# Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends

Xuerui Wang, Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
xuerui@cs.umass.edu, mccallum@cs.umass.edu

## ABSTRACT

This paper presents an LDA-style topic model that captures not only the low-dimensional structure of data, but also how the structure changes over time. Unlike other recent work that relies on Markov assumptions or discretization of time, here each topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. Thus, the meaning of a particular topic can be relied upon as constant, but the topics' occurrence and correlations change significantly over time. We present results on nine months of personal email, 17 years of NIPS research papers and over 200 years of presidential state-of-the-union addresses, showing improved topics, better timestamp prediction, and interpretable trends.

## Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*data mining*

## General Terms

Algorithms, Experimentation

## Keywords

Graphical Models, Temporal Analysis, Topic Modeling

## 1. INTRODUCTION

Research in statistical models of co-occurrence has led to the development of a variety of useful *topic models*—mechanisms for discovering low-dimensional, multi-faceted summaries of documents or other discrete data. These include models of words alone, such as Latent Dirichlet Allocation (LDA) [2, 5], of words and research paper citations [4], of word sequences with Markov dependencies [6, 17], of words and their authors [12], of words in a social network of senders and recipients [10], and of words and relations (such as voting patterns) [18]. In each case, graphical model structures are carefully-designed to capture the relevant structure and co-occurrence dependencies in the data.

Many of the large data sets to which these topic models are applied do not have *static* co-occurrence patterns; they are instead *dynamic*. The data are often collected over time, and generally patterns present in the early part of the collection are not in effect later. Topics rise and fall in prominence; they split apart; they merge to form new topics; words change their correlations. For example, across 17 years of the Neural Information Processing Systems (NIPS) conference, activity in "analog circuit design" has fallen off somewhat, while research in "support vector machines" has recently risen dramatically. The topic "dynamic systems" used to co-occur with "neural networks," but now co-occurs with "graphical models."

However none of the above mentioned topic models are aware of these dependencies on document timestamps. Not modeling time can confound co-occurrence patterns and result in unclear, sub-optimal topic discovery. For example, in topic analysis of U.S. Presidential State-of-the-Union addresses, LDA confounds Mexican-American War (1846-1848) with some aspects of World War I (1914-1918), because LDA is unaware of the 70-year separation between the two events. Some previous work has performed some post-hoc analysis—discovering topics without the use of timestamps and then projecting their occurrence counts into discretized time [5]—but this misses the opportunity for time to improve topic discovery.

This paper presents *Topics over Time (TOT)*, a topic model that explicitly models time jointly with word co-occurrence patterns. Significantly, and unlike some recent work with similar goals, our model does not discretize time, and does not make Markov assumptions over state transitions in time. Rather, TOT parameterizes a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics that simultaneously capture word co-occurrences *and* locality of those patterns in time.

When a strong word co-occurrence pattern appears for a brief moment in time then disappears, TOT will create a topic with a narrow time distribution. (Given enough evidence, arbitrarily small spans can be represented, unlike schemes based on discretizing time.) When a pattern of word co-occurrence remains consistent across a long time span, TOT will create a topic with a broad time distribution.


*KDD'06* August 20–23, 2006, Philadelphia, Pennsylvania, USA.

In current experiments, we use a Beta distribution over a (normalized) time span covering all the data, and thus we can also flexibly represent various skewed shapes of rising and falling topic prominence.

The model's generative storyline can be understood in two different ways. We fit the model parameters according to a generative model in which a per-document multinomial distribution over topics is sampled from a Dirichlet, then for each word occurrence we sample a topic; next a per-topic multinomial generates the word, *and* a per-topic Beta distribution generates the document's time stamp. Here the time stamp (which in practice is always observed and constant across the document) is associated with each word in the document. We can also imagine an alternative, corresponding generative model in which the time stamp is generated once per document, conditioned directly on the per-document mixture over topics. In both cases, the likelihood contribution from the words and the contribution from the timestamps may need to be weighted by some factor, as in the balancing of acoustic models and language models in speech recognition. The later generative storyline more directly corresponds to common data sets (with one timestamp per document); the former is easier to fit, and can also allow some flexibility in which different parts of the document may be discussing different time periods.

Some previous studies have also shown that topic discovery can be influenced by information in addition to word co-occurrences. For example, the Group-Topic model [18] showed that the joint modeling of word co-occurrence and voting relations resulted in more salient, relevant topics. The Mixed-Membership model [4] also showed interesting results for research papers and their citations.

Note that, in contrast to other work that models trajectories of individual topics over time, TOT topics and their meaning are modeled as constant over time. TOT captures changes in the occurrence (and co-occurrence conditioned on time) of the topics themselves, not changes in the word distribution of each topic. The classical view of splitting and merging of topics is thus reflected as dynamic changes in the co-occurrence of *constant* topics. While choosing to model individual topics as mutable could be useful, it can also be dangerous. Imagine a subset of documents containing strong co-occurrence patterns across time: first between birds and aerodynamics, then aerodynamics and heat, then heat and quantum mechanics—this could lead to a single topic that follows this trajectory, and lead the user to inappropriately conclude that birds and quantum mechanics are time-shifted versions of the same topic.

Alternatively, consider a large subject like medicine, which has changed drastically over time. In TOT we choose to model these shifts as changes in topic *co-occurrence*—a decrease in occurrence of topics about blood-letting and bile, and an increase in topics about MRI and retrovirus, while the topics about blood, limbs, and patients continue to co-occur throughout. We do not claim that this point of view is better, but the difference makes TOT much simpler to understand and implement.

In comparison to more complex alternatives, the relative simplicity of TOT is a great advantage—not only for the relative ease of understanding and implementing it, but also because this approach can in the future be naturally injected into other more richly structured topic models, such as the Author-Recipient-Topic model to capture changes in social network roles over time [10], and the Group-Topic model to capture changes in group formation over time [18].

We present experimental results with three real-world data sets. On more than two centuries of U.S. Presidential State-of-the-Union addresses, we show that TOT discovers topics with both time-localization and word-clarity improvements over LDA. On the 17-year history of the NIPS conference, we show clearly interpretable topical trends, as well as a two-fold increase in the ability to predict time given a document. On nine months of the second author's email archive, we show another example of clearly interpretable, time-localized topics, such as springtime faculty recruiting. On all three data sets, TOT provides more distinct topics, as measured by KL divergence.

## 2. TOPICS OVER TIME

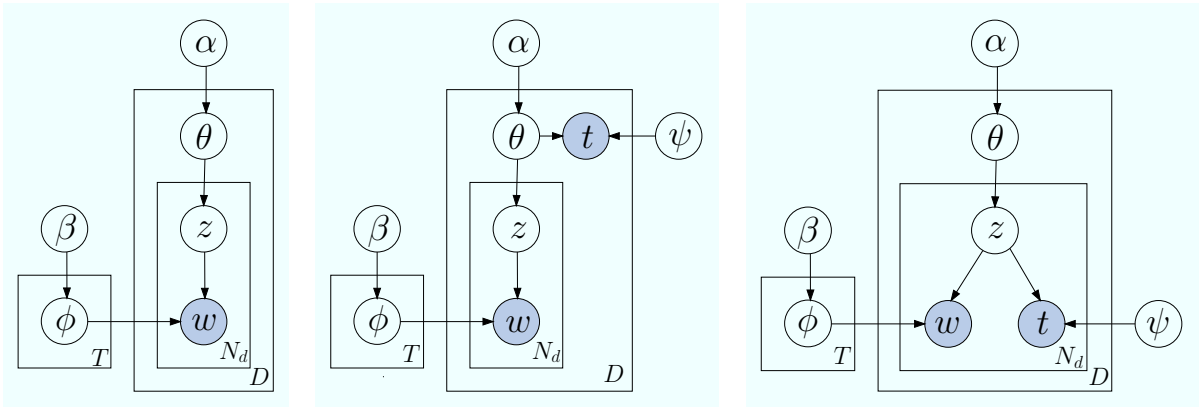
Before introducing the Topics over Time (TOT) model, let us review the basic Latent Dirichlet Allocation model. Our notation is summarized in Table 1, and the graphical model representations of both LDA and our TOT models are shown in Figure 1.

Latent Dirichlet Allocation (LDA) is a Bayesian network that generates a document using a mixture of topics [2]. In its generative process, for each document $d$, a multinomial distribution $\theta_d$ over topics is randomly sampled from a Dirichlet with parameter $\alpha$, and then to generate each word, a topic $z_{di}$ is chosen from this topic distribution, and a word, $w_{di}$, is generated by randomly sampling from a topic-specific multinomial distribution $\phi_{z_{di}}$. The robustness of the model is greatly enhanced by integrating out uncertainty about the per-document topic distribution $\theta$ and the per-topic word distribution $\phi$.

In TOT, topic discovery is influenced not only by word co-occurrences, but also temporal information. Rather than modeling a sequence of state changes with a Markov assumption on the dynamics, TOT models (normalized) absolute timestamp values. This allows TOT to see long-range dependencies in time, to predict absolute time values given an unstamped document, and to predict topic distributions given a timestamp. It also helps avoid a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance.

Time is intrinsically continuous. Discretization of time always begs the question of selecting the slice size, and the size is invariably too small for some regions and too large for others. TOT avoids discretization by associating with each topic a continuous distribution over time. Many possible parameterized distributions are possible. Our earlier experiments were based on Gaussian. All the results in this paper employ the Beta distribution (which can behave versatile shapes), for which the time range of the data used for parameter estimation is normalized to a range from 0 to 1. Another possible choice of bounded distributions is the Kumaraswamy distribution [8]. Double-bounded distributions are appropriate because the training data are bounded in time. If it is necessary to predict in a small window into the future, the bounded region can be extended, yet still estimated based on the data available up to now.

Topics over Time is a generative model of timestamps and the words in the timestamped documents. There are two ways of describing its generative process. The first, which corresponds to the process used in Gibbs sampling for parameter estimation, is as follows:




(a) LDA

(b) TOT model, alternate view

(c) TOT model, for Gibbs sampling

**Figure 1: Three topic models: LDA and two perspectives on TOT**

| SYMBOL | DESCRIPTION |
|---|---|
| $T$ | number of topics |
| $D$ | number of documents |
| $V$ | number of unique words |
| $N_d$ | number of word tokens in document $d$ |
| $\theta_d$ | the multinomial distribution of topics specific to the document $d$ |
| $\phi_z$ | the multinomial distribution of words specific to topic $z$ |
| $\psi_z$ | the beta distribution of time specific to topic $z$ |
| $z_{di}$ | the topic associated with the $i$th token in the document $d$ |
| $w_{di}$ | the $i$th token in document $d$ |
| $t_{di}$ | the timestamp associated with the $i$th token in the document $d$ (in Figure 1(c)) |

**Table 1: Notation used in this paper**

1. Draw $T$ multinomials $\phi_z$ from a Dirichlet prior $\beta$, one for each topic $z$;
2. For each document $d$, draw a multinomial $\theta_d$ from a Dirichlet prior $\alpha$; then for each word $w_{di}$ in document $d$:
   (a) Draw a topic $z_{di}$ from multinomial $\theta_d$;
   (b) Draw a word $w_{di}$ from multinomial $\phi_{z_{di}}$;
   (c) Draw a timestamp $t_{di}$ from Beta $\psi_{z_{di}}$.

The graphical model is shown in Figure 1(c). Although, in the above generative process, a timestamp is generated for each word token, all the timestamps of the words in a document are observed as the same as the timestamp of the document. One might also be interested in capturing burstiness, and some solution such as Dirichlet compound multinomial model (DCM) can be easily integrated into the TOT model [9]. In our experiments there are a fixed number of topics, $T$; although a non-parametric Bayes version of TOT that automatically integrates over the number of topics would certainly be possible.

As shown in the above process, the posterior distribution of topics depends on the information from two modalities—both text and time. TOT parameterization is

$$\begin{aligned} \theta_d|\alpha &\sim \text{Dirichlet}(\alpha) \\ \phi_z|\beta &\sim \text{Dirichlet}(\beta) \\ z_{di}|\theta_d &\sim \text{Multinomial}(\theta_d) \\ w_{di}|\phi_{z_{di}} &\sim \text{Multinomial}(\phi_{z_{di}}) \\ t_{di}|\psi_{z_{di}} &\sim \text{Beta}(\psi_{z_{di}}). \end{aligned}$$

Inference can not be done exactly in this model. We employ Gibbs sampling to perform approximate inference. Note that we adopt conjugate prior (Dirichlet) for the multinomial distributions, and thus we can easily integrate out $\theta$ and $\phi$, analytically capturing the uncertainty associated with them. In this way we facilitate the sampling—that is, we need not sample $\theta$ and $\phi$ at all. Because we use the continuous Beta distribution rather than discretizing time, sparsity is not a big concern in fitting the temporal part of the model. For simplicity and speed we estimate these Beta distributions $\psi_z$ by the method of moments, once per iteration of Gibbs sampling. One could estimate the values of the hyper-parameters of the TOT model, $\alpha$ and $\beta$, from data using a Gibbs EM algorithm [1]. For many applications, topic models are very sensitive to hyper-parameters, and it is extremely important to get the right values for the hyper-parameters. In the particular applications discussed in this paper, we find that the sensitivity to hyper-parameters is not very strong. Thus, again for simplicity, we use fixed symmetric Dirichlet distributions ($\alpha = 50/T$ and $\beta = 0.1$) in all our experiments.

In the Gibbs sampling procedure above, we need to calculate the conditional distribution $P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi)$, where $\mathbf{z}_{-di}$ represents the topic assignments for all tokens except $w_{di}$. We begin with the joint probability of a data set, and using the chain rule, we can obtain the conditional probability conveniently as

$$\begin{aligned} &P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi) \propto (m_{dz_{di}} + \alpha_{z_{di}} - 1) \\ &\times \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v} + \beta_v) - 1} \frac{(1-t_{di})^{\psi_{z_{di}1}-1} t_{di}^{\psi_{z_{di}2}-1}}{B(\psi_{z_{di}1}, \psi_{z_{di}2})}, \end{aligned}$$

where $n_{zv}$ is the number of tokens of word $v$ are assigned to topic $z$, $m_{dz}$ represent the number of tokens in document $d$ are assigned to topic $z$. Detailed derivation of Gibbs sam-

pling for TOT is provided in Appendix A. An overview of the Gibbs sampling procedure we use is shown in Algorithm 1.

**Algorithm 1** Inference on TOT

```
initialize topic assignment randomly for all tokens
for iter = 1 to N_iter do
    for d = 1 to D do
        for w = 1 to N_d do
            draw z_dw from P(z_dw | w, t, z_-dw, α, β, Ψ)
            update n_{z_dw w} and m_{d z_dw}
        end for
    end for
    for z = 1 to T do
        update ψ_z
    end for
end for
compute the posterior estimates of θ and φ
```

Although a document is modeled as a mixture of topics, there is typically only one timestamp associated with a document. The above generative process describes data in which there is a timestamp associated with each word. When fitting our model from typical data, each training document's timestamp is copied to all the words in the document. However, after fitting, if actually run as a generative model, this process would generate different time stamps for the words within the same document. In this sense, thus, it is formally a deficient generative model, but still remains powerful in modeling large dynamic text collections.

An alternative generative process description of TOT, (better suited to generate an unseen document), is one in which a single timestamp is associated with each document, generated by rejection or importance sampling, from a mixture of per-topic Beta distributions over time with mixtures weight as the per-document $\theta_d$ over topics. As before, this distribution over time is ultimately parameterized by the set of timestamp-generating Beta distributions, one per topic. The graphical model for this alternative generative process is shown in Figure 1(b).

Using this model we can predict a time stamp given the words in the document. To facilitate the comparison with LDA, we can discretize the timestamps (only for this purpose). Given a document, we predict its timestamp by choosing the discretized timestamp that maximizes the posterior which is calculated by multiplying the timestamp probability of all word tokens from their corresponding topic-wise Beta distributions over time, that is, $\arg\max_t \prod_{i=1}^{N_d} p(t|\psi_{z_i})$.

It is also interesting to consider obtaining a distribution over topics, conditioned on a timestamp. This allows us to see the topic occurrence patterns over time. By Bayes rule, $\mathrm{E}(\theta_{z_i}|t) = P(z_i|t) \propto p(t|z_i)P(z_i)$ where $P(z_i)$ can be estimated from data or simply assumed as uniform. Examples of expected topic distributions $\theta_d$ conditioned on timestamps are shown in Section 5.

Regarding parameter estimation, the two processes in Figure 1 (b) and (c) can become equivalent when we introduce a balancing hyper-parameter between the likelihood from two modalities. In the second process, not surprisingly, the generation of one timestamp would be overwhelmed by the plurality of words generated under the bag of words assumption. To balance the influence from two different modalities, a tunable hyper-parameter is needed which is responsible for the relative weight of the time modality versus the text modality. Thus we use such a weighting parameter to rescale the likelihoods from different modalities, as is also common in speech recognition when the acoustic and language models are combined, and in the Group-Topic model [18] in which relational Blockstructures and topic models are integrated. Here a natural setting for the weighting parameter is the inverse of the number of words $N_d$ in the document, which is equivalent to generating $N_d$ independent and identically distributed (i.i.d.) samples from the document-specific mixture of Beta distributions. Thus, it is probabilistically equivalent to drawing $N_d$ samples from the individual Beta distributions according to the mixture weights $\theta_d$, which exactly corresponds to the generative process in Figure 1 (c). In practice, it is also important to have such a hyper-parameter when the likelihoods from discrete and continuous modalities are combined. We find that this hyper-parameter is quite sensitive, and set it by trial and error.

## 3. RELATED WORK

Several previous studies have examined topics and their changes across time. Rather than jointly modeling word co-occurrence and time, many of these methods use post-hoc or pre-discretized analysis.

The first style of non-joint modeling involves fitting a time-unaware topic model, and then ordering the documents in time, slicing them into discrete subsets, and examining the topic distributions in each time-slice. One example is Griffiths and Steyvers' study of PNAS proceedings [5], in which they identified hot and cold topics based on examination of topic mixtures estimated from an LDA model.

The second style of non-joint modeling pre-divides the data into discrete time slices, and fits a separate topic model in each slice. Examples of this type include the experiments with the Group-Topic model [18], in which several decades worth of U.N. voting records (and their accompanying text) were divided into 15-year segments; each segment was fit with the GT model, and trends were compared. One difficulty with this approach is that aligning the topics from each time slice can be difficult, although starting Gibbs sampling using parameters from the previous time slice can help, as shown in [14]. Similarly, the TimeMines system [15] for TDT tasks (single topic in each document) constructs *overview timelines* of a set of news stories. A $\chi^2$ test is performed to identify days on which the number of occurrences of named entities or noun phrases produces a statistic above a given threshold; consecutive days under this criterion are stitched together to form an interval to be added into the timeline.

Time series analysis has a long history in statistics, much of which is based on dynamic models, with a Markov assumption that the state at time $t+1$ or $t+\Delta t$ is independent of all other history given the state at time $t$. Hidden Markov models and Kalman filters are two such examples. For instance, recent work in social network analysis [13] proposes a dynamic model that accounts for friendships drifting over time. Blei and Lafferty recently present dynamic topic models (DTMs) in which the alignment among topics across time steps is captured by a Kalman filter [3].

Continuous Time Bayesian Networks (CTBN) [11] are an example of using continuous time without discretization. A CTBN consists of two components: a Bayesian network and a continuous transition model, which avoids various granu-

larity problems due to discretization. Unlike TOT, however, CTBNs use a Markov assumption.

Another Markov model that aims to find word patterns in time is Kleinberg's "burst of activity model" [7]. This approach uses a probabilistic infinite-state automaton with a particular state structure in which high activity states are reachable only by passing through lower activity states. Rather than leveraging time stamps, it operates on a stream of data, using data ordering as a proxy for time. Its infinite-state automaton has a continuous transition scheme similar to CTBNs. However, it operates only on one word at a time, whereas TOT finds time-localized patterns in word *co-occurrences*.

TOT uses time quite differently than the above models. First, TOT does not employ a Markov assumption over time, but instead treats time as an observed continuous variable. Second, many other models take the view that the "meaning" (or word associations) of a topic changes over time; instead, in TOT we can rely on topics themselves as *constant*, while topic co-occurrence patterns change over time.

Although not modeling time, several other topic models have associated the generation of additional modalities with topics. For example, the aforementioned GT model conditions on topics for both word generation and relational links. As in TOT, GT results also show that jointly modeling an additional modality improves the relevance of the discovered topics. Another flexible, related model is the Mixed Membership model [4], which treats the citations of papers as *additional* "words", thus the formed topics are influenced by both words and citations.

# 4. DATA SETS

We present experiments with the TOT model on three real-world data sets: 9 months of email sent and received by the second author, 17 years of NIPS conference papers, and 21 decades of U.S. Presidential State-of-the-Union Addresses. In all cases, for simplicity, we fix the number of topics $T = 50$[1].

## 4.1 State-of-the-Union Addresses

The State of the Union is an annual message presented by the President to Congress, describing the state of the country and his plan for the future. Our data set[2] consists of the transcripts of 208 addresses during 1790-2002 (from George Washington to George W. Bush). We remove stopwords and numbers, and all text is downcased. Because the topics discussed in each address are so diverse, and in order to improve the robustness of the discovered topics, we increase the number of documents in this data set by splitting each transcript into 3-paragraph "documents". The resulting data set has 6,427 (3-paragraph) documents, 21,576 unique words, and 674,794 word tokens in total. Each document's time stamp is determined by the date on which the address was given.

## 4.2 A Researcher's Email

This data set consists of the second author's email archive of the nine months from January to September 2004, including all emails sent and received. In order to model only the new text entered by the author of each message, it is necessary to remove "quoted original messages" in replies. We eliminate this extraneous text by a simple heuristic: all text in a message below a "forwarded message" line or timestamp is removed. This heuristic does incorrectly delete text that are interspersed with quoted email text. Words are formed from sequences of alphabetic characters; stopwords are removed, and all text is downcased. The data set contains 13,300 email messages, 22,379 unique words, and 453,743 word tokens in total. Each document's timestamp is determined by the day and time the message was sent or received.

## 4.3 NIPS Papers

The NIPS data set (provided to us by Gal Chechik) consists of the full text of the 17 years of proceedings from 1987 to 2003 Neural Information Processing Systems (NIPS) Conferences. In addition to downcasing and removing stopwords and numbers, we also remove the words appearing less than five times in the corpus—many of them produced by OCR errors. Two letter words (primarily coming from equations), are removed, except for "ML", "AI", "KL", "BP", "EM" and "IR." The data set contains 2,326 research papers, 24,353 unique words, and 3,303,020 word tokens in total. Each document's timestamp is determined by the year of the proceedings.

# 5. EXPERIMENTAL RESULTS

In this section, we present the topics discovered by the TOT model and compare them with topics from LDA. We also demonstrate the ability of the TOT model to predict the timestamps of documents, more than doubling accuracy in comparison with LDA. We furthermore find topics discovered by TOT to be more distinct from each other than LDA topics (as measured by KL Divergence). Finally we show how TOT can be used to analyze topic co-occurrence conditioned on a timestamp. Topics presented in this section are extracted from a single sample at the 1000th iteration of the Gibbs sampler. For the address data set, 1000 iterations of the Gibbs sampler took 3 hours on a dual-processor Opteron (Linux), 2 hours for the email data set, and 10 hours for the NIPS data set.

## 5.1 Topics Discovered for Addresses

The State-of-the-Union addresses contain the full range of United States history. Analysis of this data set shows strong temporal patterns. Some of them are broad historical issues, such as a clear "American Indian" topic throughout the 1800s and peaking around 1860, or the rise of "Civil Rights" across the second half of the 1900s. Other sharply localized trends are somewhat influenced by the individual president's communication style, such as Theodore Roosevelt's sharply increased use of the words "great", "men", "public", "country", and "work". Unfortunately, space limitations prevent us from showing all 50 topics.

Four TOT topics, their most likely words, their Beta distributions over time, their actual histograms over time, as well as comparisons against their most similar LDA topic (by KL divergence), are shown in Figure 2. Immediately we see that the TOT topics are more neatly and narrowly focused in time; (time analysis for LDA is done post-hoc). An immediate and obvious effect is that this helps the reader understand more precisely when and over what length of time the topical trend was occurring. For example, in the left-

[1]It would be straightforward to automatically infer the number of topics using algorithms such as Hierarchical Dirichlet Process [16].

[2]http://www.gutenberg.org/dirs/etext04/suall11.txt

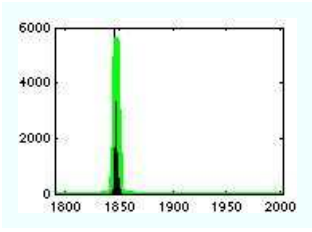
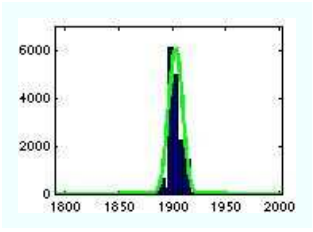
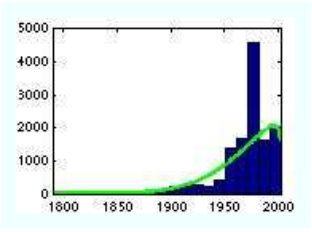
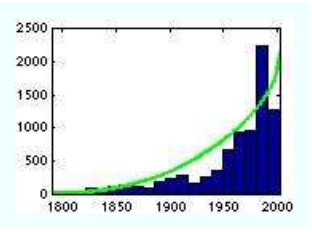
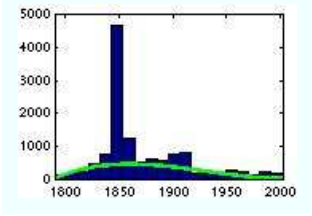
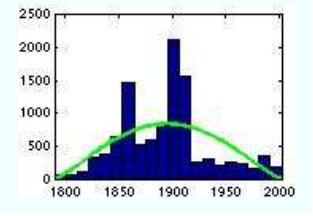
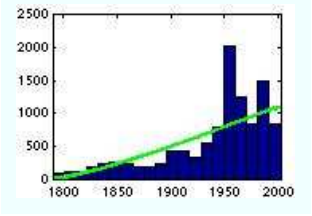
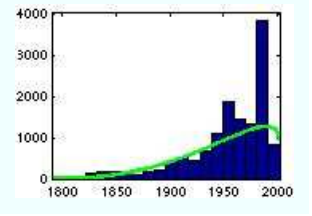
| Mexican War | | Panama Canal | | Cold War | | Modern Tech | |
|---|---|---|---|---|---|---|---|
|  | |  | |  | |  | |
| states | 0.02032 | government | 0.02928 | world | 0.01875 | energy | 0.03902 |
| mexico | 0.01832 | united | 0.02132 | states | 0.01717 | national | 0.01534 |
| government | 0.01670 | states | 0.02067 | security | 0.01710 | development | 0.01448 |
| united | 0.01521 | islands | 0.01167 | soviet | 0.01664 | space | 0.01436 |
| war | 0.01059 | canal | 0.01014 | united | 0.01491 | science | 0.01227 |
| congress | 0.00951 | american | 0.00872 | nuclear | 0.01454 | technology | 0.01227 |
| country | 0.00906 | cuba | 0.00834 | peace | 0.01408 | oil | 0.01178 |
| texas | 0.00852 | made | 0.00747 | nations | 0.01069 | make | 0.00994 |
| made | 0.00727 | general | 0.00731 | international | 0.01024 | effort | 0.00969 |
| great | 0.00611 | war | 0.00660 | america | 0.00987 | administration | 0.00957 |
|  | |  | |  | |  | |
| mexico | 0.06697 | government | 0.05618 | defense | 0.05556 | program | 0.02674 |
| government | 0.02254 | american | 0.02696 | military | 0.03819 | energy | 0.02477 |
| mexican | 0.02141 | central | 0.02518 | forces | 0.03308 | development | 0.02287 |
| texas | 0.02109 | canal | 0.02283 | security | 0.03020 | administration | 0.02119 |
| territory | 0.01739 | republic | 0.02198 | strength | 0.02406 | economic | 0.01710 |
| part | 0.01610 | america | 0.02170 | nuclear | 0.01858 | areas | 0.01585 |
| republic | 0.01344 | pacific | 0.01832 | weapons | 0.01654 | programs | 0.01578 |
| military | 0.01111 | panama | 0.01776 | arms | 0.01254 | major | 0.01534 |
| state | 0.00974 | nicaragua | 0.01381 | maintain | 0.01161 | nation | 0.01242 |
| make | 0.00942 | isthmus | 0.01137 | strong | 0.01106 | assistance | 0.01052 |

**Figure 2: Four topics discovered by TOT (above) and LDA (bottom) for the Address data set. The titles are our own interpretation of the topics. Histograms show how the topics are distributed over time; the fitted beta PDFs is shown also. (For LDA, beta distributions are fit in a post-hoc fashion). The top words with their probability in each topic are shown below the histograms. The TOT topics are better localized in time, and TOT discovers more event-specific topical words.**

most topic, TOT clearly shows that the Mexican-American war (1846-1848) occurred in the few years just before 1850. In LDA, on the other hand, the topic spreads throughout American history; it has its peak around 1850, but seems to be getting confused by a secondary peak around the time of World War I, (when "war" words were used again, and relations to Mexico played a small part). It is not so clear what event is being captured by LDA's topic.

The second topic, "Panama Canal," is another vivid example of how TOT can successfully localize a topic in time, and also how jointly modeling words and time can help sharpen and improve the topical word distribution. The Panama Canal (constructed during 1904-1914) is correctly localized in time, and the topic accurately describes some of the issues motivating canal construction: the sinking of the *U.S.S. Maine* in a Cuban harbor, and the long time it took U.S. warships to return to the Caribbean via Cape Horn. The LDA counterpart is not only widely spread through time, but also confounding topics such as modern trade relations with Central America and efforts to build the Panama Railroad in the 1850s.

The third topic shows the rise and fall of the Cold War, with a peak on the Reagan years, when Presidential rhetoric on the subject rose dramatically. Both TOT and LDA topics mention "nuclear," but only TOT correctly identifies "soviet". LDA confounds what is mostly a cold war topic (although it misses "soviet") with words and events from across American history, including small but noticeable bumps for World War I and the Civil War. TOT correctly has its own separate topic for World War I.

Lastly, the rightmost topics in Figure 2, "Modern Tech," shows a case in which the TOT topic is not necessarily

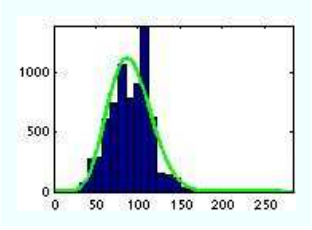
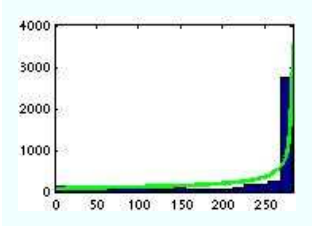
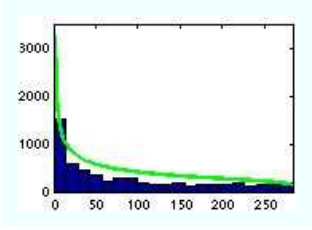
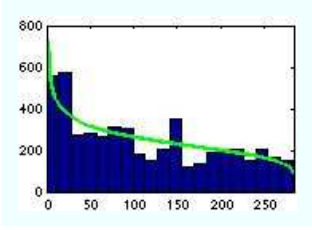
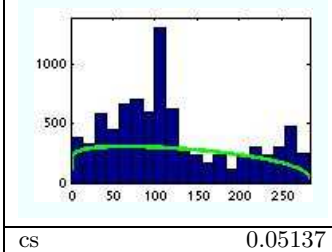
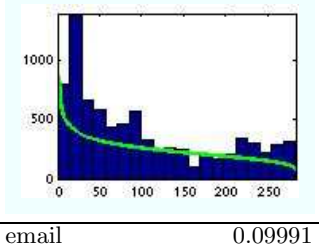
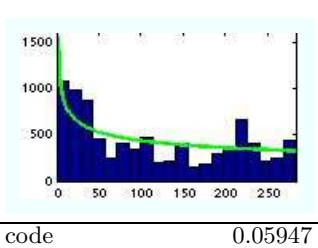
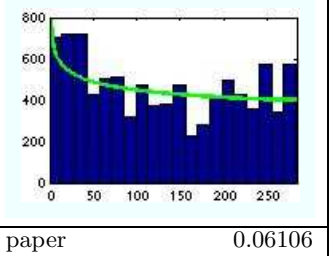
| Faculty Recruiting | | ART Paper | | MALLET | | CVS Operations | |
|---|---|---|---|---|---|---|---|
|  | |  | |  | |  | |
| cs | 0.03572 | xuerui | 0.02113 | code | 0.05668 | check | 0.04473 |
| april | 0.02724 | data | 0.01814 | files | 0.04212 | page | 0.04070 |
| faculty | 0.02341 | word | 0.01601 | mallet | 0.04073 | version | 0.03828 |
| david | 0.02012 | research | 0.01408 | java | 0.03085 | cvs | 0.03587 |
| lunch | 0.01766 | topic | 0.01366 | file | 0.02947 | add | 0.03083 |
| schedule | 0.01656 | model | 0.01238 | al | 0.02479 | update | 0.02539 |
| candidate | 0.01560 | andres | 0.01238 | directory | 0.02080 | latest | 0.02519 |
| talk | 0.01355 | sample | 0.01152 | version | 0.01664 | updated | 0.02317 |
| bruce | 0.01273 | enron | 0.01067 | pdf | 0.01421 | checked | 0.02277 |
| visit | 0.01232 | dataset | 0.00960 | bug | 0.01352 | change | 0.02156 |
|  | |  | |  | |  | |
| cs | 0.05137 | email | 0.09991 | code | 0.05947 | paper | 0.06106 |
| david | 0.04592 | ron | 0.04536 | mallet | 0.03922 | page | 0.05504 |
| bruce | 0.02734 | messages | 0.04095 | version | 0.03772 | web | 0.04257 |
| lunch | 0.02710 | data | 0.03408 | file | 0.03702 | title | 0.03526 |
| manmatha | 0.02391 | calo | 0.03236 | files | 0.02534 | author | 0.02763 |
| andrew | 0.02332 | message | 0.03053 | java | 0.02522 | papers | 0.02741 |
| faculty | 0.01764 | enron | 0.03028 | cvs | 0.02511 | email | 0.02204 |
| april | 0.01740 | project | 0.02415 | directory | 0.01978 | pages | 0.02193 |
| shlomo | 0.01657 | send | 0.02023 | add | 0.01932 | nips | 0.01967 |
| al | 0.01621 | part | 0.01680 | checked | 0.01481 | link | 0.01860 |

**Figure 3: Four topics discovered by TOT (above) and LDA (bottom) for the Email data set, showing improved results with TOT. For example, the Faculty Recruiting topic is correctly identified in the spring in the TOT model, but LDA confuses it with other interactions among faculty.**

better—just interestingly different than the LDA topic. The TOT topic, with mentions of *energy, space, science, and technology*, is about modern technology and energy. Its emphasis on modern times is also very distinct in its time distribution. The closest LDA topic also includes energy, but focuses on economic development and assistance to other nations. Its time distribution shows an extra bump around the decade of the Marshal Plan (1947-1951), and a lower level during George W. Bush's presidency—both inconsistent with the time distribution learned by the TOT topic.

## 5.2 Topics Discovered for Email

In Figure 3 we demonstrate TOT on the Email data set. Email is typically full of seasonal phenomena (such as paper deadlines, summer semester, etc.). One such seasonal example is the "Faculty Recruiting" topic, which (unlike LDA) TOT clearly identifies and localizes in the spring. The LDA counterpart is widely spread over the whole time period, and consequently, it cannot separate faculty recruiting from other types of faculty interactions and collaboration. The temporal information captured by TOT plays a very important role in forming meaningful time-sensitive topics.

The topic "ART paper" reflects a surge of effort in collaboratively writing a paper on the Author-Recipient-Topic model. Although the co-occurrence pattern of the words in this topic is strong and distinct, LDA failed to discover a corresponding topic—likely because it was a relatively short-lived phenomena. The closest LDA topic shows the general research activities, work on the DARPA CALO project, and various collaborations with SRI to prepare the Enron email data set for public release. Not only does modeling time help TOT discover the "ART paper" task, but an alternative model that relied on coarse time discretization may miss such topics that have small time spans.

The "MALLET" topic shows that, after putting in an intense effort in writing and discussing Java programming for the MALLET toolkit, the second author had less and less time to write code for the toolkit. In the correspond-

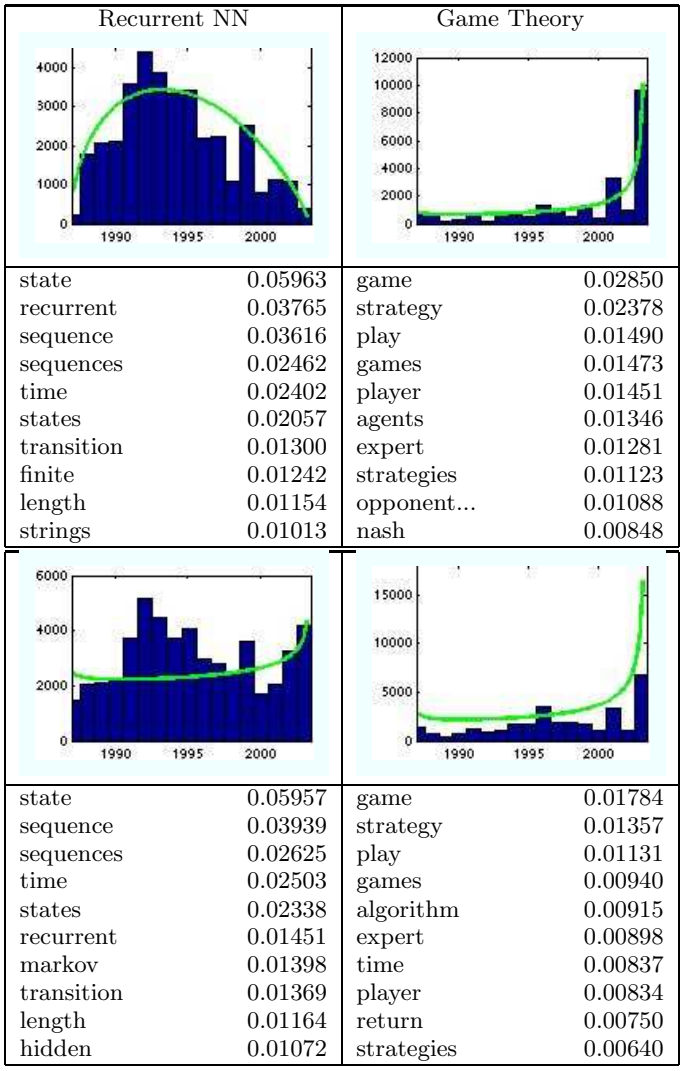



**Figure 4: Two topics discovered by TOT (above) and LDA (bottom) for the NIPS data set. For example, on the left, two major approaches to dynamic system modeling are confounded by LDA, but TOT more clearly identifies waning interest in Recurrent Neural Networks, with a separate topic (not shown) for rising interest in Markov models.**

ing LDA topic, MALLET development is confounded with CVS operations—which were later also used for managing collaborative writing of research papers.

TOT appropriately and clearly discovers a separate topics for "CVS operations," seen in the rightmost column. The closest LDA topic is the previously discussed one that merges MALLET and CVS. The second closest LDA topic (bottom right) discusses research paper writing, but not CVS. All these examples show that TOT's use of time can help it pull apart distinct events, tasks and topics that may be confusingly merged by LDA.

## 5.3 Topics Discovered for NIPS

Research paper proceedings also present interesting trends for analysis. Successfully modeling trends in the research literature can help us understand how research fields evolve, and measure the impact of differently shaped profiles in time.

Figure 4 shows two topics discovered from the NIPS proceedings. "Recurrent Neural Networks" is clearly identified by TOT, and correctly shown to rise and fall in prominence within NIPS during the 1990s. LDA, unaware of the fact that Markov models superceded Recurrent Neural Networks for dynamic systems in the later NIPS years, and unaware of the time-profiles of both, ends up mixing the two methods together. LDA has a second topic elsewhere that also covers Markov models.

On the right, we see "Games" and game theory. This is an example in which TOT and LDA yield nearly identical results, although, if the terms beyond simply the first ten are examined, one sees that LDA is emphasizing board games, such as chess and backgammon, while TOT used its ramping-up time distribution to more clearly identify game theory as part of this topic (e.g., the word "Nash" occurs in position 12 for TOT, but not in the top 50 for LDA).

We have been discussing the salience and specificity of TOT's topics. Distances between topics can also be measured numerically. Table 2 shows the average distance of word distributions between all pairs of topics, as measured by KL Divergence. In all three data sets, the TOT topics are more distinct from each other. Partially because the Beta distribution is rarely multi-modal, the TOT model strives to separate events that occur during different time spans, and in real-world data, time differences are often correlated with word distribution differences that would have been more difficult to tease apart otherwise. The MALLET-CVS-paper distinction in the email data set is one example. (Events with truly multi-modal time distributions would be modeled with alternatives to the Beta distribution.)

**Table 2: Average KL divergence between topics for TOT vs. LDA on three data sets. TOT finds more distinct topics.**

| | Address | Email | NIPS |
|---|---|---|---|
| TOT | 0.6266 | 0.6416 | 0.5728 |
| LDA | 0.5965 | 0.5943 | 0.5421 |

**Table 3: Predicting the decade, in the Address data set. L1 Error is the difference between predicted and true decade. In the Accuracy column, we see that TOT predicts exactly the correct decade nearly twice as often as LDA.**

| | L1 Error | E(L1) | Accuracy |
|---|---|---|---|
| TOT | 1.98 | 2.02 | 0.19 |
| LDA | 2.51 | 2.58 | 0.10 |

## 5.4 Time Prediction

One interesting feature of our approach (not shared by state-transition-based Markov models of topical shifts) is the capability of predicting the timestamp given the words in a document. This task also provides another opportunity to quantitatively compare TOT against LDA.

On the State-of-the-Union Address data set, we measure the ability to predict the decade given the text of the address, as measured in accuracy, L1 error and average L1

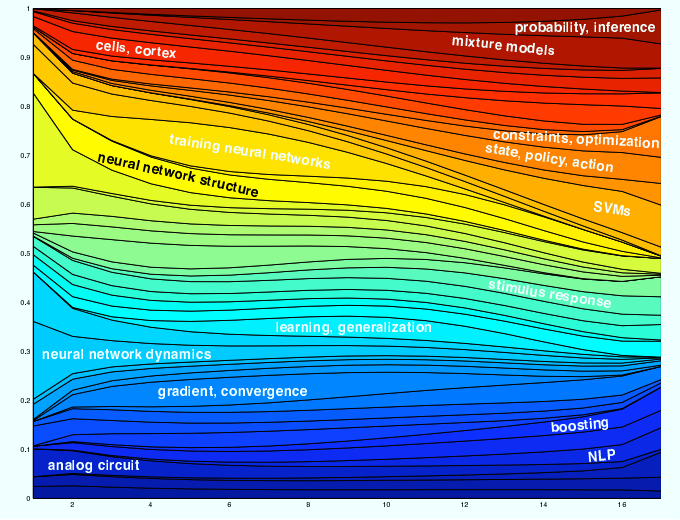



**Figure 5: The distribution over topics given time in the NIPS data set. Note the rich collection of shapes that emerge from the Bayesian inversion of the collection of per-topic Beta distributions over time.**

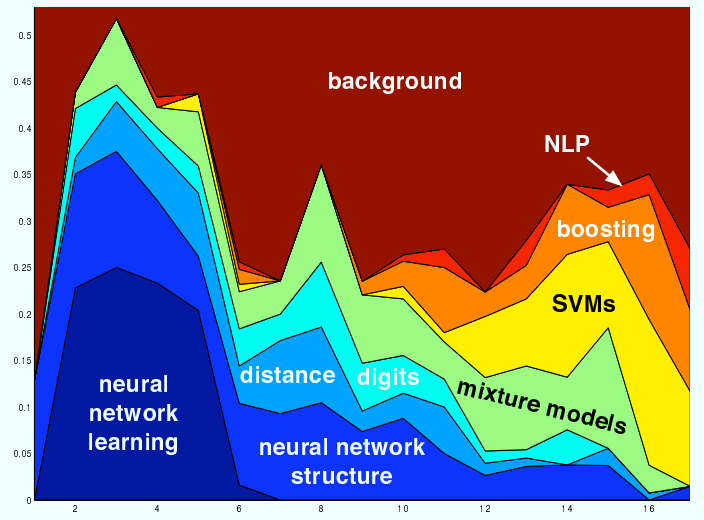



**Figure 6: Eight topics co-occurring strongly with the "classification" topic in the NIPS data set. Other co-occurring topics are labeled as a combined background topic. Classification with neural networks declined, while co-occurrence with SVMs, boosting and NLP are on the rise.**

distance to the correct decade (number of decades difference between predicted and correct decade). As shown in Table 3, TOT achieves double the accuracy of LDA, and provides an L1 relative error reduction of 20%.

## 5.5 Topic Distribution Profile over Time

It is also interesting to consider the TOT model's distribution over topics as a function of time. The time distribution of each *individual* topic is described as a Beta distribution (having flexible mean, variance and skewness), but even more rich and complex profiles emerge from the *interactions* among these Beta distributions. TOT's approach to modeling topic distributions conditioned on time stamp—based on multiple time-generating Betas, inverted with Bayes rule—has the dual advantages of a relatively simple, easy-to-fit parameterization, while also offering topic distributions with a flexibility that would be more difficult to achieve with a direct, non-inverted parameterization, (*i.e.* one generating topic distributions directly conditioned on time, without Bayes-rule inversion).

The expected topic mixture distributions for the NIPS data set are shown in Figure 5. The topics are consistently ordered in each year, and the heights of a topic's region represents the relative weight of the corresponding topic given a timestamp, calculated using the procedure described in Section 2. We can clearly see that topic mixtures change dramatically over time, and have interesting shapes. NIPS begins with more emphasis on neural networks, analog circuits and cells, but now emphasizes more SVMs, optimization, probability and inference.

## 5.6 Topic Co-occurrences over Time

We can also examine topic *co*-occurrences over time, which, as discussed in Section 1, are dynamic for many large text collections. In the following, we say two topics $z_1$ and $z_2$ (strongly) co-occur in a document $d$ if both $\theta_{z_1}$ and $\theta_{z_2}$ are greater than some threshold $h$ (we set $h = 2/T$); then we can count the number of documents in which certain topics (strongly) co-occur, and map out how co-occurrence patterns change over time.

Figure 6 shows the prominence profile over time of those topics that co-occur strongly with the NIPS topic "classification." We can see that at the beginning NIPS, this problem was solved primarily with neural networks. It co-occurred with the "digit recognition" in the middle 90's. Later, probabilistic mixture models, boosting and SVM methods became popular.

# 6. CONCLUSIONS

This paper has presented Topic over Time (TOT), a model that jointly models both word co-occurrences and localization in continuous time. Results on three real-world data sets show the discovery of more salient topics that are associated with events, and clearly localized in time. We also show improved ability to predict time given a document. Reversing the inference by Bayes rule, yields a flexible parameterization over topics conditioned on time, as determined by the interactions among the many per-topic Beta distributions.

Unlike some related work with similar motivations, TOT does not require discretization of time or Markov assumptions on state dynamics. The relative simplicity of our approach provides advantages for injecting these ideas into other topic models. For example, in ongoing work we are finding patterns in topics and group membership over time, with a Group-Topic model over time. Many other extensions are possible.

# 7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense

Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor. We also thank Charles Sutton and David Mimno for helpful discussions.

## 8. REFERENCES

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[4] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 2004.

[5] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235, 2004.

[6] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems (NIPS) 17*, 2004.

[7] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[8] P. Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46:79–88, 1980.

[9] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[10] A. McCallum, A. Corrada-Emanuel, and X. Wang. Topic and role discovery in social networks. In *Proceedings of 19th International Joint Conference on Artificial Intelligence*, 2005.

[11] U. Nodelman, C. Shelton, and D. Koller. Continuous time Bayesian networks. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 378–387, 2002.

[12] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[13] P. Sarkar and A. Moore. Dynamic social network analysis using latent space models. In *The 19th Annual Conference on Neural Information Processing Systems*, 2005.

[14] X. Song, C.-Y. Lin, B. L. Tseng, and M.-T. Sun. Modeling and predicting personal information dissemination behavior. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[15] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 73–80, 2000.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Technical report, UC Berkeley Statistics TR-653, 2004.

[17] X. Wang and A. McCallum. A note on topical n-grams. Technical report, UMass UM-CS-2005-071, 2005.

[18] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and text. In *The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Link Discovery: Issues, Approaches and Applications*, pages 28–35, 2005.

# APPENDIX

## A. GIBBS SAMPLING DERIVATION FOR TOT

We begin with the joint distribution $P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi)$. We can take advantage of conjugate priors to simplify the integrals. All symbols are defined in Section 2.

$$
\begin{aligned}
& P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi) \\
= \; & P(\mathbf{w}|\mathbf{z}, \beta) p(\mathbf{t}|\Psi, \mathbf{z}) P(\mathbf{z}|\alpha) \\
= \; & \int P(\mathbf{w}|\Phi, \mathbf{z}) p(\Phi|\beta) \mathrm{d}\Phi p(\mathbf{t}|\Psi, \mathbf{z}) \int P(\mathbf{z}|\Theta) p(\Theta|\alpha) \mathrm{d}\Theta \\
= \; & \int \prod_{d=1}^{D} \prod_{i=1}^{N_d} P(w_{di}|\phi_{z_{di}}) \prod_{z=1}^{T} p(\phi_z|\beta) \mathrm{d}\Phi \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
& \times \int \prod_{d=1}^{D} \left( \prod_{i=1}^{N_d} P(z_{di}|\theta_d) p(\theta_d|\alpha) \right) \mathrm{d}\Theta \\
= \; & \int \prod_{z=1}^{T} \prod_{v=1}^{V} \phi_{zv}^{n_{zv}} \prod_{z=1}^{T} \left( \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{zv}^{\beta_v - 1} \right) \mathrm{d}\Phi \\
& \times \int \prod_{d=1}^{D} \prod_{z=1}^{T} \theta_{dz}^{m_{dz}} \prod_{d=1}^{D} \left( \frac{\Gamma(\sum_{z=1}^{T} \alpha_z)}{\prod_{z=1}^{T} \Gamma(\alpha_z)} \prod_{z=1}^{T} \theta_{dz}^{\alpha_z - 1} \right) \mathrm{d}\Theta \\
& \times \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
= \; & \left( \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^T \left( \frac{\Gamma(\sum_{z=1}^{T} \alpha_z)}{\prod_{z=1}^{T} \Gamma(\alpha_z)} \right)^D \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(t_{di}|\psi_{z_{di}}) \\
& \times \prod_{z=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(n_{zv} + \beta_v)}{\Gamma(\sum_{v=1}^{V} (n_{zv} + \beta_v))} \prod_{d=1}^{D} \frac{\prod_{z=1}^{T} \Gamma(m_{dz} + \alpha_z)}{\Gamma(\sum_{z=1}^{T} (m_{dz} + \alpha_z))}
\end{aligned}
$$

Using the chain rule, we can obtain the conditional probability conveniently,

$$
\begin{aligned}
& P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi) \\
= \; & \frac{P(z_{di}, w_{di}, t_{di}|\mathbf{w}_{-di}, \mathbf{t}_{-di}, \mathbf{z}_{-di}, \alpha, \beta, \Psi)}{P(w_{di}, t_{di}|\mathbf{w}_{-di}, \mathbf{t}_{-di}, \mathbf{z}_{-di}, \alpha, \beta, \Psi)} \\
\propto \; & \frac{P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi)}{P(\mathbf{w}_{-di}, \mathbf{t}_{-di}, \mathbf{z}_{-di}|\alpha, \beta, \Psi)} \\
\propto \; & \frac{n_{z_{di} w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V} (n_{z_{di} v} + \beta_v) - 1} (m_{d z_{di}} + \alpha_{z_{di}} - 1) p(t_{di}|\psi_{z_{di}}) \\
\propto \; & (m_{d z_{di}} + \alpha_{z_{di}} - 1) \frac{n_{z_{di} w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V} (n_{z_{di} v} + \beta_v) - 1} \\
& \times \frac{(1 - t_{di})^{\psi_{z_{di}1} - 1} t_{di}^{\psi_{z_{di}2} - 1}}{B(\psi_{z_{di}1}, \psi_{z_{di}2})}
\end{aligned}
$$

In practice, the balancing hyper-parameter often appears as an exponential power of the last term above. Since timestamps are drawn from continuous Beta distributions, sparsity is not a big problem for parameter estimation of $\Psi$. For simplicity, we update $\Psi$ after each Gibbs sample by the method of moments, detailed as follows:

$$
\begin{aligned}
\hat{\psi}_{z1} &= \bar{t}_z \left( \frac{\bar{t}_z (1 - \bar{t}_z)}{s_z^2} - 1 \right) \\
\hat{\psi}_{z2} &= (1 - \bar{t}_z) \left( \frac{\bar{t}_z (1 - \bar{t}_z)}{s_z^2} - 1 \right)
\end{aligned}
$$

where $\bar{t}_z$ and $s_z^2$ indicate the sample mean and the biased sample variance of the timestamps belonging to topic $z$, respectively.