 sub-iterations for the topic model and 10 sub-iterations for the block assignments. The first 100 overall iterations were discarded as burn-in, the remaining 900 were thinned at intervals of 10.

5.2 Findings

The sampler converged to stationarity quickly in every parameter. To assess the mixing of the post to topic assignment, and of the blogger to block assignment, we calculated Adjusted Rand Indices (Hubert and Arabie 1985; Steinley 2004) for each iteration i compared to iteration $i-1$. The post to topic assignment was very stable, with a mean Adjusted Rand

Index of 0.806 and standard deviation 0.047. The block assignment was less stable, with a mean Adjusted Rand Index of 0.471 and standard deviation 0.031. We believe this variability is due to the fact that many bloggers tended to post on whatever news event captured their attention, making it difficult to assign them to a block with interest in no more than three topics. However, their interests were not so wide that they were reasonably assigned to the block that is interested in all topics.

All domain level parameters converged successfully. The domain rate parameter ρ_i was estimated for each domain, and the posterior means of the domain rates had a mean of 0.632 and a standard deviation of 1.67. The largest domain rate was 22.69. The distribution of domain post rates was highly skewed, with few blogs having a very high average post rate, and most blogs having a lower post rate.

The topic-specific activation parameters ψ_k converged successfully. Information on the posterior means and standard deviations is in Table 3, and were calculated after the topics had been defined. The topics Election and Republican Primary have the greatest posterior means, which suggests that these topics were more event driven than other topics.

5.2.1 Topic Results

The topic model found 22 topics, each of which had distinct subject matter. Table 1 contains the topic titles and total number of posts in each topic, as well as the three tokens that have

Table 1 Topic names and their most specific tokens

Topic name	# of posts	Highest specificity tokens		
		1	2	3
Feminism	3971	russel.saunder.juli	circumcis	femin
Keystone pipeline	4422	loan.guarante.program	product.tax.credit	tar.sand.pipelin
Birth control	2703	contracept.coverag	birth.control.coverag	religi.organ
Election	14,713	soptic	cheroke	eastwood
Mortgages	2130	estat	probat	fiduciari
Entertainment	10,555	email.read.add	olivia	free.van
Middle East	6068	mursi	morsi	fatah
LGBT rights	5425	anti.gay.right	support.equal.marriag	equal.marriag
Sensational crime	6423	zimmerman	lanza	mass.shoot
Technology	3230	mail.feel.free	pipa	ret
Supreme court	1767	commerc.claus	bork	chief.justic.robert
Bank regulation	5222	volcker	dimon	libor
National defense	1977	iaea	iranian.nuclear.weapon	warhead
Republican primary	9351	poll.mitt.romney	nation.popular.vote	romney.lead
Voting laws	7865	ohio.secretari.state	voter.registr.form	hust
Political theory	1448	bylaw	rawl	sweatshop
Eurozone	1832	standalon	troika	ecb
Taxation	8435	tax.polici.center	health.care.spend	top.tax.rate
Diet and nutrition	3057	spielberg	harlan	calori
Education	2909	chicago.teacher.union	chicago.public.school	charter.school
Global warming	2205	arctic.sea.ice	sea.ice	sea.level.rise
Terrorism	3347	kimberlin	broadwel	assang

Table 2 The most frequent words in each topic

Topic names	Most frequent words				
	1	2	3	4	5
Feminism	women	peopl	dont	person	life
Keystone pipeline	energi	oil	price	compani	industri
Birth control	right	state	law	marriag	women
Election	obama	romney	peopl	presid	polit
Mortgages	case	court	bank	judg	attorney
Entertainment	peopl	dont	good	work	game
Middle East	israel	islam	american	peopl	countri
LGBT rights	gay	peopl	marriag	homosexu	support
Sensational crime	gun	polic	report	peopl	zimmerman
Technology	compani	googl	facebook	appl	user
Supreme court	law	court	state	case	constitut
Bank regulation	bank	market	money	price	compani
National defense	iran	militari	israel	nuclear	obama
Republican primary	romney	republican	obama	poll	vote
Voting laws	state	vote	elect	voter	counti
Political theory	libertarian	peopl	right	state	govern
Eurozone	bank	debt	economi	rate	govern
Taxation	tax	state	govern	cut	obama
Diet and nutrition	peopl	dont	govern	polit	work
Education	school	student	teacher	educ	state
Global warming	climat	climat.chang	temperatur	scienc	scientist
Terrorism	report	govern	attack	inform	case

the highest predictive probability for that topic over all days. Predictive probability was calculated using Bayes' rule:

$$P(Z_d = k \mid w \in d) = \frac{P(w \in d \mid Z_d = k)P(Z_d = k)}{P(w \in d)}. \tag{31}$$

Table 2 contains the five most frequent tokens in each topic over all days. Topics were named by the authors on the basis of the most predictive tokens as well as the most frequent tokens over all days. Some of these tokens may seem obscure, but in fact they are generally quite pertinent to the identified topics.

It is beyond our scope to detail the dynamics of all 22 topics. However, a close look on one topic, Sensational Crime, shows the kind of information this analysis obtains. The posts about Sensational Crime largely concerned four events: the shooting of Trayvon Martin in February, the Aurora movie theater shooting in July, the Sikh Temple shooting in August, and the Sandy Hook shooting in December.[4]

---

[4]Trayvon Martin was a young African American man shot by George Zimmerman, in what he claimed to be an act of self defense, while Martin was walking in Zimmerman's neighborhood. The Aurora theater massacre was a mass shooting at a movie theater in Aurora, Colorado. The Sikh Temple shooting was a mass shooting at a Sikh temple in Wisconsin. The Sandy Hook massacre was a mass shooting at an elementary school in Connecticut.

To illustrate how the salience of a token changes over time, we use a weighted frequency proportion which is equal to

$$WF_{w \in k} = \frac{P(Z_{d_t} = k_t \mid w \in d) F(w \in k_t)}{\sum_{w^* \in V} P(Z_{d_t} = k_t \mid w^* \in d) F(w^* \in k_t)}, \tag{32}$$

where $F(w \in k_t)$ is the frequency of the token w in the kth topic's distribution at time t. This weighted frequency can be interpreted as the proportion of topic-specific tokens at time t that is taken up by token w, and is useful in this context as many of the tokens are shared at high frequency between topics (such as "people") and are therefore uninformative. So this quantity tracks the topic-specific information of a token over time. Recall that the topic-specific token-distributions are computed over the past 62 days, which accounts for the smoothness of the curves. The gray shading around each curve represents the 95% Bayesian credible interval.

Figure 2 presents the weighted frequency curves for tokens specifically related to the shooting of Trayvon Martin. The plot shows three interesting features. First, the prevalence of all of the tokens does not spike up at the time of the shooting (2/26/2012), but rather at the time of Obama's press statement regarding the shooting. Second, the term "zimmerman" dominates the tokens, and in fact is the most prevalent token in the whole of the Sensational Crime topic from March 22 to July 16. The gap in prevalence between the tokens of "zimmerman" and "trayvon" or "trayvon martin" is also interesting, suggesting that in this case, media attention was on the perpetrator rather than the victim.

Figure 3 tracks the major events in the Sensational Crime topic for the entire year. Notably, media focus is never as strong on the tokens related to the events as it is on "Zimmerman" specifically. Rather, the usual top terms over the course of the year are "police"
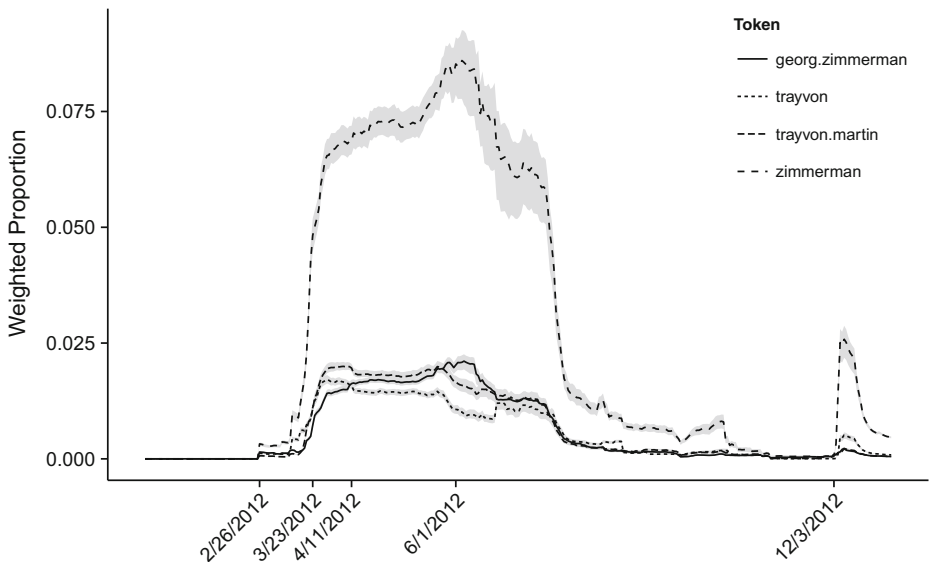

Fig. 2 Weighted proportion timeline of the Trayvon Martin shooting and the subsequent legal case. The date 2/26/2012 is when Trayvon Martin was shot, 3/23/2012 is when President Obama said that Trayvon could have been his son, 6/1/2012 was when Zimmerman's bond was revoked, and 12/3/2012 was when photos were released showing Zimmerman's injuries on the night of the shooting. (95% credible intervals shown)
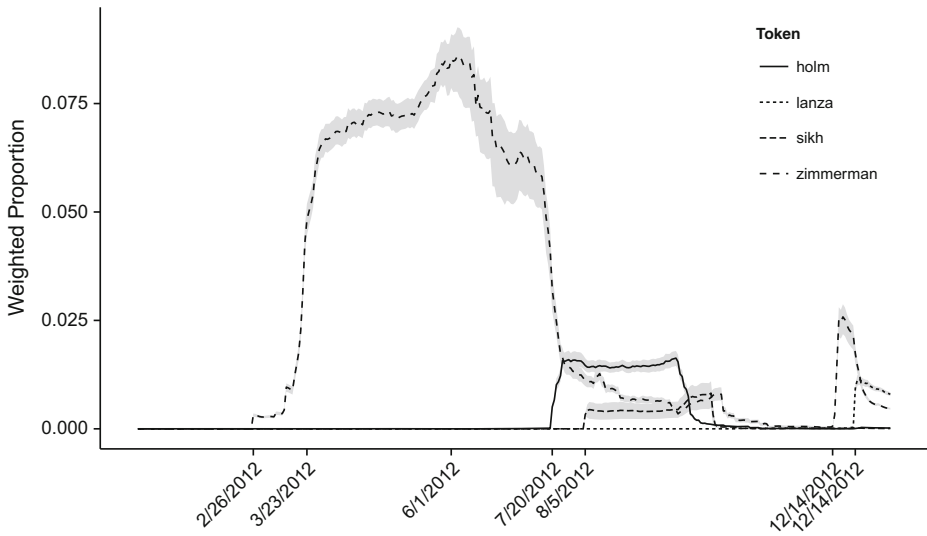
Fig. 3 Weighted proportion timeline of major events in the Sensational Crime topic. The date 2/26/2012 is when Trayvon Martin was shot, 3/23/2012 is when President Obama says that Trayvon could have been his son, 6/1/2012 is when Zimmerman's bond was revoked, 7/20/2012 is the Aurora mass shooting by James Holmes, 8/5/2012 is the Sikh Temple shooting by Michael Page, 12/3/2012 is when photos showing Zimmerman's physical injuries were released, and 12/14/2012 is when the Sandy Hook massacre occurred. (95% credible intervals shown)

and "gun." Also notable is the lack of events in the later part of the year after the media attention on the Sikh Temple Shooting receded.

5.3 The Network Results

Table 3 shows the posterior means and standard deviations of the topic specific activation parameters, while Table 4 shows the posterior means of the network parameters with 95% credible intervals. The edge parameter posterior mean indicates that the network is rather sparse at most time points. Interestingly, the 7-day lag parameter was negative, suggesting that blogs which were recently linked were less likely to link in the near future. There are two plausible explanations for this finding. First, the linkage dynamics may not be driven by recent links, but rather the links are a consequence of the events taking place. An upsurge in linking when an event occurs is followed by a decrease in the number of links as the event fades out of the news cycle. Second, if linking is done as part of a debate, then once a point has been made, the bloggers may not feel a need for back-and-forth argument.

The block parameter is strongly positive (mean = 1.058, standard deviation 0.240), suggesting that blogs which share common interests are more likely to link to each other. This is particularly important, as the block statistic was not only formed from explicit block matching, but also from blogs that did not share the same interests. The block statistic is proportional to the shared topic interests. This result directly links the network model to the topic model, and allows the analyst to make claims about the block structure as inferred from the topics.

Finally, and predictably, both the in-degree and out-degree of a blog increases the probability that the block will receive links. These parameters were included in the analysis to

Table 3 Topic-specific activation parameters ψ_k

Topic	Posterior mean	Standard deviation
Feminism	0.0034	0.0029
Keystone pipeline	0.0037	0.0017
Birth control	.00001	.00003
Election	0.3917	0.0249
Mortgages	.00001	.00005
Entertainment	0.0018	0.0014
Middle East	0.0378	0.0129
LGBT rights	0.0028	0.0026
Sensational crime	0.0208	0.0085
Technology	0.0012	0.001
Supreme court	0.0022	0.0032
Bank regulation	0.0021	0.0021
National defense	0.0014	0.0013
Republican primary	0.2267	0.0188
Voting laws	0.0375	0.0095
Political theory	0.0012	.00001
Eurozone	0.0001	.00002
Taxation	0.0135	0.0084
Diet and nutrition	0.0804	0.0139
Education	0.0011	0.0012
Global warming	0.0019	0.0011
Terrorism	0.0022	0.0019

control for the influence of highly popular blogs such as *The Blaze* and *The Huffington Post*.

We can examine the link dynamics within a topic block. There were 21 blogs whose maximum posterior probability of block assignment placed them in the block that was only interested in the Sensational Crime topic. Only 2 of these 21 blogs received any links over the course of the year, and only 1 received links within the block (*legalinsurrection.com*). While this runs counter to the idea that they form one block, recall that blogs are also more likely to link to blogs that share some of the same topic interest. There are a total of 62 blogs to which members of the Sensational Crime block link, and 15 of these blogs receive approximately 90% of the links. As such, the Sensational Crime topic block appears to be a set of "commenter" blogs that react to posts that are posted on larger blogs. Our model allows the analyst to isolate the blogs that post on a particular topic, to get a better idea of the linkage dynamics around important events. As an example, we describe how the linkage

Table 4 Posterior means and 95% credible intervals for network parameters

Parameter	Posterior mean	95% CI
Edges	− 8.524	[− 8.539, − 8.513]
7 day lag	− 0.163	[− 0.198, − 0.131]
Block	1.058	[0.638, 1.485]
Outdegree of receiver	0.330	[0.329, 0.332]
Indegree of receiver	0.497	[0.496, 0.499]

pattern changes around the occurrence of Barack Obama's speech regarding the shooting of Trayvon Martin, and also following the Aurora shooting.

Figures 4 and 5 show the link structure from the blogs in the Sensational Crime block to other blogs. The data are aggregated over 15 days. Figure 4 pertains to the days before President Obama's press conference regarding Trayvon Martin on 3/23/2013, and Fig. 5 pertains the days following his remarks. Figure 6 pertains to the period immediately before the Aurora shooting on 7/20/2012, and Fig. 7 pertains to the period immediately after.

To improve interpretability, only a subset of blogs and links are plotted. Specifically, blogs that were assigned to the block interested only in Sensational Crime, and who posted during the specified time frame, are plotted. Additionally, blogs who are part of the 15 blog subset that received 90% of the links from the Sensational Crime block, and which received links within the timeframe, are plotted. Also, links generated from blogs in the Sensational Crime block to other members of the same block, or to other blogs, are plotted. Links emanating from the 15-node subset are not plotted. These plotting constraints help enable us to discern and interpret the community structure that formed in the discussion of these events.

Comparing Figs. 4 and 5 shows that the community structure seen in the linkage pattern did not change much as a result of the press conference in which President Obama remarked that if he had a son, he would resemble Trayvon Martin. Remarkably, there was also no net increase in posting rates. It is known that there was a flurry of posts at this time, and it turns out that uptick was allocated to the Election block, as people speculated on how his remarks would affect the 2012 presidential election.
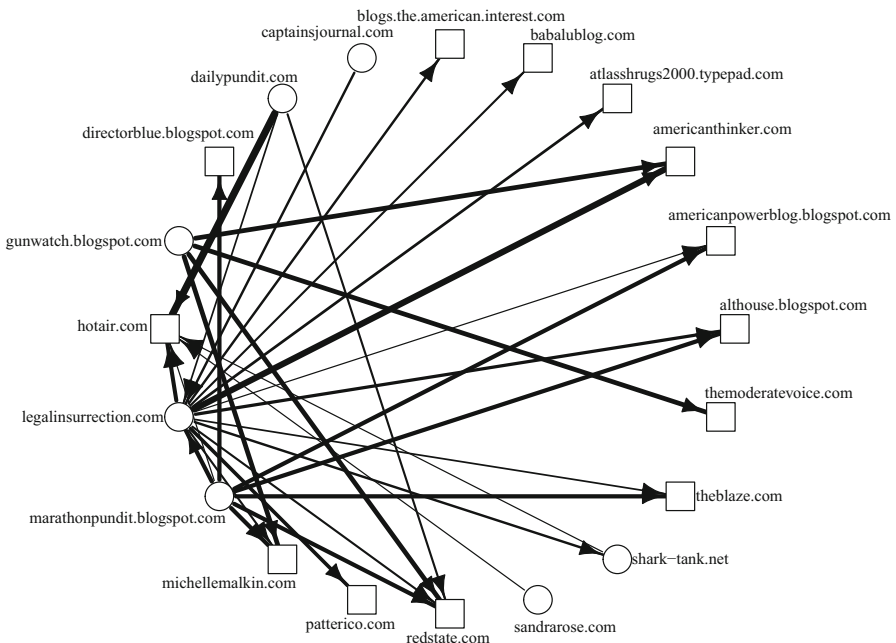

Fig. 4 Fifteen-day aggregate linkage from 3/8/2012 to 3/22/2012, immediately before President Obama's comment. The number of links, represented by line thickness, is root transformed for clarity. Circular nodes are blogs in the Sensational Crime block. Square nodes are blogs to which the Sensational Crime blogs link, and these blogs are generally in multi-topic blocks, where one of the topics is Sensational Crime
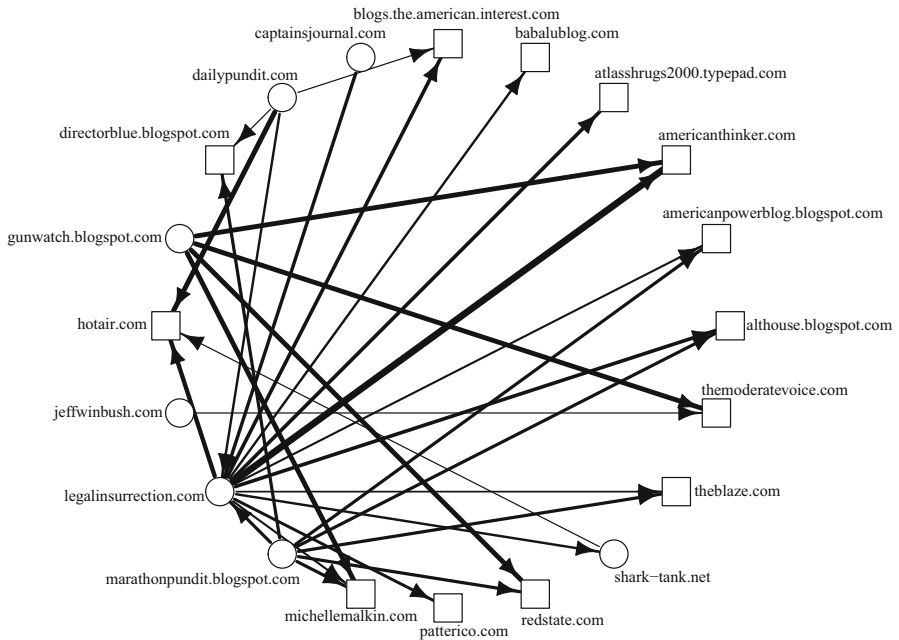
Fig. 5 This figure is constructed in the same way as Fig. 4, but for the time period from 3/23/2012 to 4/6/2012, immediately after President Obama's comment
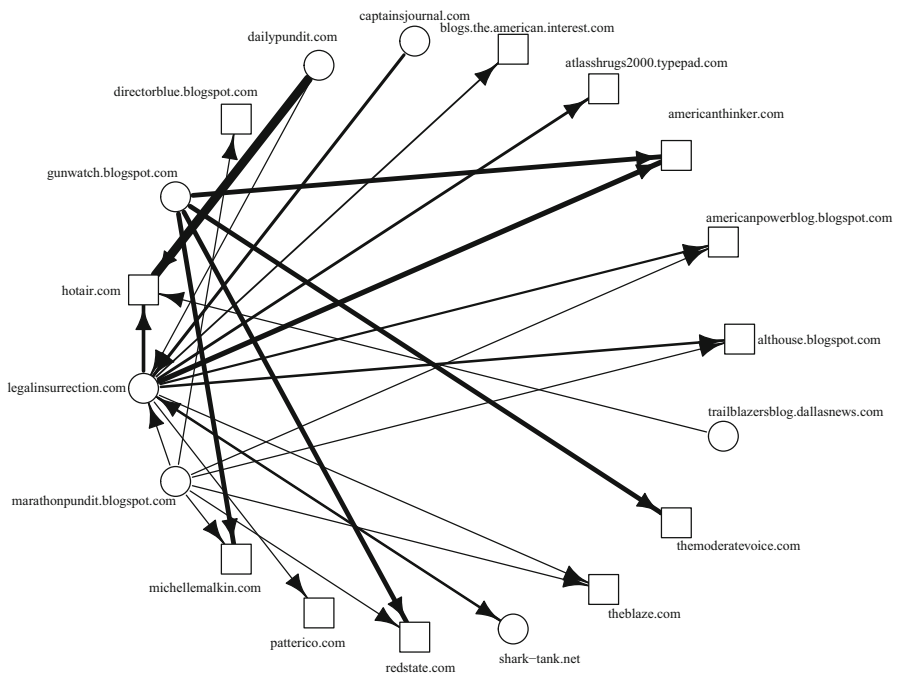

Fig. 6 This figure is constructed in the same way as Fig. 4, but for the time period from 7/5/2012 to 7/19/2012, immediately before the Aurora shooting
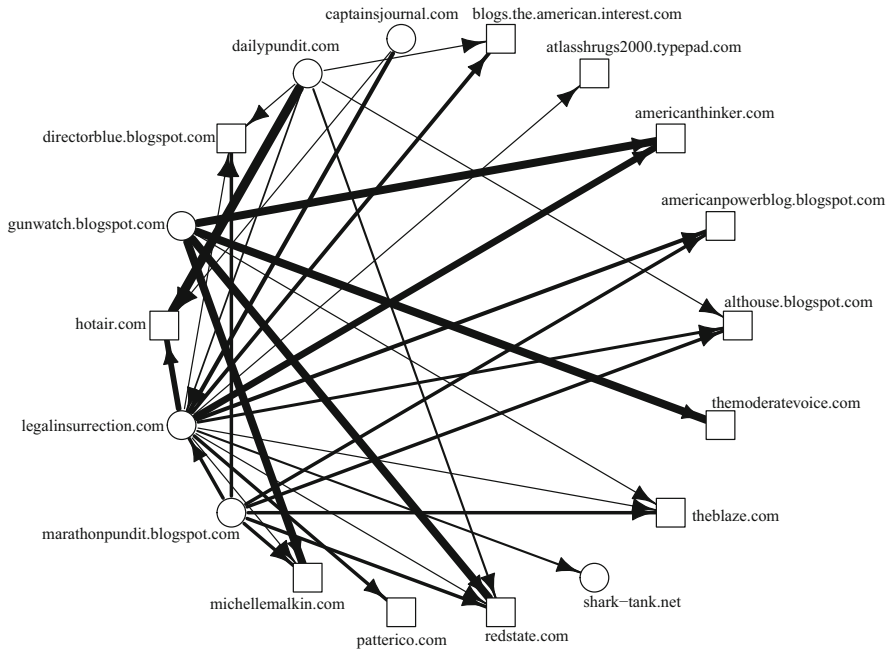
Fig. 7 This figure is constructed in the same way as Fig. 4, but for the time period from 7/20/2012 to 8/2/2012, immediately after the Aurora shooting
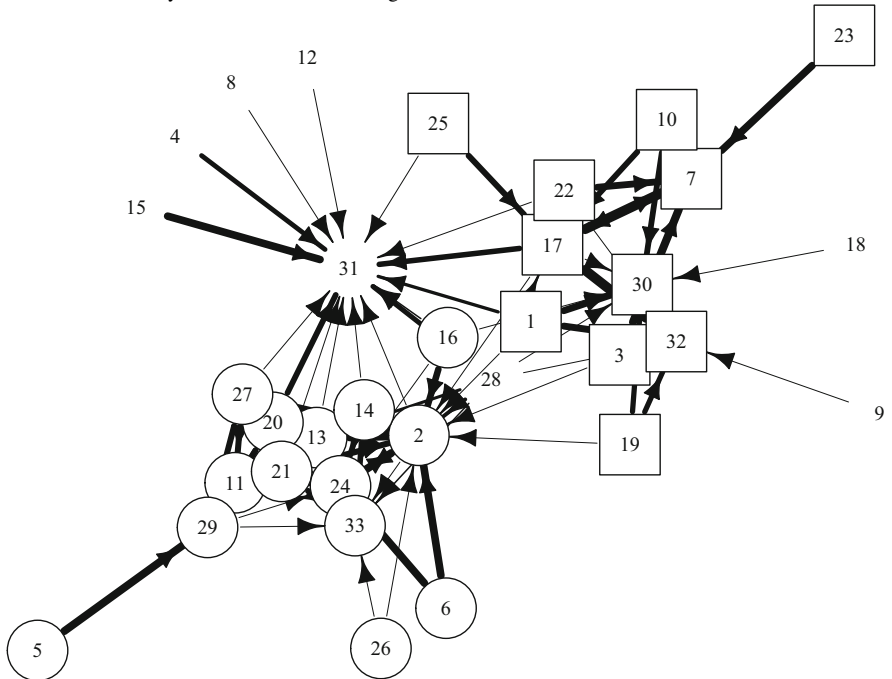

Fig. 8 Community substructure of the Election block. Circles and squares denote separate communities. The absence of shape denotes membership in a small community. The thickness of edges correspond to a log transformation of the number of links sent over the entire year

The patterns surrounding the Aurora shooting (Figs. 6 and 7) are more clear. The community structure in the discussion is essentially the same, but the amount of traffic increases conspicuously. Specifically, the number of links in the 15 days before the shooting was 197, but afterwards it was 427. Linkage rates especially increase from `gunwatch.blogspot.com`. In general, this agrees with the conclusion that the methodology is able to detect stable communities whose linkage rates are driven by news events.

To further illustrate the findings of the network model for a different block, we now present examples from the Election block. There are 52 blogs that the model assigned to the block whose only interest was the presidential election. Of these 52 blogs, 33 blogs linked

Table 5 Blog names and their community membership

Label	Blog	Community
1	afeatheradrift.wordpress.com	1
2	atlasshrugs2000.typepad.com	3
3	bleedingheartlibertarians.com	1
4	brainsandeggs.blogspot.com	2
5	citizentom.com	3
6	crethiplethi.com	3
7	crookedtimber.org	1
8	davedubya.com	4
9	dogwalkmusings.blogspot.com	5
10	driftglass.blogspot.com	1
11	greatsatansgirlfriend.blogspot.com	3
12	hennessysview.com	6
13	joshuapundit.blogspot.com	3
14	marezilla.com	3
15	mediabistro.com	2
16	michellesmirror.com	3
17	nomoremister.blogspot.com	1
18	ochairball.blogspot.com	7
19	patriotboy.blogspot.com	1
20	righttruth.typepad.com	3
21	rightwingnews.com	3
22	rogerailes.blogspot.com	1
23	rwcg.wordpress.com	1
24	sultanknish.blogspot.com	3
25	tbogg.firedoglake.com	1
26	thecitysquare.blogspot.com	3
27	therightplanet.com	3
28	thoughtsandrantings.com	2
29	varight.com	3
30	blogs.suntimes.com	1
31	mediaite.com	2
32	rightwingwatch.org	1
33	patdollard.com	3

to or received links from other blogs within this same block. And of these 33 blogs, 12 were the recipients of all links. We use random walk community detection (Pons and Latapy 2006) upon the Election block to show that the model can extract meaningful subnetworks for use in secondary analyses.

Figure 8 shows the community substructure for the Election block aggregated over the entire year. Random walk community detection indicates that seven communities optimized modularity, but two communities contained the majority of the blogs. As such, for interpretability, only these two communities are shown. The modularity of this partition is 0.49, and a 10,000 sample permutation test of the community labels indicated that this value of modularity was in the 99th percentile (the greatest modularity found in the permutation test was 0.317). This result indicates that the model found meaningful community structure, rather than sampling variability.

Table 5 contains the blog labels for the Election block network. Examination of the block structure shows that the majority of the blogs partitioned into one of two communities. Based on Technorati ratings, the community plotted as circles in Fig. 8 is politically conservative, while the other community plotted as squares is liberal. This separation of the two ends of the political spectrum has been found before in blogs (Lawrence et al. 2010). There is little communication between the two communities, but a lot of communication within those communities. Interestingly, both communities sent many links to blog 31, which was allocated into a distinct community that it shared with blogs 15, 4, and 28. Blog 31 is *mediaite.com*, a non-partisan general news and media blog, and the pattern of links from both partisan communities suggests that *mediaite.com* acts as a common source of information.

6 Conclusion

In this manuscript, we present a Bayesian model for analyzing a large dataset of political blog posts. This model links the network dynamics to topic dynamics through a block structure that informs both the topic assignment of a post and the linkage pattern of the network.

A major feature of our model is that the block structure enables interpretable associations among topics. For example, there is a two-topic block whose members are interested in both Election topic and the Republican Primary topic, but there is no block whose members are interested in just the Supreme Court and Global Warming. That pattern of shared interest conforms to what one would expect.

Another feature of our model is the flexibility of the network model. This analysis uses a limited set of predictors, but the ERGM modeling framework can easily incorporate additional covariates (Robins et al. 2001) and structural features. Additionally, if one uses a maximum pseudo-likelihood approach (Frank and Strauss 1986) as a way of approximating the likelihood, then higher order subgraph terms, such as number of triangles or geometrically weighted edgewise shared partners (Hunter et al. 2008) can account for transitivity effects. Finally, while the block structure modeled in this paper was based upon similarity in topic interest, more nuanced models are possible, and these could use information on, say, political ideology, which the analysis of the Election block found to be important in predicting linkage patterns.

Finally, one major advantage of our approach to modeling this data is the nonparametric nature of the topic dynamics. By avoiding an autoregressive specification of topic dynamics, as in Blei and Lafferty (2006), topics are able to change more freely; in particular, it is

possible for new tokens with high probability to emerge overnight. This is ideal for the blog data, since the blogs are often responding to news events.

Our analysis of the political blog dataset had an interpretable topic and block set, and analysis of the Sensational Crime block and the Election block reached reasonable conclusions. Specifically, the dominance of the token "zimmerman" across the year agrees with our sense of the tone and primacy of that discussion, and the spike following the Aurora shooting is commensurate with its news coverage. The Election block neatly split into sub-communities along partisan lines, which accords with previous research (Lawrence et al. 2010).

While our focus was on the specific application of the political blog data, the model developed here has features that can generalize to other dynamic text networks. Such as the Wikipedia and scientific citation networks. Specifically, the connection of topic and link structure through a block structure allows for document content to inform the community structure of the overall network. However, each application requires some hand fitting that captures specific aspects of the data. For example, the block structure might need to be dynamic; this would make sense is scientific citation networks, since disciplines sometimes bifurcate (e.g., the computer science of 1970 has now split into machine learning, quantum computing, algorithms, and many other subfields). Also, scientific citation is strictly directional in time—one cannot cite future articles. But the Wikipedia is not directional in time; an article posted a year ago can send a link to one posted yesterday. So specific applications will require tinkering with the model described here.

The work presented here suggests several avenues of future research. On the methodological side, the model can be generalized and extended in several ways. Specifically, the block membership could be considered a dynamic property, allowing blogs to change interest in topics over time. Additionally, building a dynamic model for link patterns would allow researchers to examine specific properties of links over time, rather than assuming the same link generating distribution at all time points. Finally, this model can be adapted to other dynamic text networks, and its performance should be compared to more traditional topic analysis and community detection procedures.

Publisher's Note   Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, *9*(2008), 1981–2014.

Arun, R., Suresh, V., Madhavan, C.V., Murthy, M.N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining* (pp. 391–402). Springer.

Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D.M., & Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM.

Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Chang, J., & Blei, D.M. (2009). Relational topic models for document networks. In *International conference on artificial intelligence and statistics* (pp. 81–88).

Faust, K., & Wasserman, S. (1992). Blockmodels: interpretation and evaluation. *Social networks*, *14*(1), 5–61.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*(395), 832–842.

Gilks, W.R., Best, N., Tan, K. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, *44*, 455–472.

Ho, Q., Eisenstein, J., Xing, E.P. (2012). Document hierarchies from text and links. In *Proceedings of the 21st international conference on World Wide Web* (pp. 739–748). ACM.

Hoff, P.D., Raftery, A.E., Handcock, M.S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098.

Hoffman, M., Bach, F.R., Blei, D.M. (2010). Online learning for latent dirichlet allocation. In *Advances in neural information processing systems* (pp. 856–864).

Holland, P.W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, *76*(373), 33–50.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Hunter, D.R., Goodreau, S.M., Handcock, M.S. (2008). Goodness of Fit of Social Network Models. *Journal of the American Statistical Association*, *103*(481), 248–258.

Krivitsky, P.N., & Handcock, M.S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 29–46.

Latouche, P., Birmelé, E., Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Annals of Applied Statistics*, *5*(1), 309–336.

Lawrence, E., Sides, J., Farrell, H. (2010). Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics*, *8*(01), 141.

McNamee, P., & Mayfield, J. (2003). Jhu/apl experiments in tokenization and non-word translation. In *Comparative evaluation of multilingual information access systems* (pp. 85–97). Springer.

Moody, J. (2004). The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.

Newman, M.E.J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *69*(22), 026113.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, *10*(2), 191.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Robins, G., Elliott, P., Pattison, P. (2001). Network models for social selection processes. *Social Networks*, *23*(1), 1–30.

Snijders, T.A., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, *14*(1), 75–100.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological Methods*, *9*(3), 386.

Technorati (2002). https://web.archive.org/web/20140420052710/http://technorati.com/.

Wang, E., Silva, J., Willett, R., Carin, L. (2011). Dynamic relational topic model for social network analysis with noisy links. In *2011 IEEE, statistical signal processing workshop (SSP)* (pp. 497–500). IEEE.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and *p**. *Psychometrika*, *61*(3), 401–425.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233–242). ACM.