A Time-Dependent Topic Model for Multiple Text Streams

Liangjie Hong* †, Byron Dom §, Siva Gurumurthy, § Kostas Tsioutsiouliklis §
† Dept. of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA
§ Yahoo! Labs, Sunnyvale, CA, USA
hongliangjie@lehigh.edu, byron_dom@yahoo.com, {shiiva,kostas}@yahoo-inc.com

ABSTRACT

In recent years social media have become indispensable tools for information dissemination, operating in tandem with traditional media outlets such as newspapers, and it has become critical to understand the interaction between the new and old sources of news. Although social media as well as traditional media have attracted attention from several research communities, most of the prior work has been limited to a single medium. In addition temporal analysis of these sources can provide an understanding of how information spreads and evolves. Modeling temporal dynamics while considering multiple sources is a challenging research problem. In this paper we address the problem of modeling text streams from two news sources - Twitter and Yahoo! News. Our analysis addresses both their individual properties (including temporal dynamics) and their inter-relationships. This work extends standard topic models by allowing each text stream to have both local topics and shared topics. For temporal modeling we associate each topic with a time-dependent function that characterizes its popularity over time. By integrating the two models, we effectively model the temporal dynamics of multiple correlated text streams in a unified framework. We evaluate our model on a large-scale dataset, consisting of text streams from both Twitter and news feeds from Yahoo! News. Besides overcoming the limitations of existing models, we show that our work achieves better perplexity on unseen data and identifies more coherent topics. We also provide analysis of finding real-world events from the topics obtained by our model.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*

General Terms

Algorithms, Experimentation, Theory

*This work was done when the first author was on a summer internship at Yahoo! Labs.

KDD'11, August 21–24, 2011, San Diego, California, USA.

Keywords

Topic models, Text streams, Temporal dynamics, News, Twitter

1. INTRODUCTION

Social-networking tools such as Facebook, LinkedIn and Twitter, have become the communication tools of choice for a large number of online users. Such tools are increasingly used for disseminating breaking news and eyewitness accounts, and even for organizing flash mobs and protest groups. For instance, Twitter was heavily used in a number of international events, such as the Iran election in 2009, the Haiti earthquakes in 2010, and the tsunami in Japan in 2011. More recently, social networking services were instrumental in facilitating the political upheavals in the Middle East. Social media as well as the on-line publishing of more established media (e.g., newspapers, magazines and televison) have attracted a lot of attention from both researchers and product developers.

This increasing use of social media has resulted in a refocusing of research activities onto related problems, many of which are new. For example, there exists an argument as to whether social media have influenced traditional media sources and in what sense, or vice versa. In addition, people are wondering whether the topics that are shared and discussed on social media significantly differ from traditional information sources and how these topics are transferred from one source to another. Moreover, questions about the differences between various types of social media (e.g., blogs, community-based questions-and-answer portals and microblogging services) have been raised continuously both in research communities and industry. Effectively addressing these issues requires the ability to analyze multiple types of information sources over time.

Problems similar to these have been attacked from various perspectives. For modeling the temporal dynamics of information Kleinberg et al. [14, 15] proposed methods to track the volume of a single term over time. Their later work (e.g., [16]) attempts to monitor the temporal dynamics of "memes" by which the authors mean sentence fragments representing concepts. In addition work has been done to study the dynamics of blogs [9], of online knowledge sharing communities [2], of news articles and stories [16], and of microblog services [13]. While most of the above-mentioned works focused on a single media source, some authors [31, 27, 28] modified Probabilistic Latent Semantic Analysis (PLSA) [11] to simultaneously model documents from different text streams. There is also some recent work in comparing social and traditional media. Zhao et al. [33] tried to obtain latent topics from Twitter and New York Times (NYT) news articles by using topic models. Two different topic models were used to learn the topics from the two sources separately and heuristics were then applied to obtain both common and local topics. Attempts have been made to extend topic

models to incorporate temporal dynamics and topic evolution (e.g., [4, 26]). In addition to research projects, commercial products also provide tools to search and browse the dynamics of queries[1], news articles and web traffic[2], and microblogging updates[3].

While existing research offers different methods to monitor and track correlated information sources over time, many of the proposed approaches suffer from significant drawbacks. For instance most of the work on tracking information sources primarily focuses on only one type of source. Given the multiplicity of media channels however, it is potentially more useful to understand multiple information sources simultaneously. Also, tracking a single word or a meme can be quite limiting. Further, most models that consider multiple text collections either have model parameters requiring manual adjustment or have theoretical limitations (see our discussion in Section 2). In addition temporal factors are either not incorporated in the models or are heuristically embedded. For temporal topic models most approaches adopt a Markovian assumption that may not be suitable for social media. Indeed, none of them utilize recent research findings of temporal variations of information in social media [16, 30].

In this paper we address the problem of modeling multiple text streams, including their temporal dynamics, in a principled manner. Our work builds on recent work in both information dynamics and topic models. More specifically, we extend topic models by allowing each text stream to have both local and shared topics. For temporal modeling, we associate each topic with a time-dependent function that characterizes its popularity over time. By combining the two models, we effectively model temporal dynamics of multiple correlated text streams in a unified framework. To summarize the contributions of this paper, the work we describe includes:

- a topic model that discovers common and uncommon topics from multiple text collections
- a temporal model that characterizes the dynamic of topics over time
- a simple and potentially scalable algorithm for mining temporal topics
- interesting results from Yahoo! News and Twitter obtained by applying our model.

The remainder of this paper is organized as follows. Section 2 provides the background and related work. In Section 3 and Section 4, we discuss our model in detail. Section 5 provides experimental results on real-world datasets. We conclude our paper with Section 6, which discusses both conclusions and future work.

2. RELATED WORK

Mining common topics and their temporal dynamics from multiple text streams can be loosely decomposed into the two independent tasks of (1) recovering topics and (2) characterizing their temporal dynamics. We review these two lines of related work.

Based on PLSA, Wang et al. [27] introduced an observed-timestamp variable into the generative model to incorporate temporal dynamics. In addition several heuristics were applied to smooth topics in consecutive time periods. Later, Wang et al. [28] followed a similar idea and used an artificial time-synchronization optimization process in their model to re-organize the time stamps of all documents so that documents with the same time stamp would share similar topics. We argue that the constraint imposed by this synchronization is unrealistic. Note that these two papers do not differentiate between common topics and topics that only occur in a single text stream. Moreover, since both models are based on PLSA, they have the tendency to overfit the data. Furthermore, both models are not well-defined generative models [5] and no assumptions on how topic distributions and per-document topic-proportion distributions change over time were made in these models. In a recent paper, Zhang et al. [32], in addressing the same problem, proposed a non-parametric model in which a Markovian assumption is made regarding the temporal dynamics of document-topic distributions. As mentioned in the previous section, however, according to recent results on information propagation and temporal variations [16, 30], this assumption may not be appropriate for social media.

Independent of temporal factors, two basic approaches to topic discovery from correlated text streams exist in the topic modeling literature. Zhai et al. [31] proposed two variants of the same idea to tackle the problem of modeling multiple text streams. One variant assumes that each document in a text stream is generated by a background language model and a set of topics. Both the background language model and topics are multinomial distributions over words shared across multiple text streams. Since they are shared across all streams, common topics are difficult to identify. The second variant also assumes that each document in a text stream is generated by a background language model and a set of topics. Once a term is chosen to be generated by topics, a topic index is first selected followed by a second-level decision regarding whether the word is generated by a common or a local topic. The model can then explicitly handle common and local topics among multiple streams. Common and local topics are aligned under the same set of indices however, forcing the total number of topics to be the same for all streams. In addition the background language is the same across all text streams, which is too inflexible for the joint modeling of disparate sources such as Twitter and Yahoo! News. Also, per-document topic-proportion parameters must be manually tuned in experiments, which is impractical for real applications. The first variant inspired models introduced in [27, 28] and the second variant was extended to a fully Bayesian formulation by Paul et al. [22, 23], in which the topic proportion parameters were automatically estimated from the inference algorithm but local topics among different text streams were forcibly put under the same set of indices. It is therefore possible that unrelated topics will be brought together under the same topic index due to this constraint.

We briefly review some of the recent extensive work on modeling temporal dynamics in topic models. Early work on incorporating temporal evolution usually made a Markovian assumption by using either a state-space model (e.g., [4, 25]) or a linear model (e.g., [24]). Besides the Markovian assumption, Wang et al. [26] introduced a beta distribution over timestamps using a non-Markovian topic model. Nallapati et al. [20] and Iwata et al. [12] focused on the problem of modeling topics spread on a timeline with multiple resolutions, namely how topics are organized in a hierarchy and how they evolve over time. Ahmed and Xing [1] proposed a non-parametric model to address the birth and death of topics over a timeline using a Markovian assumption. The datasets used in these papers are several orders of magnitude smaller than the one we used in this paper.

3. CORRELATED TEXT STREAMS

3.1 Model Description

Our correlated-text-stream model (`Collection Model`) is an extension of Latent Dirichlet Allocation [5] (LDA). In our

[1] http://www.google.com/insights/search/
[2] http://www.google.com/trends
[3] http://www.google.com

`Collection Model`, we have a set S of n text streams. Associated with each stream $s \in S$ is a set T_s of *local* topics and associated with all streams is a set T_c of *common* topics. Thus the total number of topics in the model is $(\sum_s |T_s|)+|T_c|$. As in LDA, each topic k is defined as a multinomial distribution over a fixed vocabulary V, denoted as ϕ_k. Local topics $\phi^{(s)}$ are drawn from stream-dependent Dirichlet distributions $\text{Dir}(\beta^{(c)})$ while common topics $\phi^{(c)}$ are drawn from a stream-independent Dirichlet distribution $\text{Dir}(\beta^{(c)})$. Each document d in a stream s, has an associated Bernoulli distribution with parameter $\eta_{d,s} \sim \text{Beta}(\gamma_s^{(s)}, \gamma_s^{(c)})$, indicating how likely the document is to choose local rather than common topics. For convenience we let $\eta_{d,c}$ (where $\eta_{d,c} = 1 - \eta_{d,s}$) represent how likely a document d is to choose common topics. The random variable $x_{d,i} \sim \text{Bernoulli}(\eta_{d,s})$ takes on one of the two values "local" or "common" for each word position i in document d. In addition, each document has two multinomial distributions with parameter vectors $\theta_d^{(s)} \sim \text{Dir}(\alpha_s)$ and $\theta_d^{(c)} \sim \text{Dir}(\alpha_c)$ over T_s and T_c respectively, where α_s and α_c represent the two Dirichlet parameter vectors. The document generation process associated with this model is as follows:

1. For all common topics T_c, draw $\phi^{(c)} \sim \text{Dir}(\beta^{(c)})$
2. For a particular stream s
   (a) For all local topics T_s, draw $\phi^{(s)} \sim \text{Dir}(\beta^{(s)})$
   (b) For each document d in s
      i. Draw Bernoulli parameter $\eta_{s,d} \sim \text{Beta}(\gamma_s^{(s)}, \gamma_s^{(c)})$
      ii. Draw $\theta_d^{(s)} \sim \text{Dir}(\alpha_s)$
      iii. Draw $\theta_d^{(c)} \sim \text{Dir}(\alpha_c)$
      For each word position i in document d
         A. Draw $x_{di} \sim \text{Bernoulli}(\eta_{s,d})$
         B. Draw a topic $z_{di} \sim \text{Multinomial}(\theta_d^{(x_{di})})$
         C. Draw a word $w_{di} \sim \text{Multinomial}\ (\phi_{z_{di}}^{(x_{di})})$

3.2 Inference via Collapsed Gibbs Sampling

In order to estimate the hidden parameters in the model, we apply collapsed Gibbs sampling using the following updating rules:

$$p(x_{di} = s, z_{di} = t) \propto \frac{c_{d,s-i} + \gamma_s^{(s)}}{N_d + \gamma_s^{(s)} + \gamma_s^{(c)} - 1} \frac{m_{d,z-i} + \alpha_z}{\sum_{z \in T_s} m_{d,z-i} + \alpha_z} \frac{n_{z,w-i} + \beta_w^{(s)}}{\sum_w^V n_{z,w-i} + \beta_w^{(s)}}$$
$$p(x_{di} = c, z_{di} = t) \propto \frac{c_{d,c-i} + \gamma_s^{(c)}}{N_d + \gamma_s^{(s)} + \gamma_s^{(c)} - 1} \frac{m_{d,z-i} + \alpha_z}{\sum_{z \in T_c} m_{d,z-i} + \alpha_z} \frac{n_{z,w-i} + \beta_w^{(c)}}{\sum_w^V n_{z,w-i} + \beta_w^{(c)}} \quad (1)$$

where $c_{d,s-i}$ is the number of words in document d assigned to local topics (excluding w_{di}), $m_{d,z-i}$ is the number of words in document d assigned to topic z (excluding the current one) and $n_{z,w-i}$ is the number of occurrences of term w assigned to topic z (excluding the current one). By using the samples from Gibbs sampling, parameters $\{\theta_d^{(s)}, \theta_d^{(c)}\}$, $\{\phi_s, \phi_c\}$ and $\{\eta_{d,s}, \eta_{d,c}\}$ can be effectively estimated as follows:

$$\theta_{d,z}^{(x)} = \frac{m_{d,z} + \alpha_z}{\sum_{z \in T_x} m_{d,z} + \alpha_z} \quad , x \in \{s, c\} \quad (2)$$

$$\phi_{z,w}^{(x)} = \frac{n_{z,w} + \beta_w}{\sum_{z \in T_x} n_{z,w} + \beta_w} \quad , x \in \{s, c\} \quad (3)$$

$$\eta_{d,x} = \frac{c_{d,x} + \gamma_s^{(x)}}{N_d + \gamma_s^{(s)} + \gamma_s^{(c)}} \quad , x \in \{s, c\} \quad (4)$$

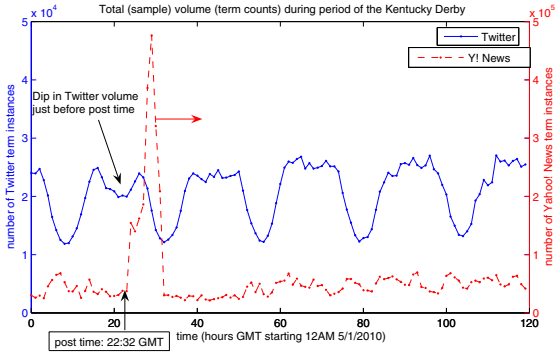

Figure 1: This figure shows the total volume (term instance counts) of Twitter and Yahoo! News as functions of time over the first 120 hours (GMT) of May 2010.

The formalism of our model resembles in spirit that of Chemudugunta et al. [6] where each term is "split" into corpus-level "background" topics, document-level "special" topics and normal topics. However, their work is only for a single corpus while our model fits multiple collections. Hyper-parameters like β, α and γ can be estimated using standard methods introduced by Minka [19].

4. MODELING TEMPORAL DYNAMICS

4.1 Temporal Dynamics for Topics

In this section we review a temporal model for news articles, introduced in [16] and present an alternate derivation. Before proceeding however, it bears pointing out that, as stated in [16], "rigorous analysis of the proposed model appears to be quite complex." The referred-to model embodies two driving forces for news-article publishing which the authors refer to as *imitation* and *recency*. The authors assert that this pair constitutes a minimum set for the purpose of explaining the temporal dynamics $n(t)$ of news-article publishing, but that the real situation is undoubtedly more complicated. We agree that there are factors beyond just these two. For example consider the Twitter and Yahoo! News total-volume data plotted in Figure 1. Note the enormous surge in the volume of news articles beginning after the Kentucky Derby[4], the premier American thoroughbred-horse race, and continuing for several hours. This clearly demonstrates significant elasticity in the volume capacity of the various sources contributing to Yahoo! News.

We start by assuming the following setting, which is a response that looks like a *proportional controller*[5] except that the "control point" $n_{\max}$ is not a constant.

$$\frac{dn}{dt} = \lambda\,[\,n_{\max} - n(t)\,], \quad (5)$$

where $n_{\max}$ is a function of both t and n. The form (5) captures the *saturation* effect mentioned above. The saturation value $n_{\max}$ varies with both n and t, however. We assume that it is the product of a term $(\zeta n(t))$ embodying the *imitation* effect mentioned above and a term (νt^{-1}) embodying the *recency* effect, where ζ and ν are adjustable parameters. Substituting the resulting expression for $n_{\max}$ into (5), we obtain:

$$\frac{dn}{dt} = \lambda\, n(t)\,[\,\zeta \nu t^{-1} - 1\,] \quad (6)$$

[4] `http://en.wikipedia.org/wiki/Kentucky_Derby`
[5] See `http://en.wikipedia.org/wiki/Proportional_control`.

Figure 2: Overall Algorithm

Initialize Gibbs Sampler
while Not Converging **do**
  E-step
  For all documents in all text streams, update topic assignments using (1)
  M-step
  Update α, β and γ values through the method introduced in [19]
  for Each local and common topic **do**
    1) Fit gaussian function to α values
    2) Fit "temporal gamma" function by using the parameters from the previous step
    3) Re-calculate α values for topic k by using fitted function
  end for
end while

Next we solve this differential equation, assuming that the event occurs at $t = 0$ for convenience. For an event occurring at time t_0 let $t \rightarrow t - t_0$. We must also ensure that our solution satisfies the following boundary conditions.

1. $n(t) = 0$ for $t \leq 0$.
2. $n(t) \geq 0$ for $t > 0$.
3. $n(t) \rightarrow 0$ as $t \rightarrow \infty$.

The solution of (6) proceeds in the following steps.

$$\begin{aligned} \frac{1}{n}\frac{dn}{dt} &= \lambda\,[\,\zeta\nu t^{-1} - 1\,], \\ \int_1^t \frac{1}{n}\frac{dn}{dt}\,dt &= \lambda\zeta\nu \int_1^t t^{-1}\,dt - \lambda\int_1^t dt, \\ \ln n(t) &= \ln n(t=1) + \lambda\zeta\nu \ln t - \lambda t + \lambda, \\ \ln n(t) &= \ln A + q \ln t - \lambda t, \\ n(t) &= A\,t^q\,e^{-\lambda t}, \end{aligned} \tag{7}$$

where $A := n(t = 1)\,e^{\lambda}$ and $q := \lambda\zeta\nu$. Next we apply our boundary conditions to the solution given in (7). First, to enforce condition 1 we multiply the solution of (7) by the Heaviside unit step function $u(t)$, which equals 0 for $t < 0$ and 1 for $t > 0$. Thus, we have

$$n(t) = u(t)\,A\,t^q\,e^{-\lambda t}. \tag{8}$$

Condition 2 requires that $A > 0$ and Condition 3 requires that $\lambda > 0$. The form of (8) has been demonstrated to capture spikes of news articles and social-media blogs [16, 30].

4.2 Incorporating Temporal Dynamics

In this section we describe how to incorporate the temporal model described above into our `Collection Model` and then introduce the inference approach to estimate the parameters in the model. We assume that the temporal dynamics of each topic are independent of each other. In other words, the popularity of one topic does not affect that of the other topics. We realize that this is a simplified assumption. The basic intuition behind embeding temporal dynamics into the model is to allow certain topics to have a higher probability of being selected. For example, during the Soccer World Cup in June and July of 2010, news articles and Twitter messages may naturally be more likely to talk about the World Cup, rather than politics. We encode this notion by associating the Dirichlet parameters for each topic with a time-dependent function. This function governs the variation of those parameters and thus indirectly controls the popularity of the associated topics. More specifically, for all common topics (with parameters α_c) and local topics (with parameters α_s), we let each dimension α_k in Dirichlet parameters α to be associated with the following time-dependent function.

$$\alpha_k(t) = f_k(t) = A_k t^{q_k} e^{-\lambda_k t} \tag{9}$$

where $f_k(t)$ is the temporal model described in Section 4.1. However, if we naïvely associate α_k with f_k, the model may face difficult problems since the temporal model unrealistically assumes that the starting point of the time t for all topics is time stamp 0. In other words, different topics should have different starting times t_0. Thus, we modify it into the following form:

$$f_k(t) = C_k + u(t - t_0^k)(A_k|t - t_0^k|^{q_k} \exp(-\lambda_k|t - t_0^k|)) \tag{10}$$

where t_0 is the starting time stamp of the topic, A_k controls the height of the prior knowledge, q_k indicates how quickly the topic would rise to the peak, λ_k controls the rate of decay and C_k is the "noise" level of the topic. We refer to the right hand side of (10) as the "temporal gamma function".

The absolute-value function guarantees that the time-dependent part is only active when t is larger than t_0. Additionally, $u(t - t_0)$ is a step function that is 1 for $t \geq t_0$ and 0 otherwise. In our implementation, a "soft" version of the step function as $u(t - t_0^{(k)}) = 1/(1 + \exp(-(t - t_0^{(k)}))$ is used. Intuitively, this equation states that the prior knowledge of each topic is fixed over time (by the "noise" level C_k) until a starting point t_0 and from that point on it follows a temporal gamma function controlled by three parameters, A_k, q_k and λ_k. The crux of the problem is to estimate the values of these five parameters from the data.

The absolute-value function and the parameter t_0 in (10) present challenges to model fitting (parameter estimation). To address this, rather than directly fitting $f_k(t)$, we use the following heuristic similar to that used in [18]. We first fit the following Gaussian function:

$$\alpha_k(t) \approx g_k(t) = C_k^{'} + A_k^{'} \exp(-(t - \mu_k)^2/2\sigma_k^2), \tag{11}$$

where μ_k is the mean and σ_k^2 is the variance. The resulting parameter values are then used to obtain initial parameter values for fitting the temporal gamma function of (10). This Gaussian function is straightforward to fit and its symmetric form allows us to obtain t_0 easily. We set the initial values of C_k and A_k in (10) to those obtained by fitting the Gaussian function and we fix $t_0^{(k)} = \mu_k - \sigma_k$. This process simplifies our inference algorithm. Note that the Gaussian approximation is only used to find initial values of the parameters including t_0. In our later experiments we find that this approximation gives reasonable initial values.

The outline of our inference algorithm is shown in Figure 2. Overall, we incorporate the functional optimization problem with Gibbs sampling into a stochastic EM framework (e.g., similar to [7]). In the E-step we gather topic assignments and useful counts by Gibbs sampling through (1). In the M-step we optimize the proposed objective functions to obtain the updated hyper-parameters for the next iteration. More specifically, the first step is to estimate the Dirichlet parameters α from counts obtained from Gibbs Sampling. This can be done in several ways [19]. We use Newton's method in this step. The second step is to use these α values to fit the Gaussian function (11) and then, using the parameters from the fitted Gaussian function as initial values, to fit our temporal gamma function (10). For both problems we minimize the following objective functions:

$$\underset{\boldsymbol{g_k}}{\operatorname{argmin}}\, G_k = \frac{1}{2}\sum_t \Big(\alpha_k(t) - g_k(t)\Big)^2 \tag{12}$$

$$\underset{\boldsymbol{f_k}}{\operatorname{argmin}}\, F_k = \frac{1}{2}\sum_t \Big(\alpha_k(t) - f_k(t)\Big)^2 \tag{13}$$

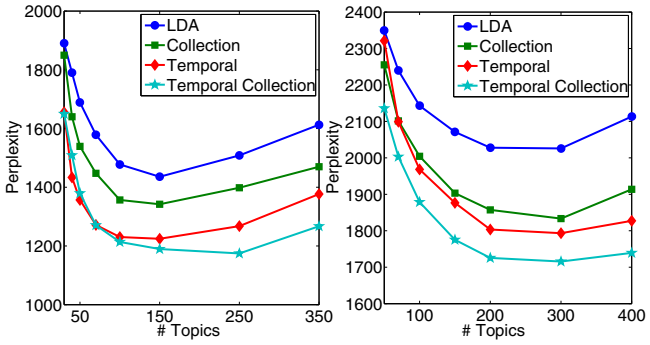

Figure 3: Perplexity Comparison Between Multiple Models. Left-hand plot: random 80%/20% train/test split; Right-hand plot: past/future 20-days/10-days train/test split.

We use the L-BFGS algorithm[17] implemented in GNU/GSL Library[8], which only requires the first-order gradients to obtain the optimal values for the parameters in both functions. Note that the method proposed here is potentially scaleable to very large datasets. For example LDA-style Gibbs sampling has been scaled to very large dataset sizes by [21], which is particularly useful for our E-step. For our M-step a stochastic gradient descent can be used instead of the usual Newton's method. We denote the whole algorithm as the `Temporal Collection` model.

5. EVALUATION

We utilize a real-world dataset consisting of Yahoo! News and Twitter messages from May 2010 to evaluate our method. Since the original dataset is quite large, we sample news articles and Tweets proportional to the total volume of each hour in May, resulting in 233,488 news articles and 1,736,350 Twitter messages in total. We use each hour as a time unit, which starts from 0, the first hour of May 1, 2010 to 720, the last hour of May 30, 2010. All the experiments are based on this dataset. The models used are (1) Latent Dirichlet Allocation (`LDA`), (2) Correlated Stream Model (`Collection`), introduced in Section 3, (3) Temporal Dynamics Topic Model (`Temporal`), introduced in Section 4.2 but ignoring multiple collection effects, and (4) Correlated Collection Model with Temporal Dynamics (`Temporal Collection`), introduced in Section 4.2. For `Collection` and `Temporal Collection`, we set the number of common topics to 20 $\sim$ 50 (depending on the total number of topics) and equally divide the remaining topics into all other streams, as local topics. We do not compare with other similar methods because `Collection` and `Temporal Collection` can essentially represent the two major directions of previous work discussed in Section 2.

5.1 Perplexity Evaluation

Following common practice for comparing topic models, we use *perplexity* of the held-out test data as our goodness-of-fit measure. Perplexity is defined as $\exp\Big(-\frac{\sum_{d=1}^{D}\sum_{i=1}^{N_d}\log p(w_{d,i}|\mathrm{M})}{\sum_{d=1}^{D}N_d}\Big)$ where $w_{d,i}$ represents the i^{th} term in document d, M is the model and N_d is the number of words in document d. First, we randomly sample 80% of the data as the training data and use the remaining 20% as the test data. Although this is a common evaluation procedure for topic models, it may not reflect real-world scenarios for temporal text collections because it may give additional undesirable advantages to models knowing the "future." All models are trained on the same training set and evaluated using the same test set. In the training phase we obtain topic distributions ϕ and all other hyper-parameters. In the testing phase we fix them and perform 100 Gibbs-sampling iterations for each document in the test set, obtaining θ_d. Using these newly estimated θ_d, we calculate $p(w_{d,i}|\mathrm{M}) = \sum_z^K \phi_{k,w}\theta_{d,k}$ and then compute perplexity. The result is shown on the left-hand side in Figure 3. The second setting we choose is closer to real-world scenarios. We train all models on the first 20 days in May and test the perplexity on the remaining 10 days, shown on the right-hand side in Figure 3. As is evident in the figure, the perplexity exhibits a minimum with respect to the number of topics in both settings. As the number of topics is increased beyond that minimum, overfitting appears to set in, as was also observed in [10]. For both settings `Temporal Collection` significantly outperforms the others.

5.2 Common Topics and Local Topics

Here we manually compare the topics obtained by `Temporal Collection` and by `LDA` to determine which topics are meaningful and to see if any interesting patterns are discovered by the model. As we described previously, the advantage of `Temporal Collection` is to identify common topics among multiple text collections in a principled manner. Since `LDA` does not provide any mechanism for retrieving common topics explicitly, we use the following heuristic ranking method to indicate the prevalence of a topic T on both News and Twitter: $\frac{1}{2}\Big[\frac{n(z_T,\text{News})}{\sum_{T'} n(z_{T'},\text{News})} + \frac{n(z_T,\text{Twitter})}{\sum_{T'} n(z_{T'},\text{Twitter})}\Big]$ where $n(z_T,\text{News})$ is the number of tokens assigned to topic T in News and $n(z_T,\text{Twitter})$ is the number of tokens assigned to the same topic in Twitter. Basically, this simple heuristic measures how likely a topic is to be assigned to a token in both News and Twitter on average. The higher this value is, the more likely this topic will appear, on average. We rank all the topics obtained by `LDA` through this method and show the top 5 on the left top part of Table 1. For `Temporal Collection`, since common topics are identified automatically, we just need to rank all common topics and extract the top ones, by the following criterion: $\frac{1}{2}\Big(\mathbb{E}[\theta_i^{\text{News}}]+\mathbb{E}[\theta_i^{\text{Twitter}}]\Big)$ where $\mathbb{E}[\theta_i^{\text{News}}]$ is the expected value of θ_i for common topic t_i on news and similarly for Twitter. This equation can be interpreted as the average of the expected value of topic k appearing in a document on both collections. The quantity $\mathbb{E}[\theta_i]$ can be easily computed by $\frac{\alpha_i}{\sum_k \alpha_k}$ and normalized across all time epochs. The top 5 common topics are listed on the right top part of the same table.

The first column and the third column of the Table 1 show the *title* of the topics, a label given by the authors for easier interpretation. All topics (the second and the fourth column) are represented by the top ranked terms by $\phi_{z,w}$. Note that all these models are fit in an unsupervised manner in which no explicit human labels are available beforehand. From the results it is clear that both methods rank some potential common topics highly, such as "Oil Spill" and "Financial Crisis". However, it is also noticeable that simple ranking heuristics may not give appropriate scores to the topics. For instance the ranking scheme may prefer the topics from a collection that is significantly larger than the other, even if a topic only appears in one collection. For example, the two "junk" topics shown on the left are examples of this situation. In addition, if two topics are common to both data collections but one is popular among a lot of short documents (e.g., Twitter messages) and the other is prevalent in a relative small number of long documents (e.g., news articles), some sort of normalization schemes is clearly needed. Although there exist some sophisticated ranking heuristics

Table 1: Example Topics from Our Dataset

Comparison of Top Ranked Common Topics between LDA (Left) and Temporal Collection (Right)			
Title	**Top Terms**	**Title**	**Top Terms**
"finance"	percent billion bank market greece financial banks debt	"finance"	percent billion bank greece financial debt banks euro crisis
"crime"	police car times vehicle found york square street bomb	"oil spill"	oil gulf spill coast drilling mexico water louisiana
"junk"	link cont via #jobs #fb album super live wii #tcot #news	"world cup"	world cup team league final players south season club
"oil spill'	oil gulf spill coast mexico gas drilling sea water	"health care"	health medical care cancer hospital patients study research
"junk"	dont people cant thats youre bad look tell talk	"UK election"	minister party prime cameron political leader president
Comparison of Local Topics between News (Left) and Twitter (Right)			
News		**Twitter**	
Title	**Top Terms**	**Title**	**Top Terms**
"crime"	police car times vehicle found york square street	"social media"	blog video post check news via twitter online facebook
"US election"	election party law president vote political campaign	"hash tags"	#fb info #quote #fail #ge #lol #ff #twibbon cont
"China"	minister china south india north chinese korea indian	"non-English"	les pas pour sur une cest est qui avec bien suis tout faire
"jobs"	budget tax million money pay bill federal increase cuts	"junk"	cant this wait watch next believe gonna watching just
"education"	school students schools board education district college	"junk"	that would have could never were wish there

[3], we argue that our model can handle these issues in a more principled way by modeling common topics explicitly. For instance the α values for common topics can shed light on how popular these topics are, either in one of the data collections or in all of them.

Since similar ranking heuristics do not work well for LDA to provide local topics for Twitter and news, we only report the local topics found by our method, shown in the lower part of Table 1. On the left-hand side top ranked local topics on news are presented while on the right hand side top local topics found on Twitter are shown. Interesting observations can be made based upon these results. First, news articles tend to have more "formal" topics, such as politics, education and economy, whereas a large fraction of the Twitter stream consists of personal chat and opinions. Therefore, besides the common topics (e.g., Table 1) in both news and Twitter, local topics for Twitter seem less understandable and coherent. Indeed, throughout several experiments conducted on May's data and on other months as well, we observe that most of the local topics on Twitter are not very interesting. On the other hand, based on our experiments, some local topics (e.g "Crime") are on news but seldom picked up on Twitter. Many different kinds of criminal incidents are reported on a variety of news sources but not many of them really trigger interest on Twitter. Note that we understand that these results are preliminary and more thorough experiments are required. Nevertheless, our method provides a tool to investigate these interesting phenomena which are difficult or were impossible to be examined before [33].

5.3 Case Study on A Common Topic

Besides finding common and local topics on news and Twitter, our model also provides information about the temporal dynamics of these topics. Here we take a topic related to "Kentucky Derby" as an example to show the usefulness of our method. The Kentucky Derby[6] is the premier annual American horse race and has a significant international following. In 2010 it took place on May 1st. We try to identify the topics related to this event from the results obtained by our model. Remember that topics are only distributions over words. In order to find potential topics, we check the ranking positions of a list of terms which are known to be related to the event (e.g., "horse", "race", "kentucky", "derby"). If these terms are ranked highly in a particular topic, we consider that topic to be about the Kentucky Derby. We list the top 5 ranked terms of the topic we found by this simple heuristic just described: "derby", "race", "borel", "kentucky" and "horse". The topic we matched is a **common** topic and therefore it has the same distribution over words for both News and Twitter, meaning that once an article in News or a message in Twitter refers to this topic, the same word distribution is used to generate words, which is guaranteed by the model. However, the difference between News and Twitter on how this topic would be selected in a document is controlled by a stream-specific prior $\sim \text{Dir}(\alpha_t^{(c)})$ and further governed by a stream-dependent temporal gamma function.

In order to show the time series of the topic on news and Twitter, we transform the counts into a valid distribution by calculating a $p(t|z) = \frac{p(z|t)p(t)}{\sum_{t'} p(z|t')p(t')}$ where t is a time epoch. Then, $p(z|t)$ is estimated by the number of tokens assigned to topic z in time epoch t divided by the total number of tokens in time epoch t and $p(t)$ is estimated by the total number of tokens in time t divided by the overall number of tokens across all time epochs. Basically, the probability $p(t|z)$ tells us how likely the topic would appear in time epoch t. The results are shown in Figure 4. We first show the topic on the whole timeline (720 hours in May) on the top and show the first 120 hours at the bottom of the figure. The first observation is that the topic has two major peaks on both news and Twitter, shown in the upper part. This may reflect that "Kentucky Derby" is indeed a popular sports event. From the first 120-hour view of the topic, it is interesting to see that the topic first exhibited a peak on News and exhibited another peak on Twitter several hours later. This is a concrete example demonstrating the potential usage of our model to analyze common topics on multiple text streams in the timeline. A similar kind of analysis is conducted in [33] using sophisticated heuristics to find common topics and to view the timelines of topics.

5.4 Case Study on Hashtags

Hashtags, a type of community convention[7] which starts with a "#" sign, have been heavily used as annotations to represent events and topics on Twitter. We select several hashtags that can act as indicators for certain events where each hashtag is clearly associated to some events in May, 2010. More specifically, we choose #mothersday for "Mothers Day", #memorialday for "Memorial Day", #bp for "Oil Spill",#kentuckyderby for "Kentucky Derby", #gaga for "Lady Gaga" and #justinbieber for "Justin Bieber". We wish to see whether these events can be discovered by different models and how well these topics can be presented. We believe these hashtags represent a large range of social events and therefore are representative. In order to make a fair comparison, we

[6] http://en.wikipedia.org/wiki/Kentucky_Derby

[7] http://twitter.pbworks.com/Hashtags

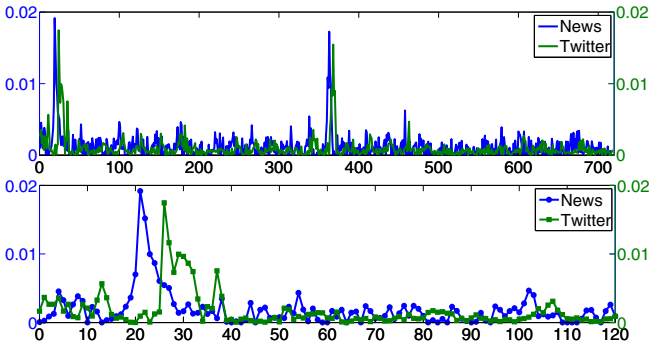

Figure 4: Temporal dynamics of "Kentucky Derby" on News and Twitter (X-axis is the hours in May, 2010. Y-axis is $p(t|z)$).

Table 2: Hashtag-to-Topic Mappings

Hashtag	Top Terms of Mapped Topic
[a] Hashtag Mapping for LDA model	
#mothersday	family home life children mother son friends
#memorialday	event june call center community club park
#bp	oil gulf spill coast mexico gas drilling
#kentuckyderby	race car track kentucky win top cars
#gaga & #justinbieber	justin lady super try bieber ider rio gaga jonas
[b] Hashtag Mapping for Temporal Collection model	
#mothersday	family children day home life church mother
#memorialday	memorial event day june community center
#bp	oil gulf spill coast drilling mexico water louisiana
#kentuckyderby	derby race borel kentucky horse super
#gaga & #justinbieber	bieber music video song gaga album lady

[c] KL Divergence between Hashtags and Matched Topics	
Hashtag	LDA vs. Temporal Collection
#mothersday	1.1911 / 0.7714
#memorialday	1.4331 / 0.9365
#bp	0.3958 / 0.1577
#kentuckyderby	1.9924 / 0.8183
#gaga & #justinbieber	2.2391 / 1.1754

transform the volume of these hashtags over time into distributions by using a technique similar to those introduced above. The first question we want to ask is whether the models can identify topics that reflect the events behind these hashtags. We map hashtags onto the topics obtained by the models and top ranked terms in these topics are examined to see whether these terms have any relationships with the underlying events. To map the hashtags, we calculate the following probability $p(z|w) = \frac{p(w|z)p(z)}{\sum_{z'} p(w|z')p(z')}$ where $p(w|z)$ is exactly $\phi_{z,w}$, provided by the trained models and $p(z)$ can be easily estimated by the counts. Intuitively, this probability tells us how likely a topic is to be selected, given the term. For the Temporal Collection model, all topics (including common topics and local topics) are treated as candidates to be matched. We map hashtags to topics for both LDA and our model, shown in the upper part of Table 2. Both models map #gaga and #justinbieber together onto a single topic, indicating that topics obtained by these models do not strictly correspond to real world events. Although some top ranked terms are similar for both models, the results from the Temporal Collection model are arguably better, in terms of *interpretation* of these terms. For instance, the Temporal Collection model explicitly ranks terms "memorial" and "day" highly in the list, implying this topic has much closer relationship with "Memorial Day", while LDA only has terms with broader connections with this kind of event. Similarly, the Temporal Collection model ranks more specific terms highly for 'Kentucky Derby" (e.g., "borel", "horse", "pletcher") while the topic obtained by LDA is essentially related to many races including "car races" and "horse races".

Table 3: Evaluation on Retrieval Performance

Method	MAP
TF-IDF	0.673
TF-IDF + Plain LDA	0.685
TF-IDF + Collection	0.703
TF-IDF + Temporal Collection	0.732

We can also compare the time-series of topics and hashtags to determine whether they are similar. The assumption is that, if they behave similarly on the timeline, the topics might be good choices for *explaining* the underlying events. Note that we are **not** seeking the **exact** match here since the topics have many more terms rather than a single hashtag and it may explain multiple events. Again, we transform the volumes into probabilities. We plot the time series of selected hashtags and the time series of selected topics in the same plots, shown in Figure 5. For each hashtag we compare its time series, obtained using LDA, with those obtained from our model. Although top ranked terms may look similar, the time series of these topics behave significantly differently. For LDA, because of the fixed Dirichlet hyper-parameter α over time, the models may give inappropriate "pseudo counts" for certain topics in the timeline. Indeed, one property of Dirichlet distribution can shed some lights on the observation: $\mathbb{E}[\theta_k] = \frac{\alpha_k}{\sum_{k'} \alpha_{k'}}$ where the expected value of θ_k, the proportion of topic k represented in a document, is the ratio of the Dirichlet parameter α_k over the sum of all α values. Since α values are fixed over time, this expected value will also be constant over time, leading to the fact that the topic assignments fluctuate around a certain value, though with variance, which is exactly shown in our experiments. This drawback of LDA may lead to difficulty in identifying the peaks of these topics. On the other hand, in our model, since the hyper-parameters α are controlled by the temporal gamma functions, the rise and fall of these values may give good hints for the model to assign topics to words, yielding better modeling temporal dynamics. Also, from the results, our models can better match the peaks of hashtags, indicating that the method can better reflect real events.

We can further compare these time series quantitatively. Since these time series are valid distributions over time, KL divergence is employed to measure their "distance" as follows: $\sum_t p(t|w_1) \log \frac{p(t|w_1)}{p(t|w_2)}$. KL divergence is non-negative and the smaller the value is, the more similar two distributions are. We calculate the KL divergence between hashtag time series and topic time series for both LDA and our model. The results are shown in the bottom part of Table 2. From the results it is obvious that the time series of topics obtained by the Temporal Collection model better match the corresponding hashtag time series, yielding lower KL divergence scores. This also validates the visual evidence from Figure 5.

5.5 Performance on Retrieval

As a further demonstration of the utility and effectiveness of our model, we apply it in a toy application that uses it as part of an information-retrieval relevance measure. For a query q and document d, the idea is to use the probability $p(q|d)$ that q was generated by d's generating model as a measure of the relevance of d to q. In a scheme similar to that used in [29] we use a relevance measure $S(d|q)$ that is a linear combination of $p(q|d)$ and a simple TF-IDF-

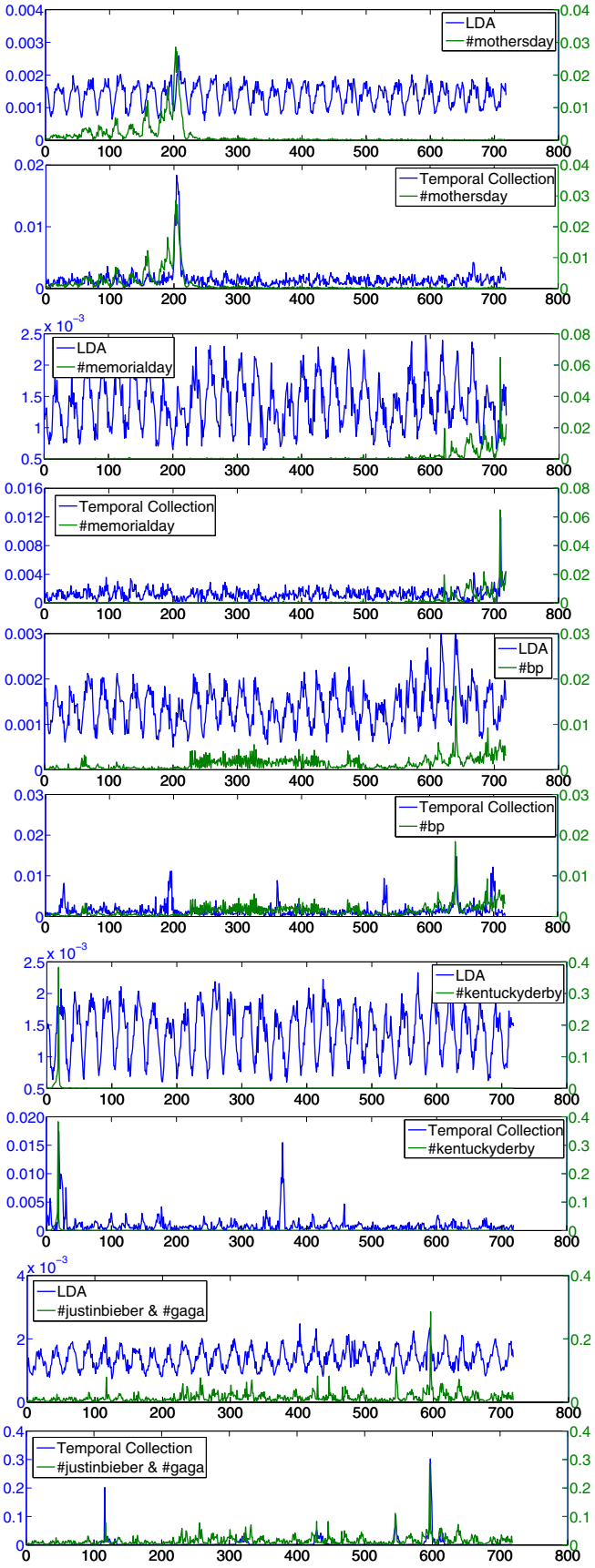

Figure 5: The distributions $p(t|z)$ of mapped topics in May. X-axis is the hour number. Y-axis is the probability.

based cosine-similarity score $\tau(d, q)$. That is

$$S(d|q) \;=\; \lambda\,\tau(d,q) \;+\; (1-\lambda)\,p(q|d). \qquad (14)$$

For our experiment we select the top 20 queries from `GoogleInsights` in the time period of May 2010, corresponding to our datasets. To select retrieval candidates we compute $\tau(d, q)$ for all tweet-query pairs (d, q) and use these scores to rank the tweets for all queries. We then select the top 50 tweets from that ranking for each query. These tweets and queries are then submitted to Amazon Mechanical Turk for manual relevance judgements, which we use as ground truth. These judgements are assigned using a three-level scale consisting of "relevant", "neutral" and "non-relevant." For each pair (d, q) three judges are assigned to assess the relevance and only the pairs on which at least two workers agree are kept, leaving a total of 922 tweets.

Mean Average Precision (MAP) for the top 20 positions is used for retrieval-accuracy characterization. These top-20 MAP scores are computed for each of four combinations of the TF-IDF measure $\tau(d, q)$ and a topic model. For each combination the parameter λ of (14) is varried over the range $[0, 1]$ and the optimal (highest MAP) value is determined. The corresponding MAP values are shown in Table 3. From these results we see that the choice of topic model used affects retrieval accuracy, with the highest retrieval accuracy being associated with the combination of TF-IDF and `Temporal Collection` scores.

6. CONCLUSION & FUTURE WORK

Modeling the temporal dynamics of topics is still a challenge, especially on multiple data collections. In this paper we propose a model for use in automatically analyzing multiple correlated text streams with their temporal behavior in a principled way. Our method bridges the recent advances in topic-modeling and information cascading in social media. We extend topic models by allowing each text stream to have local topics and shared topics, overcoming several theoretical problems of previously proposed models for similar problems. For temporal modeling we associate each topic with a time-dependent function that characterizes its popularity over time. By combining the two models, we can effectively model the temporal dynamics of multiple correlated text streams in a unified framework. Compared to related work our method is easy to implement and can potentially scale to large datasets. Additionally our method provides a new tool for browsing and mining a variety of types of social media simultaneously. For future work it will be interesting to utilize Bayesian non-parametric techniques to automatically learn the number of topics from the dataset. This is especially valuable for our model where the number of common topics and local topics must be manually assigned in current settings. In addition in order to better reflect real events, topics can be linked with named entities such that each topic is forced to contain a certain number of entities. It is also interesting to see hierarchical modeling of topics with temporal dynamics, which permits users to "zoom in" and "zoom out" on large topics (e.g. "oil spill") and track their evolution over time.

Acknowledgements

We thank Suju Rajan and Brian D. Davison for helping to improve the quality of the paper.

7. REFERENCES

[1] A. Ahmed and E. P. Xing. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th*

International Conference on Conference on Uncertainty in Artificial Intelligence (UAI), pages 20–29, 2010.

[2] A. Aji and E. Agichtein. Deconstructing interaction dynamics in knowledge sharing communities. In *International Conference on Social Computing, Behavioral Modeling, and Prediction*, pages 273–281, 2010.

[3] L. Alsumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 67–82, 2009.

[4] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 113–120, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248, 2006.

[7] G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 281–288, 2009.

[8] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi. *GNU Scientific Library Reference Manual - Third Edition (v1.12)*. Network Theory Ltd., 2009. `http://www.gnu.org/software/gsl/`.

[9] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[10] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, pages 5228–5235, 2004.

[11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[12] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 663–672, 2010.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD*, pages 56–65, 2007.

[14] J. Kleinberg. Bursty and hierarchical structure in streams. *Journal Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[15] J. Kleinberg. Temporal dynamics of on-line information streams. In *Data Stream Management: Processing High-Speed Data Streams*, 2005.

[16] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

[17] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[18] T. Masada, D. Fukagawa, A. Takasu, T. Hamada, Y. Shibata, and K. Oguri. Dynamic hyperparameter optimization for Bayesian topical trend analysis. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1831–1834, 2009.

[19] T. P. Minka. Estimating a Dirichlet distribution. Technical report, 2009. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`.

[20] R. M. Nallapati, S. Ditmore, J. D. Lafferty, and K. Ung. Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 520–529, 2007.

[21] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.

[22] M. Paul. Cross-collection topic models: Automatically comparing and contrasting text. Master's thesis, UIUC, 2009.

[23] M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1408–1417. Association for Computational Linguistics, 2009.

[24] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin. Hierarchical Bayesian modeling of topics in time-stamped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:996–1011, June 2010.

[25] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 579–586, 2008.

[26] X. Wang and A. McCallum. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.

[27] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 784–793, 2007.

[28] X. Wang, K. Zhang, X. Jin, and D. Shen. Mining common topics from multiple asynchronous text streams. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM)*, pages 192–201, 2009.

[29] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, 2006.

[30] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM International Conference on Web search and Data Mining (WSDM)*, pages 177–186, 2011.

[31] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 743–748, 2004.

[32] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1079–1088, 2010.

[33] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and traditional media using topic models. In *ECIR*, pages 338–349, 2011.