A nonparametric mixture model for topic modeling over time

Avinava Dubey* Ahmed Hefny* Sinead Williamson* Eric P. Xing*

Abstract

A single, stationary topic model such as latent Dirichlet allocation is inappropriate for modeling corpora that span long time periods, as the popularity of topics is likely to change over time. A number of models that incorporate time have been proposed, but in general they either exhibit limited forms of temporal variation, or require computationally expensive inference methods. In this paper we propose nonparametric Topics over Time (npTOT), a model for time-varying topics that allows an unbounded number of topics and flexible distribution over the temporal variations in those topics' popularity. We develop a collapsed Gibbs sampler for the proposed model and compare against existing models on synthetic and real document sets.

1 Introduction

Latent variable models, such as latent Dirichlet allocation (LDA, Blei et al., 2003), are popular choices for modeling text corpora. Documents are modeled as a distribution over a shared set of topics, which are themselves distributions over words. Each word in a document is assumed to be generated by one of these topics.

Most topic models assume that the documents are *exchangeable*, or in other words, that the order in which they appear is irrelevant. This is often not a reasonable assumption – the distribution over topics in today's newspaper is likely to be more similar to the distribution over topics in yesterday's newspaper than to the distribution over topics in a newspaper from a year ago. Similarly, popular topics on Twitter are likely to vary with both time and geographic location.

A number of models have been proposed to address this. Dependent Dirichlet processes (MacEachern, 1999) are distributions over collections of distributions, each indexed by a location in some covariate space (e.g. time), such that distributions that are close together in that space tend to be similar. Various forms of dependent Dirichlet process have been used to construct time-dependent topic models. Many of these models are limited in the form of variation obtained – for example the in the models of Lin et al. (2010) and Rao and Teh (2009) the probability of seeing a topic as a function of time is restricted to be unimodal. Moreover, these models are difficult to apply to higher dimensional spaces, and often rely on the discretization of time. More flexible models, such as those proposed by Srebro and Roweis (2005) and MacEachern (2000), tend to lose the desirable conjugacy properties of the corresponding stationary model, making inference challenging.

An alternative approach is seen in a model known as Topics over Time (TOT, Wang and McCallum, 2006). Unlike the previously discussed models, which define a distribution over topics conditioned on a time, TOT models the text and the time-stamp of a document jointly. This allows us to consider the time-stamp as a random variable, rather than a fixed parameter. Such a framework allows us to incorporate non-Markovian dynamics while maintaining reasonable inference requirements. It also means that we can incorporate data without covariate information (for example, documents with no time-stamp), something that is not easily achieved in conditional models such as dependent Dirichlet processes.

Like the conditional models, Topics over Time suffers from a number of shortcomings. The distribution over times for each topic is assumed to be unimodal, while in real life we often see topics vary in popularity in a more flexible manner. For example, Figure 1 shows the popularity of the search term "NHL" as a Google query. The popularity waxes and wanes with the hockey season, and occasionally peaks due to a major news event. In addition, the number of topics must be fixed a priori, which can involve expensive model comparison.

In this paper, we propose a nonparametric extension to the Topics over Time model (npTOT). This model extends TOT to allow an unbounded number of topics, each of which can peak in popularity an unbounded number of times. In addition, npTOT induces correlations between the temporal variations in topic popularity, so that related topics trend in similar manners. Because, like TOT, npTOT is a joint model of both text and time, document/time-stamp pairs can be considered exchangeable and we can make use of tractable exchangeable distributions to develop a Gibbs sampling scheme. We compare npTOT with its para-

*Machine Learning Department, Carnegie Mellon University, {akdubey, ahefny, sinead, epxing}@cs.cmu.edu

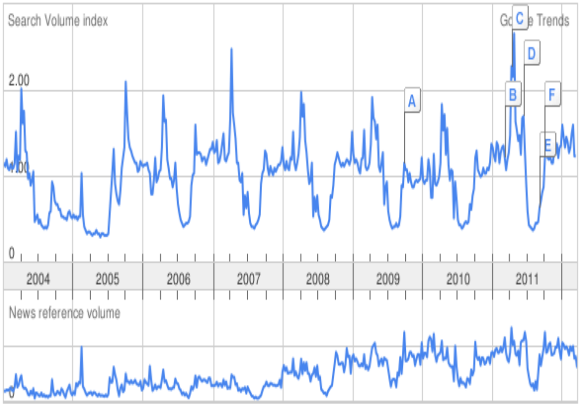


Figure 1: Search and news trends for ”NHL” obtained from Google Trends. This shows that interest in this topic rises and falls multiple times.

metric counterpart, plus several baselines, and show that the added flexibility translates into qualitatively and quantitatively better performance.

2 Related Work

Traditional topic models such as LDA have two main shortcomings. Firstly, they are parametric models that assume a fixed prespecified number of topics regardless of the data. Secondly, they assume that the probability of seeing a topic is independent of the time at which a document is written. In this section, we consider existing models that address one or both of these limitations.

2.1 Nonparametric topic models To relax the assumption of a fixed number of topics, nonparametric topic models have been proposed. Rather than the fixed, finite number of topics specified by LDA, such models allow a countably infinite number of topics a priori, meaning that a random number of topics will be used to represent a given dataset. The most widely used nonparametric topic model replaces the collection of Dirichlet distributions used to model the per-document distributions over topics in LDA with a hierarchical Dirichlet process (HDP, Teh et al., 2006). Here, the distribution over topics in a given document is given by a Dirichlet process. The document-specific Dirichlet processes are coupled using a shared base measure, which is itself a Dirichlet process.

2.2 Dependent Dirichlet processes The HDP assumes that the documents in our corpus are exchangeable. A class of models referred to as dependent Dirichlet processes (DDPs, MacEachern, 1999) relaxes this assumption. In topic models based on DDPs, each document is associated with a value in some covariate space, for example time. As in the HDP, the topic distribution of each document is marginally distributed according to a Dirichlet process. Unlike the HDP, documents that are close together in covariate space tend to have similar distributions.

A number of DDPs have been used in topic modeling. The recurrent Chinese restaurant process (Ahmed and Xing, 2010) creates a Markov chain of distributions; however the model is non-exchangeable so we cannot make use of conjugacy in inferring the topic proportions. In addition, the model is only applicable to covariate spaces of a single dimension. A number of related models (Caron et al., 2007; Lin et al., 2010; Rao and Teh, 2009) maintain some of the conjugacy of the original model, but do not allow as flexible variation in topic probability. A number of DDPs can exhibit more flexible, non-Markovian variation in topic probabilities (Srebro and Roweis, 2005; MacEachern, 2000), but inference in such models scales very poorly.

2.3 Topics over Time The DDP models mentioned in the previous section are examples of conditional models – the covariate is assumed fixed, and the model defines a distribution over topics conditioned on this covariate value. The Topics over Time (TOT, Wang and McCallum, 2006) model takes a different tack, assuming that the covariate values are also random, and that the latent topics describe a distribution both over words and over times. This model is exchangeable if we consider a data point to consist of both a document's text and its time-stamp, meaning we can make use of conjugacy.

TOT is a form of supervised LDA (Blei and McAuliffe, 2007), where the label is the time-stamp of the document. TOT assumes the following generative process for a corpus of documents and their associated timestamps:

1. For each topic $k = 1, \dots, K$
   (a) Sample a distribution over words $\phi_k|\beta \sim$ Dirichlet(β).
   (b) Choose a set of parameters ψ_k to parametrize a beta distribution.
2. For each document $j = 1, \dots, D$
   (a) Sample a distribution over topics, $\theta_j|\alpha \sim Dir(\alpha)$.
   (b) For each word $i = 1, \dots, N_j$
      i. Sample a topic indicator $z_{ji}|\theta_j \sim \theta_j$.
      ii. Sample a word $w_{ji}|z_{ji} \sim$ Mult$(\phi_{z_{ji}})$.

iii. Sample a time-stamp $t_{ji}|z_{ji} \sim \text{Beta}(\psi_{z_{ji}})$.

This model exhibits non-Markovian variations in topic probabilities, but has a number of drawbacks. The beta distribution used to model the time-varying probability is unimodal, and means that times must be bounded. This limits the form of temporal variation available, and precludes prediction outside of the bounded time-frame or extension to higher dimensionalities. Moreover, the lack of a prior on ψ_k means it must be estimated using an approximate method. In addition, the number of topics must be defined a priori.

A recent extension, Topics over Nonparametric Time (TONPT, Walker et al., 2012) breaks the assumption that the distribution over document times is bounded and unimodal by sampling the timestamps from a Dirichlet process mixture of Gaussians. While this allows a much more flexible distribution over the times at which a topic appears, it still assumes that the number of topics is given a priori. Moreover, it assumes that per-topic distributions over time are independent, which as we will describe in the next section, is a poor assumption. Finally, Walker et al. (2012) assume that all mixture components for each topic share a common variance. While this restriction is easily removed, it limits the form of temporal variability obtainable.

3 Nonparametric Topic Over Time (npTOT)

In this section we address the two problems identified in TOT: Inflexible topic probability variation, and a fixed number of topics. The resulting model employs nonparametric distributions to generate both the distribution over topics, and the distribution over timestamps; therefore, we refer to this model as nonparametric Topics over Time (npTOT).

We follow TOT in assuming that each document (indexed by j) consists of a (unordered) set of tokens (indexed by i). Each token $x_{ji} := (w_{ji}, t_{ji})$ is defined to be an ordered pair of a word w_{ji} and a time-stamp t_{ji} [1]. We assume that each document is generated by a distribution over multiple topics.

The restriction to a fixed number of topics can be avoided by replacing the Dirichlet distribution over topics with a hierarchical Dirichlet process. This allows an unbounded number of topics a posteriori, and ensures the topics are shared across documents.

The form of temporal variation can be modified by replacing the beta distribution in the TOT model with another choice of distribution. One way to model multimodal variation on an unbounded timeframe while maintaining tractable inference, as used by Walker et al. (2012), is to sample the timestamps from a Dirichlet process mixture of Gaussians. However, this ignores the possibility of correlations between the trending patterns of topics, something that is not addressed in much of the dynamic topic modeling literature. For example, topics to do with sports players and sports fans are likely to have similar temporal variation. We address this by allowing the components of our mixture of Gaussians to be shared between topics. This is achieved by sampling the mixture components from a hierarchical Dirichlet process.

Let GEM indicate the distribution over probability measures associated with the Dirichlet process. The generative process, represented by plate diagram 2, is defined as follows:

1. Sample a global base distribution over topic proportions, $J_0|\gamma \sim \text{GEM}(\gamma)$.
2. Sample a global base distribution over time component proportions, $L_0|\lambda \sim \text{GEM}(\lambda)$.
3. For each topic $k = 1, 2, \ldots,$
   (a) Sample a distribution over words, $\phi_k|\beta \sim \text{Dirichlet}(\beta)$.
   (b) Sample a topic-specific distribution over time components, $L_k|\alpha_1, L_0 \sim DP(\alpha_1, L_0)$.
4. For each time component $l = 1, 2, \ldots$
   (a) Sample a distribution over time, $(\mu_l, \sigma_l^2)|\Theta \sim$ Normal-inverse Gamma(Θ) (where Θ are fixed hyperparameters).
5. For each document $j = 1, \ldots, D$,
   (a) Sample a distribution over topics, $J_j|\alpha_0, J_0 \sim DP(\alpha_0, J_0)$.
   (b) For each word $i = 1, \ldots, N_j$,
      i. Sample a topic indicator $z|J_j \sim J_j$.
      ii. Sample a word $w_{ji}|\phi_{z_{ji}} \sim \text{Mult}(\phi_{z_{ji}})$.
      iii. Sample a time component indicator $\omega_{ji}|L_j \sim L_j$.
      iv. Sample a time-stamp $t_{ji}|\mu_{\omega_{ji}}, \sigma_{\omega_{ji}} \sim \mathcal{N}(\mu_{\omega_{ji}}, \sigma_{\omega_{ji}})$.

4 Inference

We propose a Gibbs sampler based on the Chinese restaurant franchise (CRF, Teh et al., 2006). Our model requires *two* restaurant franchises, one for the word HDP and the other for the time HDP. In the CRF interpretation, each word indicator z indexes a "dish" in a word-related restaurant franchise, and each time

[1] In practice, a document has a single time-stamp which we duplicate for each word during inference.

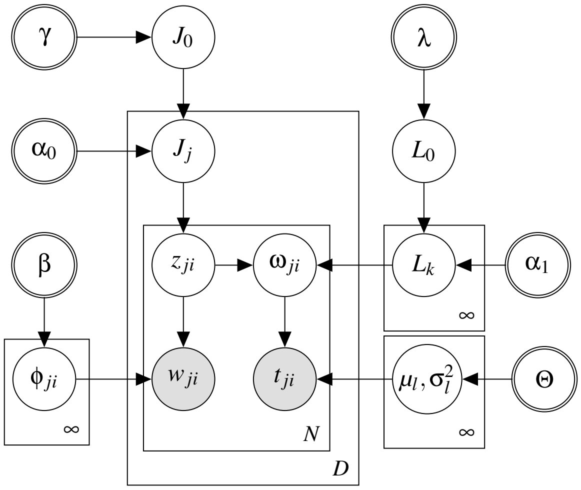


Figure 2: Plate diagram for nonparametric Topics over Time

component indicator ω indexes a dish in a time-related restaurant franchise.

The CRF associated with the word HDP mimics that described by Teh et al. (2006): each restaurant corresponds to a document and each dish corresponds to a topic. The CRF associated with the time HDP has a different interpretation: the restaurants correspond to the topics, and the dishes correspond to "time components", which are associated with a Gaussian distribution over time. We terms such as "time table", "time dish", "word table", and "word dish" to distinguish between the two franchises.

Each token $x_{ji} := (w_{ji}, t_{ji})$ is associated with a word table τ_{ji}^{word} and a time table τ_{ji}^{time}. Each word table a in document j is associated with a word dish (topic) d_{ja}^{word}. Each time table b in a topic k is associated with a time dish (time component) d_{kb}^{time}. The topic indicator for the ith word in the jth document is therefore $z_{ji} = d_{j\tau_{ji}^{word}}^{word}$, and similarly the time component indicator is $\omega_{ji} = d_{z_{ji}\tau_{ji}^{time}}^{time}$.

We define n_{ja} as the number of tokens in document j associated with word table a; m_k as the number of word tables serving word dish k; q_{kb} as the number of tokens associated with topic k and time table b; r_c as the number of time tables serving time dish c, and $f(v,k)$ as the number of times the word v is associated with topic k. We let K_+ be the current number of utilized word dishes, and C_+ be the current number of utilized time dishes. We use the notation $m_. = \sum_k m_k$.

At each iteration of our Gibbs sampler, we need to sample, for each token i in document j, both the corresponding word table τ_{ji}^{word} and time table τ_{ji}^{time}. We also need to sample the topic d_{ja}^{word} corresponding to each word table a in document j and the time component d_{kb}^{time} corresponding to each time table b in topic k. We describe these steps in detail in the remainder of this section.

4.1 Sampling τ_{ji}^{word} Recall that each word table is associated not just with a distribution over words, but also with a distribution over time tables. If we were to sample the word table for a token conditioned on that token's time table, our sampler would mix very slowly. Instead, we marginalize over τ_{ji}^{time} in order to sample τ_{ji}^{word}, and then sample τ_{ji}^{time} conditioned on τ_{ji}^{word} as described in Section 4.3.

The resulting distribution over time tables is given by

$$\begin{aligned} &p(\tau_{ji}^{word} = a|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{rest}_{-ji}) \\ &\quad\propto p(\tau_{ji}^{word} = a|\boldsymbol{\tau_{-ji}^{word}}) \\ &\quad\quad p(w_{ji}, t_{ji}|\tau_{ji}^{word} = a, \boldsymbol{x}_{-ji}, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \end{aligned} \tag{4.1}$$

where $\boldsymbol{rest}_{-ji} = (\boldsymbol{d^{word}}, \boldsymbol{d^{time}}, \boldsymbol{\tau_{-ji}^{word}}, \boldsymbol{\tau_{-ji}^{time}})$. The component terms of Equation 4.1 are given by

$$p(\tau_{ji}^{word} = a|\boldsymbol{\tau_{-ji}^{word}}) \propto \begin{cases} n_{ja}^{-ji} & \text{if } a \text{ is an existing table} \\ \alpha_0 & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} &p(w_{ji}, t_{ji}|\tau_{ji}^{word} = a, \boldsymbol{x}_{-ji}, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \\ =&\frac{(\beta + f(w_{ji}, k))}{V\beta + \sum_{v=1}^{V} f(v,k)} \cdot \\ &p(t_{ji}|\tau_{ji}^{word} = a, d_{ja}^{word} = k, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \end{aligned}$$

if a is an existing table serving dish k, or

$$\begin{aligned} &p(w_{ji}, t_{ji}|\tau_{ji}^{word} = a, \boldsymbol{x}_{-ji}, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \\ =&\sum_{k=1}^{K_+} \frac{m_k}{m_. + \gamma} \frac{\beta + f(w_{ji}, k)}{V\beta + \sum_{v=1}^{V} f(v,k)} \cdot \\ &p(t_{ji}|\tau_{ji}^{word} = a, d_{ja}^{word} = k, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \\ &+ \frac{\gamma}{m_. + \gamma} \frac{1}{V} \\ &p(t_{ji}|\tau_{ji}^{word} = a, d_{ja}^{word} = k_{new}, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \end{aligned}$$

if a is a new table. In both cases, we can write

$$\begin{aligned} &p(t_{ji}|\tau_{ji}^{word} = a, d_{ja}^{word} = k, \boldsymbol{t}_{-ji}, \boldsymbol{rest}_{-ji}) \\ =&\Bigg\{ \sum_b \frac{q_{kb}^{-ji}}{q_{k.}^{-ji} + \alpha_1} g_{d_{kb}^{time}}^{\boldsymbol{t}_{-ji}}(t_{ji}) \\ &+ \frac{\alpha_1}{q_{k.}^{-ji} + \alpha_1} \Bigg(\sum_{c=1}^{C_+} \frac{r_c}{r_. + \lambda} g_c^{\boldsymbol{t}_{-ji}}(t_{ji}) \\ &\qquad + \frac{\lambda}{r_. + \lambda} g_{c_{new}}(t_{ji}) \Bigg) \Bigg\}, \end{aligned}$$

where $g_c^{\boldsymbol{t}_{-ji}}$ denotes the posterior predictive time distribution for token x_{ji} conditioned on a time component c and other timestamps associated with that time component. For the Gaussian model described here, the posterior predictive distribution is a t-distribution.

If a new word table is created for token x_{ji} then we sample its corresponding word dish (topic) from the global word DP.

4.2 Sampling d_{ja}^{word} In order to resample the topic assignment d_{ja}^{word} for an entire word table, we need to marginalize over the time table assignments of *all* the tokens (denoted $\boldsymbol{x}_{ja} = (\boldsymbol{w}_{ja}, \boldsymbol{t}_{ja})$) associated with that word table. Since the number of tokens at the word table might be large, summing over all possible assignments is infeasible, so we approximate $p(\boldsymbol{t}_{ja}|d_{ja}^{word} = k, \boldsymbol{t}_{-ja}, \boldsymbol{rest}_{-ja})$ by sampling sets of table topic assignments. We use the resulting estimate $\hat{p}(\boldsymbol{t}_{ja})$ to approximate the true Gibbs sampling probabilities:

$$\begin{aligned} & p(d_{ja}^{word} = k|\boldsymbol{d_{-ja}^{word}}, \boldsymbol{w}, \boldsymbol{rest}_{-ja}) \\ & \propto \begin{cases} m_k^{-ja} p(\boldsymbol{w}_{ja}|d_{ja}^{word} = k, \boldsymbol{w}_{-ja})\hat{p}(\boldsymbol{t}_{ja}) & \text{existing topic,} \\ \gamma p(\boldsymbol{w}_{ja}|d_{ja}^{word} = k, \boldsymbol{w}_{-ja})\hat{p}(\boldsymbol{t}_{ja}) & \text{otherwise.} \end{cases} \end{aligned}$$

4.3 Sampling τ_{ji}^{time} and d_{kb}^{time} Given the topic assignments, the distribution over the timestamps is independent of the rest of the model, and we can perform inference using Gibbs sampling as in Teh et al. (2006).

5 Evaluation

The goal of this paper was to increase the flexibility of TOT, an existing joint model for documents and their timestamps, by allowing correlated multimodal variation in topic popularity, and by learning the number of topics. In this section, we present experimental results that demonstrate that we can capture more flexible variation than TOT, and learn an appropriate number of topic components. Moreover, we show that this added flexibility translates into improved log likelihood on test datasets.

5.1 Evaluation on Synthetic Data To demonstrate the ability of npTOT to recover the temporal variation of topics, we trained the model on a synthetic dataset, where ground truth is available. We generated a dataset of D documents from K topics, each associated with a multinomial distribution over V words obtained by discretizing Gaussian distributions with means sampled uniformly on $[0, V]$. Each topic is also associated with a continuous distribution over time, distributed according to a mixture of C Gaussians with means at $0.5 + k/K$, $k = 1, \ldots, K$. Each component has equal variance σ such that $3 * C * \sqrt{\sigma} = 1$. For each time-stamp we generate one document. Each document is associated with a distribution over topics which is proportional to the probability of that document generating the time-stamp. Topics and words were sampled according to the LDA generative procedure. We set D to 100, K to 30, V to 100 and C to 10. An example of a single topic and the corresponding distribution over times is shown in Figure 3.

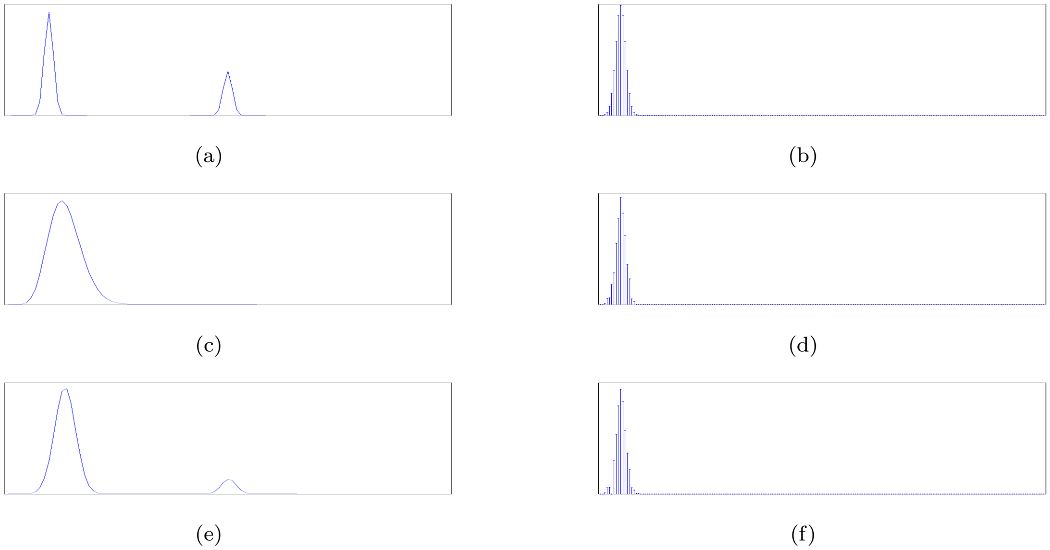
We trained TOT and npTOT on the generated data. The number of topics in TOT was set to the true number of topics, and as we see in Figure 3(b,d,e), the distributions over words obtained were a good match for the generating data. The npTOT model found 27 topics, very close to the true value. As we see in Figure 3(c), TOT was unable to capture the variation in topics. Conversely, npTOT was able to capture the multimodality of their distribution with respect to time (Figure 3(e)).

5.2 Real-world Data Experiments To show that npTOT is able to capture the temporal variation in real documents, we performed experiments on three datasets:

- **Twitter Subset.** This dataset consists of tweets originating from Egypt in the time period from January through March 2011. We selected tweets given by active users where an active user is a user who has more than 200 tweets. As preprocessing, we removed words that are less than 3 characters long. We then removed the most frequent 40 words as well as words that occurred less than 10 times. Finally, we aggregate the tweets of each user in each day in a single document and remove documents that are less than 20 words long. The preprocessed dataset contains 6,072 documents, 9,080 unique words and 324,298 word tokens in total. Because these tweets originated from Egypt, they contain both Arabic and English words.

- **State of the Union Address dataset.** The State of the Union dataset[2] contains the transcripts of 208 State of the Union addresses from 1790 to 2002. We followed Wang and McCallum (2006) in processing the dataset. Namely, we divided each speech into three-paragraph documents, and removed stop words and numbers. This resulted in 5,897 documents, 22,620 unique words and 800,399 word tokens in total.

[2] http://www.gutenberg.org/dirs/text04/suall11.txt


(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: (a) shows the actual distribution over time for a particular topic on the synthetic dataset, (b) shows the distribution over words for that particular topic, (c) shows the distribution over time of the TOT-detected topic closest to the original (d) shows the probability of the top-10 words for the TOT topics in (c), (e) shows the distribution over time for corresponding topic found by npTOT and (f) shows the word distribution for the npTOT topic in (e).

- **NIPS dataset.** The NIPS dataset[3] consists of the full text of the 12 years of proceedings from 1987 to 1999 Neural Information Processing Systems (NIPS) Conferences. The dataset is already preprocessed as described in Globerson et al. (2007) and it consists of 1,740 research papers, 13,946 unique words and 2,301,375 word tokens in total.

In addition to TOT, which we will refer to as TOT-Unimodal in this section, we evaluated against three other baselines:

- **LDA-Unimodal** Here we ran LDA on the text of the documents, and then fit the temporal variation of each topic with a single Gaussian distribution.

- **LDA-Multimodal** Here we ran LDA on the text of the documents, and then fit the temporal variation of each topic with a mixture of Gaussians.

- **TOT-Multimodal.** Here, we restricted npTOT to have a fixed number of topics, in order to disambiguate the effect of an unbounded number of topics from the effect of using a more flexible distribution over time. This is similar to the model described in Walker et al. (2012) but with component parameters drawn from an HDP rather than drawing the means from a DP and assuming a single variance per topic.

5.3 Real-world Data Experiments To show that npTOT is able to capture the temporal variation in these datasets, we performed a qualitative analysis of the topics found, and a quantitative analysis of the predictive performance of the models.

5.3.1 Qualitative analysis To see how npTOT can capture a wider variety of temporal variation than TOT, consider topics found using both models. Figure 4 shows topics found in the Twitter and the State of the Union addresses. We hand-picked topics that addressed the same themes for the purpose of this comparison. On the Twitter dataset, we see a topic that arises with the outbreak of revolution in Egypt on January, 25, 2011. Both models capture a sharp peak in this topic at that time, but the slow decay shown by the npTOT model is more realistic than the sharp decline in interest implied by the TOT model. On a subset of the State of the Union dataset, we show a topic concerned with conflict involving the US and Britain. Both models show a sharp peak in this topic around the time of the War of 1812, but the nonparametric model is able to reuse this topic to describe tensions between the US and Britain leading to the declaration of the war, such as the Embargo Act of 1807.

[3] http://cs.nyu.edu/ roweis/data.html

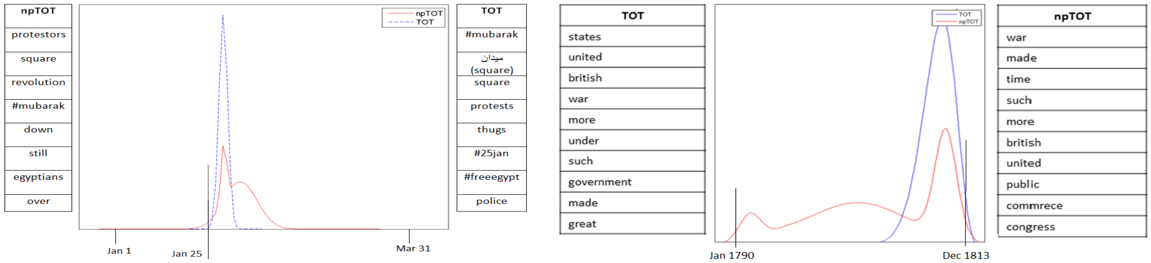


Figure 4: **Left:** Top eight most probable words in topics from TOT and npTOT corresponding to the Egyptian revolution (started on Jan 25) in Twitter dataset. **Right:** Top ten most probable words in topics from TOT and npTOT corresponding to conflicts involving the US and Britain in the State of the Union address dataset.

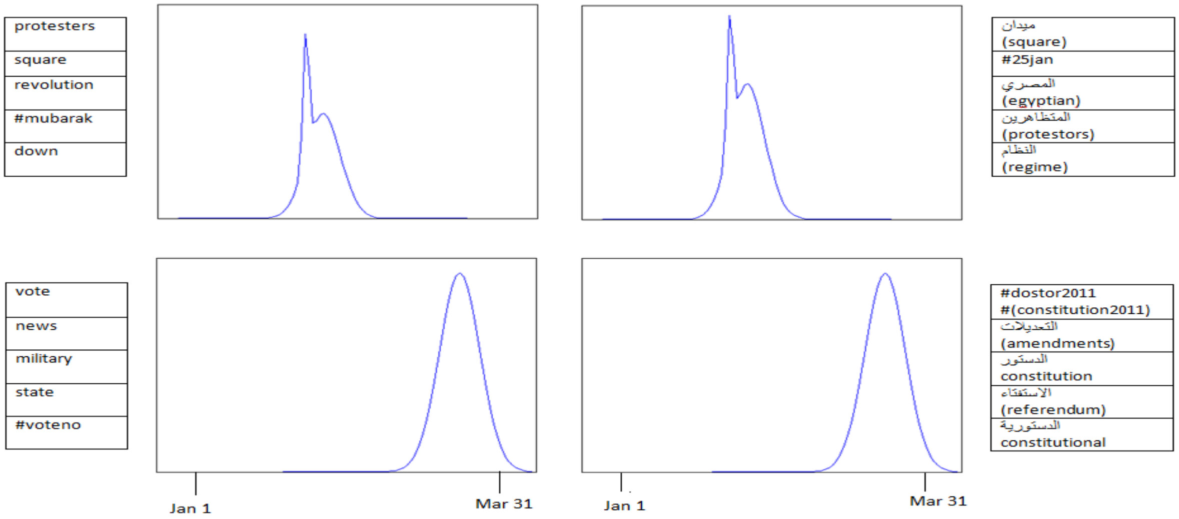


Figure 5: Distributions over time for two discovered English topics (left) and their Arabic counterparts (right), showing that they share time components.The top topic is about the Egyptian revolution outbreak. The bottom topic is about a referendum on constitutional amendments. Each panel shows words selected from the top twelve most probable words in the corresponding topic. Words in parentheses are translated from Arabic.

Figure 5 shows how related topics can share time components to give similar temporal variation. Since Twitter data is bilingual, we expect pairs of topics that address similar issues but in different languages. The figure shows two such pairs, demonstrating that they share the same time components. This behavior would not be exhibited if the component parameters are drawn from a DP (as in Walker et al., 2012).

5.3.2 Quantitative analysis We evaluated the performance of npTOT and its competitors using two methods: Joint likelihood of a document and its time-stamp, and perplexity of the second half of a document, conditioned on its time-stamp and the first half of the text. The joint likelihood gives a general measure of how well the various methods are able to model the corpora. The perplexity task demonstrates how well we are able to make use of temporal information to predict the content of a document.

In each case, we randomly split each dataset into training and test sets using a 70:30 split, and learned all four models on the training set. The LDA models were run for 1000 iterations, and npTOT and TOT were run until the percentage of changed tokens was below 5%. The joint log likelihood was obtained using the harmonic mean method, as described in Wallach et al. (2009), by sampling topic assignments $\boldsymbol{z}_d^{(s)} \sim \boldsymbol{z}_d|\boldsymbol{\Phi}, \boldsymbol{w}_d, \boldsymbol{t}_d$, (where $\boldsymbol{\Phi}$ denotes the estimated model parameters) and taking the harmonic mean of the conditional likelihoods $P(\boldsymbol{w}_d, t_d|\boldsymbol{z}_d^{(s)}, \boldsymbol{\Phi})$ over 200 samples.

We evaluated perplexity using the estimated θ method described in Wallach et al. (2009): for each test document d we sample topic assignments $\boldsymbol{z}_d \sim \boldsymbol{z}_d|\boldsymbol{\Phi}, \boldsymbol{w}_d^{(1)}, \boldsymbol{t}_d$, where $\boldsymbol{\Phi}$ denotes the estimated model

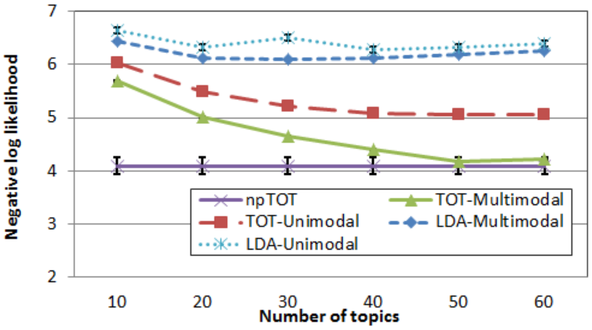


Figure 6: Average per-token negative log likelihood on test set for Twitter dataset.

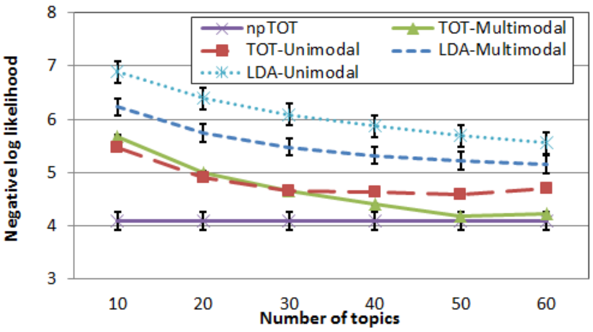


Figure 7: Average per-token negative log likelihood on test set for State-of-the-union-address dataset.

parameters, t_d and $\boldsymbol{w}_d^{(1)}$ denote time stamp of document d and the words in the first half of the document respectively. For each sample of $\boldsymbol{z}_d^{(s)}$ we estimate $\hat{\theta}_{dk}^{(s)} = P(k|\boldsymbol{z}_d^{(s)}, \alpha)$, which we use to estimate the likelihood of the second half of the document $P(\boldsymbol{w}_d^{(2)}|\boldsymbol{w}_d^{(1)}, \boldsymbol{\Phi}, \hat{\boldsymbol{\theta}}^{(s)}) = \prod_{i=1}^{N_d} \sum_k P(w_{di}|z_i = k, \boldsymbol{\Phi})\hat{\theta}_{dk}^{(s)}$. Taking the product over all document and then averaging over samples of $\boldsymbol{z}^{(s)}$ gives an estimate of the document completion likelihood. The perplexity score we report is evaluated as exp(−completion log likelihood/N), where N is the total number of words in the test set.

Figures 6, 7 and 8 show the resulting log joint likelihoods on the three datasets. In each case, npTOT gives the best likelihood, and the baseline LDA-Unimodal and LDA-Multimodal models perform poorly. The two TOT models, TOT-Multimodal and TOT-Unimodal, perform comparably, and approach the performance of npTOT as the number of topics reaches that found by npTOT. This is not surprising; a parametric model with the "right" number of topics should perform as well as a nonparametric model. The advantage of a nonparametric model such as npTOT is that we do not need to specify the number of topics a priori, or perform expensive model comparisons, to obtain good results.

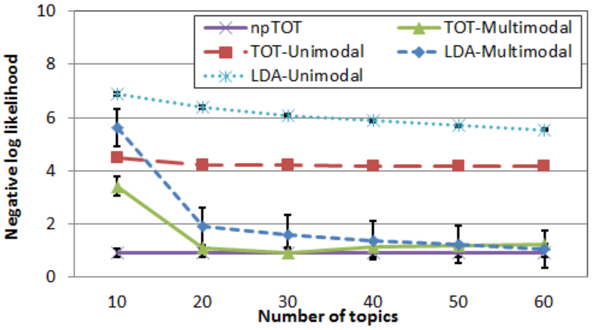


Figure 8: Average per-token negative log likelihood on test set for NIPS dataset.

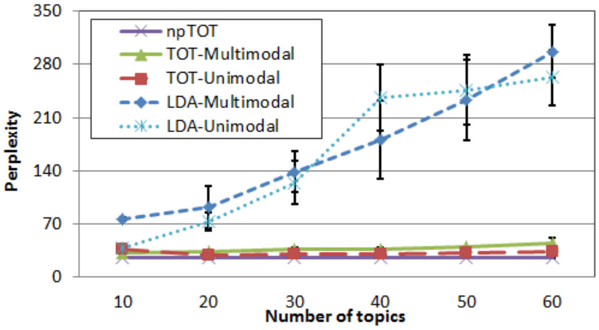


Figure 9: Perplexity on test set for Twitter dataset.

Figures 9, 10 and 11 show the perplexity obtained through document completion. Again, we find that npTOT obtains lower perplexity, indicating that it is better able to predict held-out text. In particular, note that the LDA models learned without an explicit model of time perform very poorly. As expected, having information about when a document is written, and having a model sophisticated enough to make use of this information, allows us to make better guesses about the content of that document.

6 Conclusions and Future Work

The goal of this paper was to develop a flexible model for capturing time-varying topics in text corpora where the total number of topics is not known a priori. By extending the TOT model to incorporate nonparametric distributions over both words and timestamps, we have presented a model that is able to find interpretable topics and achieve good predictive performance on held-out data.

One advantage of the npTOT model described herein is that it can easily be extended to higher

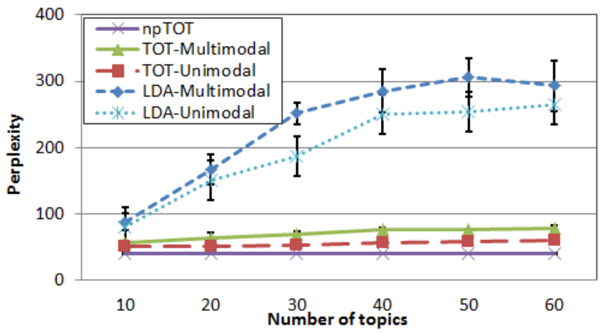


Figure 10: Perplexity on test set for State-of-the-union-address dataset.

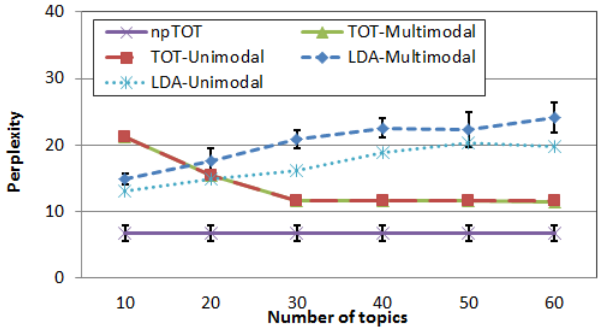


Figure 11: Perplexity on test set for NIPS dataset.

dimensional covariate values. This would enable us to model geographical variations in topic popularity. In addition to modeling documents, topic models have been used to model images (Fei-Fei and Perona, 2005). This is another area where spatially dependent topic models, based on npTOT, could be employed.

While nonparametric models offer greater flexibility than their parametric counterparts, inference tends to be slow. One direction for future research might be to develop faster inference algorithms based on variational methods Wang et al. (2011) or parallelization Williamson et al. (2013) and Asuncion et al. (2008).

References

Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *UAI*.

Asuncion, A., Smyth, P., and Welling, M. (2008). Asynchronous distributed learning of topic models. In *NIPS*.

Blei, D. and McAuliffe, J. (2007). Supervised topic models. In *NIPS*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR*, 3:993–1022.

Caron, F., Davy, M., and Doucet, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures. In *UAI*.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *CVPR*.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. (2007). Euclidean Embedding of Co-occurrence Data. *JMLR*, 8:2265–2295.

Lin, D., Grimson, E., and Fisher, J. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *NIPS*.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proc. Sec. Bayes. Statist. Sci.*

MacEachern, S. (2000). Dependent Dirichlet processes. Technical report, Ohio State University.

Rao, V. and Teh, Y. (2009). Spatial normalized gamma processes. In *NIPS*.

Srebro, N. and Roweis, S. (2005). Time-varying topic models using dependent Dirichlet processes.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *JASA*, 101:1566–1581.

Walker, D. D., Seppi, K., and Ringger, E. K. (2012). Topics over nonparametric time: A supervised topic model using bayesian nonparametric density estimation. In *UAI Applications Workshop*.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *ICML*.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *AISTATS*.

Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *KDD*.

Williamson, S. A., Dubey, A., and Xing, E. P. (2013). Parallel markov chain monte carlo for nonparametric mixture models. In *ICML*.