A Dual Markov Chain Topic Model for Dynamic Environments

Ayan Acharya
CognitiveScale Inc
Austin, Texas
aacharya@utexas.edu

Joydeep Ghosh
University of Texas at Austin
Austin, Texas
ghosh@ece.utexas.edu

Mingyuan Zhou
University of Texas at Austin
Austin, Texas
mzhou@utexas.edu

ABSTRACT

The abundance of digital text has led to extensive research on topic models that reason about documents using latent representations. Since for many online or streaming textual sources such as news outlets, the number, and nature of topics change over time, there have been several efforts that attempt to address such situations using dynamic versions of topic models. Unfortunately, existing approaches encounter more complex inferencing when their model parameters are varied over time, resulting in high computation complexity and performance degradation. This paper introduces the **DM-DTM**, a **d**ual **M**arkov chain **d**ynamic **t**opic **m**odel, for characterizing a corpus that evolves over time. This model uses a gamma Markov chain and a Dirichlet Markov chain to allow the topic popularities and word-topic assignments, respectively, to vary smoothly over time. Novel applications of the Negative-Binomial augmentation trick result in simple, efficient, closed-form updates of all the required conditional posteriors, resulting in far lower computational requirements as well as less sensitivity to initial conditions, as compared to existing approaches. Moreover, via a gamma process prior, the number of desired topics is inferred directly from the data rather than being pre-specified and can vary as the data changes. Empirical comparisons using multiple real-world corpora demonstrate a clear superiority of DM-DTM over strong baselines for both static and dynamic topic models.

CCS CONCEPTS

• **Mathematics of computing** → **Bayesian networks**; **Bayesian nonparametric models**; **Time series analysis**; • **Information systems** → **Document topic models**;

KEYWORDS

dynamic topic model; CRT augmentation; Gibbs sampling

ACM Reference Format:
Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2018. A Dual Markov Chain Topic Model for Dynamic Environments. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219995

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219995

1 INTRODUCTION

Analysis of dyadic data, which represent the relationships between two different sets of entities, such as documents and words *or* users and items, has been a prolific domain of research over the past decade, driven largely by applications in diverse areas such as topic modeling [5, 9, 17], recommender systems [13, 16, 29], e-commerce [38] and bio-informatics [36]. Successful as these analysis techniques are, a major limitation of most of them is that they are static models and ignore the temporal correlation and evolution of the relationships between entities – an attribute present in most real-world dyadic data. Text mining researchers have developed a handful of techniques for analyzing corpora that evolve over time by modeling them as a sequence of document-by-word count matrices. Some of these techniques employ Kalman filtering based inference and a nonlinear transformation of the latent states to the discrete observations [8, 31, 45], while others [5, 6] use a temporal Dirichlet process and make arguably simplistic assumptions to calculate an intractable posterior. Since the inference techniques for linear dynamical systems are well-developed, one usually is tempted to connect a count-valued observation to a latent Gaussian random variable. However, such approaches often incur heavy computation cost, fail to exploit the natural sparsity of the data and lack interpretation of the latent states as the components of these states may take negative values. This is also true for models in recommendation systems that exploit temporal correlation [28, 48] but hypothesize that the observation is generated from an interaction of latent factors that assume a normal distribution. Clearly, such an assumption is restrictive for count-valued dyadic data unless some nonlinear transformation is used, which again makes the inference intractable [12]. This critical problem of non-conjugacy arising from latent Gaussian variables and their subsequent non-linear transformation to model count-valued observations can be further mitigated using the Pólya-Gamma augmentation trick [19, 31]. However, such augmentation does not necessarily improve the empirical performance, as evidenced in Section 4.

The objective of this paper is to model a set of documents that evolve over time and provide an inference mechanism without making crude approximations. To that end, we introduce **DM-DTM**; a novel **d**ual **M**arkov chain based **d**ynamic **t**opic **m**odel. A critical aspect of DM-DTM is that, unlike the standard techniques adopted in both text mining and recommender system problems, the observations are modeled using a Poisson distribution and the latent factors/topics are allowed to vary smoothly over time using the gamma and Dirichlet distributions. To be more specific, two separate Markov chains are introduced – a gamma Markov chain and a Dirichlet Markov chain. The gamma Markov chain models the temporal evolution of the popularities of the topics. The Dirichlet Markov chain, on the other hand, is employed to adapt the topic-word assignment with time. Gibbs sampling is adopted for

inference where the conditional posteriors are all available in closed form. This is made possible by the use of an augmentation trick associated with the negative binomial distribution together with a forward-backward sampling algorithm, each step of which assumes a posterior that is easy to sample from [1]. Using the gamma process [20] to generate a countably infinite number of weighted latent factors in the prior, the model can infer a parsimonious set of topics from the data in the posterior. Empirical comparisons in terms of held-out perplexity indicate the clear superiority of DM-DTM over two of the most widely-used temporal topic models [8, 31].

The remainder of the paper is organized as follows. Section 2 provides a detailed description of the modeling assumptions and the inference techniques of DM-DTM. Related works are outlined in Section 3. Empirical results with real-world data are reported in Section 4. Finally, the conclusion and future works are listed in Section 5.

2 MODEL AND INFERENCE

In what we present below, vectors and matrices are denoted by bold-faced lowercase and capital letters respectively. Scalar variables are written in italic font, and sets are denoted by calligraphic uppercase letters. Dir(), Gam(), Pois() and mult() stand for the Dirichlet, gamma, Poisson and multinomial distributions, respectively. For a tensor $\boldsymbol{X} \in \mathbb{Z}^{K_1 \times K_2 \times K_3}$ the $(k_1, k_2, k_3)^{\text{th}}$ entry is denoted by $x_{k_1 k_2 k_3}$. Also, $x_{k_1 k_2 \cdot} = \sum_{k_3=1}^{K_3} x_{k_1 k_2 k_3}$ and $x_{k_1 \cdot \cdot} = \sum_{k_2=1}^{K_2} \sum_{k_3=1}^{K_3} x_{k_1 k_2 k_3}$.

Consider a collection of matrices $\{\boldsymbol{X}_t \in \mathbb{Z}^{|\mathcal{D}_t| \times V}\}_{t=1}^{T}$ that are sequentially observed. These matrices are the bag-of-words representation of a corpus that evolves over time. In what is proposed hereafter, each document in the corpus appears in only one time-slice. In particular, let t_d denote the time-stamp of the d^{th} document and $\mathcal{D}_t$ denote the set of documents that appear in the t^{th} time-stamp.

Further, consider a gamma process [20] $G \sim \Gamma\text{P}(c, G_0)$ defined on the product space $\mathbb{R}_+ \times \Omega$, with scale parameter c and a finite and continuous base measure G_0 over a complete separable metric space Ω, such that $G(A_i) \sim \text{Gam}(G_0(A_i), 1/c)$ are independent gamma random variables for disjoint partition $\{A_i\}_i$ of Ω. The Lévy measure of the gamma process can be expressed as:

$$\nu(drd\omega) = r^{-1} e^{-cr} dr G_0(d\omega).$$

A gamma process based model has an inherent shrinkage mechanism, as in the prior the number of atoms with weights greater than $\tau \in \mathbb{R}_+$ follows a Poisson distribution whose parameter is given by $H(\Omega) \int_{\tau}^{\infty} r^{-1} \exp(-cr) dr$. The value of this parameter decreases as τ increases. A draw from the gamma process is expressed as $G = \sum_{k=1}^{\infty} r_{0k} \delta_{\boldsymbol{\beta}_{1k}}$, where $\boldsymbol{\beta}_{1k} \in \Omega$ is a V-dimensional atom drawn from $\boldsymbol{\beta}_{1k} \sim \text{Dir}(\eta)$ and $r_{0k} = G(\boldsymbol{\beta}_{1k})$ is the associated weight.

We associate each atom $\boldsymbol{\beta}_{1k}$ with an r_{1k} and generate a *gamma Markov chain* by letting:

$$r_{tk} | r_{(t-1)k} \sim \text{Gam}(r_{(t-1)k}, 1/c), \; t \in \{1, \ldots, T\}.$$

The parameter r_{tk} models the global popularity of the latent factor k at time t. Similarly, we generate a *Dirichlet Markov chain* by letting:

$$\boldsymbol{\beta}_{tk} | \boldsymbol{\beta}_{(t-1)k} \sim \text{Dir}(\eta V \boldsymbol{\beta}_{(t-1)k}), \; t \in \{2, \ldots, T\}.$$

The parameter $\boldsymbol{\beta}_{tk}$ models the distribution of the words within the k^{th} latent factor at time t. Additionally, each atom $\boldsymbol{\beta}_{tk}$ is associated with an atom $(\theta_{dk})_{d \in \mathcal{D}_t}$ which is a D_t-dimensional distribution characterized as $(\theta_{dk})_{d \in \mathcal{D}_t} \sim \prod_{d=1}^{D_t} \text{Gam}(r_{tk}, 1/c_d)$. The $(d, w)^{\text{th}}$ entry of $\boldsymbol{X}_t$ is assumed to be generated from a sum of latent counts as: $x_{tdw} \sim \text{Pois}\left(\sum_k \lambda_{tdwk}\right)$ where $\lambda_{tdwk} = \theta_{dk} \beta_{twk}$. One may consider λ_{tdwk} as the strength of the k^{th} latent factor that dictates the relation between the d^{th} document and the w^{th} word at time t. Each of these latent counts is composed of two parts – θ_{dk} models the affinity of the d^{th} document to the k^{th} latent factor and β_{twk} models the popularity of the w^{th} word among the k^{th} latent factor at time t. Each latent factor contributes such a count and the total count is the aggregate of the countably infinite latent factors.

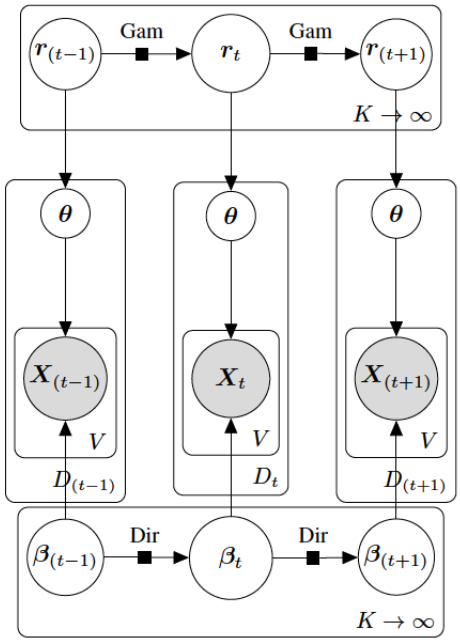

Figure 1: Graphical Model of DM-DTM

To complete the generative process, we put Gamma priors over c, c_d, γ_0 and η as: $c \sim \text{Gam}(a_0, 1/b_0)$, $c_d \sim \text{Gam}(c_0, 1/d_0)$, $\gamma_0 \sim \text{Gam}(e_0, 1/f_0)$ and $\eta \sim \text{Gam}(s_0, 1/t_0)$. In the formulation above, we assume that the global popularity of the latent factors evolves over time using a gamma Markov chain. At the t^{th} time instance, the proximity of the d^{th} document to the k^{th} latent factor is given by θ_{dk}, which in turn is generated from a Gamma distribution with scale r_{tk}. Therefore, the evolution of r_{tk} may capture the changes in the semantic themes (or topics) over time that these documents talk about. Additionally, the words that describe the topics can also evolve smoothly over time using the Dirichlet Markov chain imposed on the $\boldsymbol{\beta}_{tk}$'s. Moreover, using the gamma process prior, the model adjusts its capacity automatically as the number of active topics vary with time. The corresponding plate diagram of DM-DTM is shown in Fig. 1.

Though DM-DTM supports a countably infinite number of latent factors, in practice, it is impossible to instantiate all of them. Therefore, a finite approximation of the infinite model is considered by truncating the number of factors to K which approaches the original infinite model as $K \to \infty$. The sampling proceeds as follows.

Sampling of x_{tdwk} : The sampling of the latent rates x_{tdwk} follows from the relation between Poisson and multinomial distribution and can be derived as:

$$((x_{tdwk})_{k=1}^{K}|-) \sim \text{Mult}\left(x_{tdw}; \frac{(\theta_{dk}\beta_{twk})_{k=1}^{K}}{\sum_{k=1}^{K}\theta_{dk}\beta_{twk}}\right).$$

Sampling of r_{tk} : The difficulty in inferring the shape parameter of the gamma distribution and the unique construction of the gamma Markov chain make the sampling of the r_{tk}'s non-trivial. To that end, we introduce the negative binomial (NB) distribution. The NB distribution $m \sim \text{NB}(r,p)$, with probability mass function (PMF) $P(M=m) = \frac{\Gamma(m+r)}{m!\Gamma(r)}p^m(1-p)^r$ for $m \in \mathbb{Z}$, can be augmented into a gamma-Poisson construction as $m \sim \text{Pois}(\lambda)$, $\lambda \sim \text{Gam}(r, p/(1-p))$, where the gamma distribution is parameterized by its shape r and scale $p/(1-p)$. It can also be augmented under a compound Poisson representation as $m = \sum_{t=1}^{l} u_t, u_t \overset{iid}{\sim} \text{Log}(p), l \sim \text{Pois}(-r\ln(1-p))$, where $u \sim \text{Log}(p)$ is the logarithmic distribution [26]. Consequently, we have the following Lemma.

Lemma 2.1 ([54]). *If $m \sim \text{NB}(r,p)$ is represented under its compound Poisson representation, then the conditional posterior of l given m and r has PMF:*

$$P(l=j \mid m,r) = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,j)|r^j,\ j=0,1,\cdots,m,$$

where $|s(m,j)|$ are unsigned Stirling numbers of the first kind. We denote this conditional posterior as $(l|m,r) \sim \text{CRT}(m,r)$, a Chinese restaurant table (CRT) count random variable, which can be generated via $l = \sum_{n=1}^{m} z_n, z_n \sim \text{Bernoulli}(r/(n-1+r))$.

The following lemma is a consequence of Lemma 2.1.

Lemma 2.2 ([1]). *If $x_i \sim \text{Pois}(m_i r_2)$, $r_2 \sim \text{Gam}(r_1, 1/d)$, $r_1 \sim \text{Gam}(a, 1/b)$, then $(r_1|-) \sim \text{Gam}(a+\ell, 1/(b-log(1-p)))$ where $(\ell|x,r_1) \sim \text{CRT}(\sum_i x_i, r_1)$ and $p = \sum_i m_i/(d+\sum_i m_i)$.*

For $t=T$, we augment $\ell_{Tdk} \sim \text{CRT}(x_{Td.k}, r_{Tk})$ and then sample r_{Tk} according to Lemma 2.2 as:

$$(r_{Tk}|-) \sim \text{Gam}\Big(r_{(T-1)k} + \textstyle\sum_{d\in\mathcal{D}_T}\ell_{Tdk},\\ 1/(c+\textstyle\sum_{d\in\mathcal{D}_T}\log(1+1/c_d))\Big).$$

We now describe how the posterior is calculated for $t=(T-1)$. The same process recursively applies for deriving the posteriors for $1 \le t \le (T-2)$. Note that, using Lemma 2.1, one can write $\ell_{Tdk} \sim \text{Pois}(r_{Tk}\log(1+1/c_d))$. Further, using the additive property of the Poisson distribution, we have $\sum_{d\in\mathcal{D}_T}\ell_{Tdk} \sim \text{Pois}(r_{Tk}\sum_{d\in\mathcal{D}_T}\log(1+1/c_d))$. Therefore, for deriving the posterior for $r_{(T-1)k}$, we can integrate out r_{Tk} and obtain $\sum_{d\in\mathcal{D}_T}\ell_{Tdk} \sim \text{NB}(r_{(T-1)k}, q_{Tk})$, where

$$q_{Tk} = \frac{\sum_{d\in\mathcal{D}_T}\log(1+1/c_d)}{c+\sum_{d\in\mathcal{D}_t}\log(1+1/c_d)}.$$

We then augment $L_{Tk} \sim \text{CRT}(\sum_{d\in\mathcal{D}_T}\ell_{Tdk}, r_{(T-1)k})$, $\ell_{(T-1)dk} \sim \text{CRT}(x_{(T-1)d.k}, r_{(T-1)k})$ and using Lemmas 2.1 and 2.2, one can now sample $r_{(T-1)k}$ as:

$$(r_{(T-1)k}|-) \sim \text{Gam}\Big(r_{(T-2)k} + \textstyle\sum_{d\in\mathcal{D}_{(T-1)}}\ell_{(T-1)dk} + L_{Tk},\\ 1/(c+\textstyle\sum_{d\in\mathcal{D}_{(T-1)}}\log(1+1/c_d) - \log(1-q_{Tk}))\Big).$$

For $1 \le t \le (T-2)$, we augment $\ell_{tdk} \sim \text{CRT}(x_{td.k}, r_{tk})$, $L_{(t+1)k} \sim \text{CRT}(\sum_{d\in\mathcal{D}_{(t+1)}}\ell_{(t+1)dk}, r_{tk})$, apply Lemma 2.1 and 2.2 repeatedly, and then sample:

$$(r_{tk}|-) \sim \text{Gam}(r_{(t-1)k} + \sum_{d\in\mathcal{D}_t}\ell_{tdk} + L_{(t+1)k},\\ 1/(c+\sum_{d\in\mathcal{D}_t}\log(1+1/c_d) - \log(1-q_{(t+1)k}))),$$

where $q_{(t+1)k} = \frac{\sum_{d\in\mathcal{D}_{(t+1)}}\log(1+1/c_d)-\log(1-q_{(t+2)k})}{(c+\sum_{d\in\mathcal{D}_{(t+1)}}\log(1+1/c_d)-\log(1-q_{(t+2)k}))}$. For $t=0$, augment $L_{1k} \sim \text{CRT}(\sum_{d\in\mathcal{D}_1}\ell_{1dk}, r_{0k})$ and according to Lemma 2.2 sample:

$$(r_{0k}|-) \sim \text{Gam}(\gamma_0/K + L_{1k}, 1/(c-\log(1-q_{1k}))),$$
$$q_{1k} = \frac{\sum_{d\in\mathcal{D}_1}\log(1+1/c_d) - \log(1-q_{2k})}{(c+\sum_{d\in\mathcal{D}_1}\log(1+1/c_d) - \log(1-q_{2k}))}.$$

Sampling of θ_{dk} : Sampling of these variables are straightforward and follows from Bayes' rule:

$$(\theta_{dk}|-) \sim \text{Gam}\left(r_{t_dk} + x_{t_dd.k}, 1/\left(c_d+1\right)\right).$$

Sampling of c_d and c : To derive the updates of these parameters, we make use of the conjugacy of a gamma distribution with another gamma distribution for the scale parameter. The sampling for c_d and c then follows as:

$$(c_d|-) \sim \text{Gam}\left(c_0 + r_{t_d.}, 1/\left(d_0+\theta_{d.}\right)\right),$$
$$(c|-) \sim \text{Gam}\left(\gamma_0 + a_0 + \sum_{k=1}^{K}\sum_{t=0}^{(T-1)} r_{tk}, 1/\left(\sum_{k=1}^{K}\sum_{t=0}^{T} r_{tk} + b_0\right)\right).$$

Sampling of β_{twk} : For $t=T$, the conditional posterior is relatively easy to calculate. However, the unique construction of the Dirichlet Markov chain makes the inference for β_{twk}'s very difficult for $1 \le t \le (T-1)$. To make the inference tractable, we first observe that if $x_w \sim \text{Pois}(m\beta_w)\ \forall w \in \{1,2,\cdots,V\}$ and $\boldsymbol{\beta} \sim \text{Dir}(\boldsymbol{\eta})$, then $(\boldsymbol{\beta}|-) \sim \text{Dir}(\eta_1+x_1,\cdots,\eta_V+x_V)$. This follows directly from the relation between the Poisson and multinomial distributions and Bayes' rule. In addition, we introduce the Dirichlet-multinomial distribution: Let $\boldsymbol{x} = (x_w)_{w=1}^{V}$ be a random vector of category counts sampled from a multinomial distribution as $\boldsymbol{x} \sim \text{mult}(\boldsymbol{\beta})$. Additionally, let $\boldsymbol{\beta} \sim \text{Dir}(\boldsymbol{\eta})$. The marginal distribution of $\boldsymbol{x} = (x_w)_{w=1}^{V}$ obtained by integrating out $\boldsymbol{\beta}$ has the pdf of a Dirichlet-multinomial (Dirmult) distribution as given below:

$$f(\boldsymbol{x}|\boldsymbol{\eta}) = (\textstyle\sum_w x_w)!\frac{\Gamma(\sum_w \eta_w)}{\Gamma(\sum_w x_w+\sum_w \eta_w)}\prod_{w=1}^{V}\frac{\Gamma(x_w+\eta_w)}{x_w!\Gamma(\eta_w)}.$$

The introduction of the Dirichlet-multinomial distribution leads to the following Lemma which we utilize for computing the closed-form inference with the Dirichlet Markov chain.

Lemma 2.3. *([53]) If* $\boldsymbol{\beta} \sim Dir(\eta)$, $\eta \sim Gam(s_0, 1/t_0)$, $(x_w)_{w=1}^V \sim mult((\beta_w)_{w=1}^V; \sum_w x_w)$ *then*

$$(\eta|-) \sim Gam(s_0 + \sum_w \xi_w, 1/(t_0 - V log(1-\zeta))),$$

where $\xi_w \sim CRT(x_w, \eta)$ *and* $\zeta \sim Beta(\sum_w x_w, \eta V)$.

The sampling for $t = T$ is easy and follows as:

$$((\beta_{Twk})_{w=1}^V|-) \sim \text{Dir}\left(\left(\eta V \beta_{(T-1)wk} + x_{T.wk}\right)_{w=1}^V\right).$$

For $2 \leq t \leq (T-1)$, the sampling is non-trivial due to the Dirichlet Markov chain. However, from the relation between the Poisson and multimnomial distributions, it follows that

$$(x_{(t+1).wk})_{w=1}^V \sim \text{mult}\left(\left(\beta_{(t+1)wk}\right)_{w=1}^V; x_{(t+1)..k}\right).$$

Since $(\beta_{(t+1)wk})_{w=1}^V \sim \text{Dir}(\eta V(\beta_{twk})_{w=1}^V)$, we may integrate out $\beta_{(t+1)wk}$ and according to the definition of the Dirichlet-multinomial distribution, we have

$$\left(x_{(t+1).wk}\right)_{w=1}^V \sim \text{Dirmult}\left(\eta V\,(\beta_{twk})_{w=1}^V\right).$$

The Dirichlet-multinomial likelihood is further augmented with $\zeta_{(t+1)k} \sim \text{Beta}(x_{(t+1)..k}, \eta V)$ and according to Lemma 2.3, the joint distribution takes the following form:

$$f((x_{(t+1).wk})_{w=1}^V, \zeta_{(t+1)k}) \propto \prod_{w=1}^V \text{NB}(x_{(t+1).wk}; \eta V, \zeta_{(t+1)k}).$$

We now augment $\xi_{(t+1)wk} \sim \text{CRT}(x_{(t+1).wk}, \eta\beta_{twk})$ and using the results of 2.3 sample β_{twk} as:

$$((\beta_{twk})_{w=1}^V|-) \sim \text{Dir}\left(\left(\eta V \beta_{(t-1)wk} + x_{t.wk} + \xi_{(t+1)wk}\right)_{w=1}^V\right).$$

This augmentation trick is illustrated in further details in the proof of Lemma 2.3 [53]. For $t = 1$, the sampling follows almost the same pattern except that the prior is changed:

$$((\beta_{1wk})_{w=1}^V|-) \sim \text{Dir}\left((\eta + x_{1.wk} + \xi_{2wk})_{w=1}^V\right).$$

Sampling of γ_0 : We augment $\ell_{0k} \sim \text{CRT}(\ell_{1k}, \gamma_0/K)$ and use Lemma 2.2 to derive:

$$(\gamma_0|-) \sim \text{Gam}\left(e_0 + \textstyle\sum_k \ell_{0k}, 1/\left(f_0 - 1/K \sum_k \log(1-p_{0k})\right)\right),$$

$$p_{0k} = \frac{\log(1-p_{1k})}{(\log(1-p_{1k}) - c)}.$$

Sampling of η : Sampling of η follows from an application of Lemma 2.2 and the Bayes' rule as:

$$(\eta|-) \sim \text{Gam}\left(s_0 + \textstyle\sum_{t,w,k} \xi_{twk}, 1/\left(t_0 - \sum_{t,k} \log\left(1-\zeta_{tk}\right)\right)\right).$$

The sequence in which the sampling is performed is concisely presented in Algorithm 1. For the temporal correlation in the latent variables, the sampling needs to follow a backward step and a forward step in every epoch, which is designated by $s \in \{1, 2, \cdots, S\}$ in Algorithm 1. The variables are all indexed by an additional superscript (s) just to highlight the specific epoch. Note that the run-time complexity of Algorithm 1 is dictated by the number of non-zero entries in the observed corpus $\{\boldsymbol{X}_t \in \mathbb{Z}^{|\mathcal{D}_t| \times V}\}_{t=1}^T$.

We would like to emphasize further that both the model and the inference are novel contributions of this paper. That the NB augmentation trick can be utilized for an efficient inference procedure in hierarchical graphical models was first proposed in Zhou and Carin [54], however, it was first utilized for modeling time-evolving count vectors in Acharya et al. [1]. Such adoption of the NB trick was non-trivial, as is the case with the current paper that further uses it for modeling two separate Markov chains–the gamma Markov chain and the Dirichlet Markov chain–to yield closed-form solution for Gibbs sampling. These samples converge to a meaningful representation only when a precise order of sampling is followed, as suggested in Algorithm 1. Due to the introduction of the CRT distributed random variables, the backward sampling step must precede the forward sampling step, the precise explanation of which can be found in Acharya et al. [1]. Also, we strongly believe that the simplicity in the final form of the updates leads to superior empirical results. Note that none of the existing works on the temporal topic model has such assumptions that naturally fit the overdispersed count data, facilitates interpretability of the latent states, has closed-form and straight-forward updates in inference and exhibits such superior empirical performance. Moreover, as mentioned in Section 4, the performance of DM-DTM is least sensitive to the initialization of the parameters, a flexibility absent in any existing implementation of temporal topic model.

Algorithm 1: Forward Backward Gibbs Sampling

Result: $\{r_{tk}^{(s)}\}_{s=1}^S$, $\{\theta_{dk}^{(s)}\}_{s=1}^S$, $\{\beta_{twk}^{(s)}\}_{s=1}^S$;
for $s \in \{1, 2, \cdots, S\}$ **do**
  for $d \in \{\mathcal{D}_t\}_{t=1}^T$ **do**
    sample $\{x_{tdwk}^{(s)}\}$ and $c_d^{(s)}$;
  end
  backward sampling: initialize $t = T$;
  while $t > 0$ **do**
    sample $\{\ell_{tdk}^{(s)}\}$, $\{L_{tk}^{(s)}\}$, $\{\zeta_{tk}^{(s)}\}$ and $\{\xi_{twk}^{(s)}\}$;
    cache $\{q_{tk}^{(s)}\}$ to use in forward sampling;
    $t = (t-1)$;
  end
  forward sampling: initialize $t = 1$;
  while $t \leq T$ **do**
    sample $\{r_{tk}^{(s)}\}$, $\{\theta_{d\in\mathcal{D}_t k}^{(s)}\}$ and $\{\beta_{twk}^{(s)}\}$;
    $t = (t+1)$;
  end
  sample $c^{(s)}$, $\gamma_0^{(s)}$, $\eta^{(s)}$;
end

3 RELATED WORK

Poisson Factor Analysis: Since the document-by-word observation matrices in DM-DTM are modeled using Poisson factorization, a brief discussion of Poisson factor analysis is necessary. A large number of discrete latent variable models for count matrix factorization can be united under Poisson factor analysis (PFA) [1–3, 55, 56], which factorizes a count matrix $\mathbf{Y} \in \mathbb{Z}^{D\times V}$ under the Poisson likelihood as $\mathbf{Y} \sim \text{Pois}(\boldsymbol{\Theta}\boldsymbol{\beta})$, where $\boldsymbol{\Theta} \in \mathbb{R}_+^{D\times K}$ is the factor loading matrix or dictionary, $\boldsymbol{\beta} \in \mathbb{R}_+^{K\times V}$ is the factor score matrix. For

example, non-negative matrix factorization [11, 30], with the objective to minimize the Kullback-Leibler divergence between $\mathbf{N}$ and its factorization $\boldsymbol{\Theta\beta}$, is essentially PFA solved with maximum likelihood estimation. LDA [9] is equivalent to PFA, in terms of both block Gibbs sampling and variational inference [55, 56], if Dirichlet distribution priors are imposed on both $\boldsymbol{\theta}_k \in \mathbb{R}_+^D$, the columns of $\boldsymbol{\Theta}$, and $\boldsymbol{\beta}_k \in \mathbb{R}_+^V$, the columns of $\boldsymbol{\beta}$.

Temporal Topic Models: One of the notable contributions towards a dynamic topic model leverages the well-known concept of Gaussian state space evolution. In Blei and Lafferty [8], a Kalman filter is used to infer temporal updates to the state space parameters, which are then mapped to the topic simplex. Wang et al. [45] allow a continuous time state space sampling, but still employ a Gaussian distribution and a mapping to the topic space thereafter using a logistic-normal distribution. These models also require the number of topics to be specified in advance. Elibol et al. [19], Linderman et al. [31] employ the Pólya-Gamma augmentation trick [37] to conquer the non-conjugacy that arises from the Gaussian state space evolution and likelihood for modeling count-valued observations.

Ahmed and Xing [5, 6] use a temporal Dirichlet process and make arguably simplistic assumptions to calculate an intractable posterior. In particular, the framework of temporal Dirichlet process, first introduced in Ahmed and Xing [5], is combined with the Hierarchical Dirichlet Process (HDP) [42] to facilitate smooth temporal evolution and admixture modeling. In such formulation, the base measures of the HDP's for different time slices are modeled using a temporal Dirichlet process and the documents for a given time slice are assumed to be generated following an HDP with the corresponding base measure. The non-conjugacy that arises in such a modeling assumption requires one to use a Metropolis-Hastings sampler for inferring the word-topic assignments. However, to their credit, Ahmed and Xing [6] model both the genesis and death of topics and Wang et al. [46] further model nonlinear evolutionary traces in temporal data, which we avoid in this paper but plan to incorporate in a later submission. Iwata et al. [24], Nallapati et al. [35] emphasize on the problem of modeling topics spread on a timeline with multiple resolutions, namely how topics are organized in a hierarchy and how they evolve over time. Similarly, Srebro and Roweis [41] use the framework of the Dependent Dirichlet Process (DDP) [34] to model more flexible, non-Markovian variation in topic probabilities, but inference in all such models scales very poorly. Bhadury et al. [7] adopt the framework of stochastic gradient Langevin dynamics [32] to accelerate the inference based on Gibbs sampling in the original formulation of dynamic topic model [8]. Some of the other online algorithms [4, 23, 51] explicitly model temporal evolution by making Markovian assumptions.

Different from the works mentioned above, the topics over time (TOT) model [47] assumes that the topics define a distribution over words as well as time slices. Though TOT and some of its extensions [18, 44] can model non-Markovian variations in the topic probabilities and enjoy inference that is computationally tractable, they do not explicitly evolve the parameters of the model with time. Though the modeling assumptions are interesting, there has not been much empirical comparison between these two different sets of algorithms.

Relevant Temporal Models for Count Data: Time-evolving dyadic data is also prevalent in applications of recommender systems and social network analysis. Though such applications are not the focus of the current paper, we discuss a few algorithms for completeness. Both Bayesian Probabilistic Tensor factorization (BPTF) [48] and Dynamic Poisson factorization (DPF) [12] model the temporal evolution using normal distribution. While BPTF models the count data using the normal distribution itself, DPF uses an exponential function to convert the latent rates to nonnegative values, a transformation that makes the inference intractable. To impose temporal smoothness in the frequency domain for audio processing, Virtanen et al. [43] consider chaining latent variables across successive time frames *via* the Gamma scale parameters. Jerfel et al. [25] model the evolution of the latent factors in the context of recommender systems *via* the Gamma scale parameters. Similarly, Févotte et al. [21] propose a gamma Markov chain using the scale parameters for applications in audio and speech. Most of the works in dynamic social network analysis [22, 27, 49] employ similar temporal evolution using a normal distribution to model time-varying *binary* matrices. This paper borrows some of the technical ideas from Acharya et al. [1, 2], Schein et al. [39], which introduce gamma Markov chain for analyzing count and binary data with temporal correlation.

4 EMPIRICAL EVALUATION

4.1 Experiments with Synthetic Data

To illustrate the working principles of DM-DTM, we created a synthetic corpus that has three different time slices. The document-by-word matrices corresponding to each of these time slices are presented in the first column of Fig. 2. Note that each document-*by*-word matrix, denoted by $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, has a clearly defined structure where some documents have the exact same words and some words only appear in a given set of documents. The appearance of the words is varied smoothly from one time slice to the next, replicating the temporal evolution that we may see in a real-world corpus. The reconstructed matrices, denoted by $\hat{\mathbf{X}}_1$, $\hat{\mathbf{X}}_2$, and $\hat{\mathbf{X}}_3$, are presented in the second column which precisely reflect the original observations. The third, fourth and the fifth column display the derived parameters of the model corresponding to different time slices. Note that the r_{tk}'s, which represent the popularities of the topics, only have few components that are dominant and span across time-slices, implying the temporal smoothness discovered by the model, which is a consequence of using both the gamma Markov chain and the Dirichlet Markov chain. Similarly, the θ_{dk}'s (document-topic assignments) and the β_{twk}'s (topic-word assignments) also have a temporal correlation, as is evident from the heat-maps.

4.2 Experiments with Real-world Data

4.2.1 *Description of Datasets.*

- **NIPS Corpus:** The NIPS corpus consists of papers that appeared in the NIPS conference from the years 1987 to 1999. After standard pre-processing and removal of most frequent and least frequent words, the size of the corpus is reduced to 1383 documents and 1636 words. Documents were divided into 13 epochs based on the publication year.
- **Business News Corpora:** We create three additional corpora for our experimental analysis by crawling the Bloomberg News

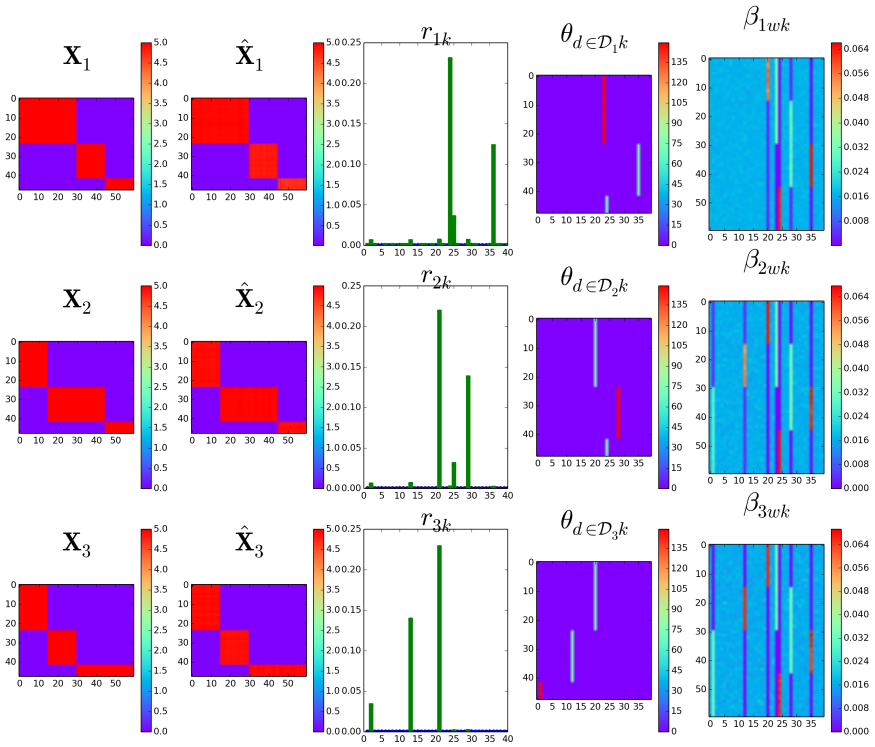

Figure 2: Performance of DM-DTM on Synthetic Data

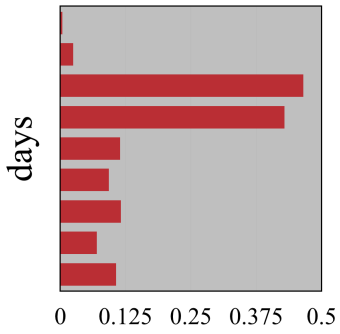

Figure 3: Evolution of r_k

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9
Egypt	Egypt	Paris	Paris	Paris	Paris	Paris	Paris	Paris
flight	kill	kill	France	France	France	France	France	France
Russia	flight	security	security	terror	security	suspect	Islam	terror
crash	Russia	gunman	gunman	security	Islam	terror	suicide	Islam
plane	crash	Islam	Belgium	Islam	kill	Islam	bomb	gunman
kill	plane	terror	Islam	bomb	stadium	soccer	Abaaoud	Syria
sinai	sinai	stadium	Europe	Belgium	soccer	François	Brussel	Europe
bomb	tourism	bomb	terror	Islam	Europe	Syria	terror	raid
airline	gunman	train	bomb	stadium	militant	suicide	militant	Abaaoud
resort	Islam	holland	stadium	train	François	Europe	raid	bomb

Table 1: Evolution of Topic-Word Assignments

portal. In particular, we are only interested in the news article that has a mention of the companies that belong to the list of Financial Times Stock Exchange (FTSE) 250. Each of these corpora consists of news articles from 9 successive days. The three corpora, termed as Business News Corpus 1, 2, and 3, has news articles starting from November 1st 2015, November 12th 2015, and November 22nd 2015 respectively. After standard pre-processing and removal of most frequent and least frequent words, the first news corpus consists of 3271 documents and 1636 words, the second corpus consists of 2935 documents and 1570 words, and the third corpus consists of 2234 documents and 1352 words. The datasets used in these experiments are listed here (https://goo.gl/uVB1f7)[1].

4.2.2 Qualitative Evaluation of Topics. For qualitative understanding of how DM-DTM works with real-world data, we consider the Business News Corpus 2 whose documents span from Nov 12th, 2015 to Nov 20th, 2015. The significance of this corpus is due to the unfortunate event of terrorist attacks in Paris, France late in the evening on Nov 13th, 2015, which triggered massive socio-economic impact worldwide. The news articles published Nov 14th onwards convey information about the incidents and their impacts on the

[1]We are indebted to Matt Sanchez, CTO of CognitiveScale, for curating these datasets.

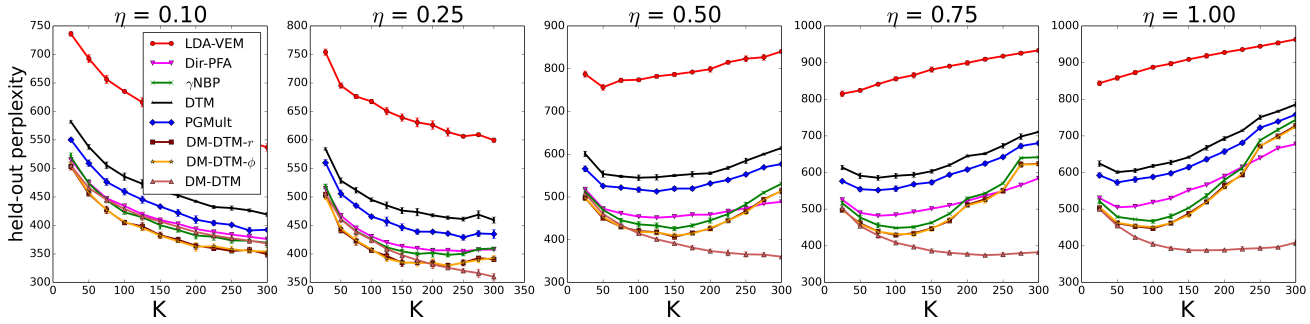

Figure 4: Performance Comparison on NIPS Corpus

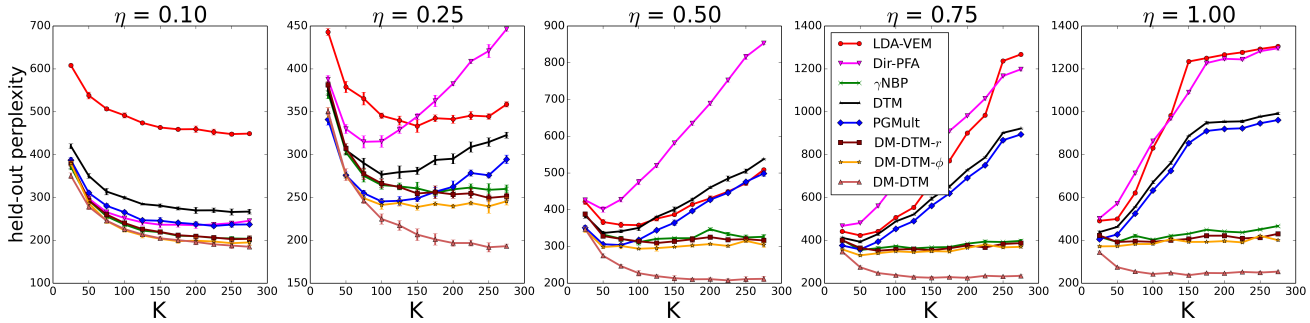

Figure 5: Performance Comparison on Business News Corpus 1

global economy. In Fig. 3, we show the temporal evolution of the strength of one of the 50 topics that are used to model this corpus in one of the experiments. In Fig. 1, the top 10 words corresponding to this topic for all the time slices are also displayed. One can clearly see how the semantics of this topic change over time. Note that, before Nov 14$^{\text{th}}$ (day 3), the composition of the topic is significantly different. However, the terrorist attacks on the evening of the 13$^{\text{th}}$ (day 2), enhance the strength of the topic and change its composition. As time advances, the topic incorporates words like “Syria" and “Abbaaoud" linking the origin and the perpetrators of the terrorist attack. Interestingly, the former French President François Hollande also appears in the topic as he is quoted condemning the genocide in the news articles.

4.2.3 Quantitative Results. We randomly hold out p fraction of the data ($p \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$), train a model with the rest and then predict on the held-out set. For comparing multiple models with different assumptions and inference mechanisms the per-word perplexity on the held-out words is considered, which is defined as:

$$\text{Perplexity} = \exp\left(-\frac{1}{y_{\cdot\cdot}}\sum_{d=1}^{D}\sum_{w=1}^{V} y_{dw}\log f_{dw}\right),$$

where $y_{\cdot\cdot} = \sum_{d,w} y_{dw}$. For models where inference is carried out using Gibbs sampling, f_{dw} is defined as:

$$f_{dw} = \sum_{s,k}\theta_{dk}^{(s)}\phi_{wk}^{(s)} / \sum_{s,w,k}\theta_{dk}^{(s)}\phi_{wk}^{(s)},$$

where $s \in \{1, \cdots, S\}$ are the indices of collected samples. For models that employ variational methods for inference,

$$f_{dw} = \sum_{k}\bar{\theta}_{dk}\bar{\phi}_{wk} / \sum_{w,k}\bar{\theta}_{dk}\bar{\phi}_{wk},$$

where $\bar{\theta}_{dk}$ and $\bar{\phi}_{wk}$ are the point estimates of the respective parameters obtained from the variational inference. Note that the per-word perplexity is equal to V if $f_{dw} = 1.0/V$, thus it should be no greater than V for a topic model that works appropriately. The final results are averaged over five random training/testing partitions.

For concrete empirical comparison, we use several models as baselines, the first of which is the original LDA model [9] that is learned using variational EM. We address this model as LDA-VEM. The second and the third models are γ-NB Process (γNBP) and Dir-PFA [55], both of which are learned using Gibbs sampling. Note that both Dir-PFA and LDA [9] have the same block Gibbs sampling and variational Bayes inference equations. Hence, we use Dir-PFA to facilitate an inference with Gibbs sampling in the original LDA model. All of these models, however, ignore the temporal information in the corpus. Note that the γNBP model used as a baseline is as strong as the HDP (inferred with the Chinese restaurant franchise representation), as shown in Zhou and Carin [55]. In fact, one can show that a normalized γNBP can be reduced to HDP.

The dynamic topic model (DTM) [8] and the Pólya-Gamma multinomial dynamic topic model (PGMult) [31], which are capable of incorporating the time-stamps associated with each document, are

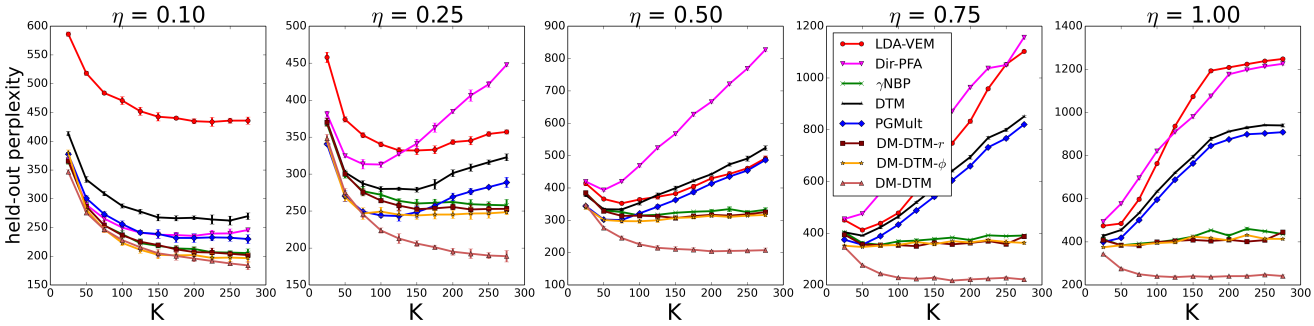

Figure 6: Performance Comparison on Business News Corpus 2

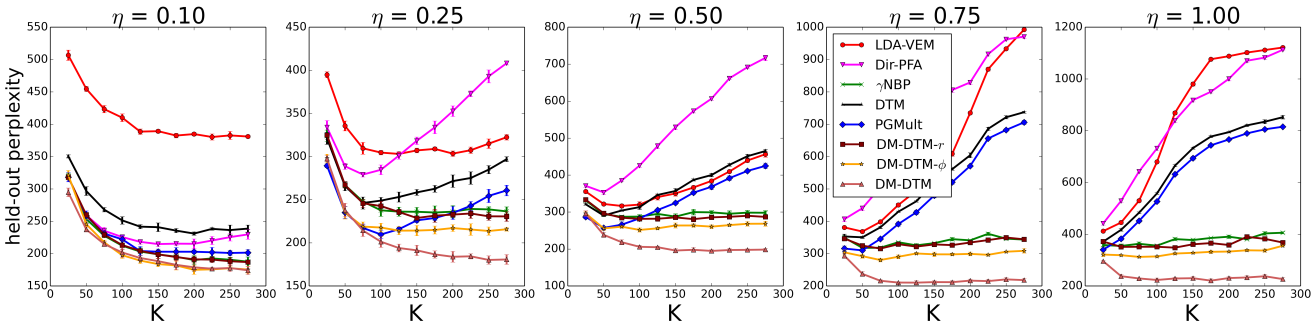

Figure 7: Performance Comparison on Business News Corpus 3

used as stronger baselines. Despite the extensive literature on temporal topic models as listed in the related work section, we, unfortunately, did not find an open-source implementation for rest of the models. For a thorough understanding of the effect of the two chains – the gamma Markov chain and the Dirichlet Markov chain, we also tried two different ablations of the DM-DTM model– DM-DTM-r and DM-DTM-ϕ. In DM-DTM-r, we just maintain the gamma Markov chain (on the r_{tk}'s) and assume a global set of topics (ϕ_{wk}'s) that explain the observations in all time slices. This implies that the global popularities of the topics change with time, but the topic-word assignments do not. In DM-DTM-ϕ, we only maintain the Dirichlet Markov chain (on the $\boldsymbol{\phi}_{tk}$'s) and assume a single set of topic strengths $(r_k)_{k=1}^K$ that explain all the observations. In such model, the topic popularities do not change with time, but the topic-word assignments do. Comparison with these two models is expected to prove the utility of both the chains, instead of the isolated use of either of them.

The performances corresponding to $p = 0.70$ on the corpora listed in the previous section are compactly presented in Fig. 4, 5, 6 and 7. The results corresponding to other values of p are similar and omitted here to avoid redundancy. Each of these plots represents the average performances over 10 different runs of the aforementioned models with different values of the parameter $\eta \in \{0.10, 0.25, 0.50, 0.75, 1.00\}$ which is the parameter of the Dirichlet prior on the topic-word distribution. For both LDA-VEM and Dir-PFA the parameter α, which is the parameter for the Dirichlet

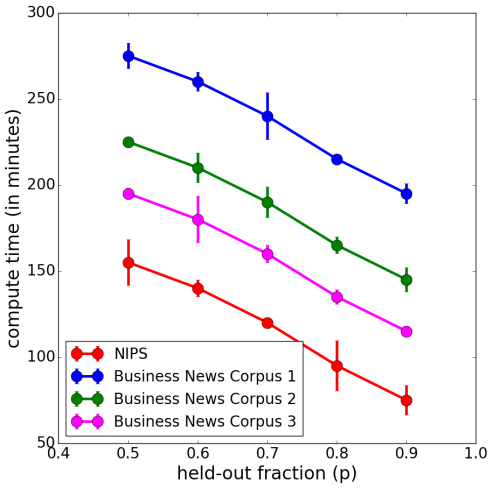

Figure 8: Compute Times for DM-DTM

prior over the document-topic distribution, is set at $50.0/K$ according to the popular choice. For all other models, all the relevant parameters (the ones that do not have any prior imposed on them) are set to 1.0. For both DTM and PGMult, we use 30 iterations for initialization of the parameters with the LDA model where the corresponding η is initialized with one of the five different values

mentioned above. For inference with the variational Kalman filtering (VKF) in DTM and the Pólya-Gamma augmentation trick in PGMult, we use 50 iterations. For LDA-VEM we use 50 iterations for variational EM. For all other models that use Gibbs sampling, we use 500 iterations for burn-in and 500 for collection. Note that, unlike in DTM and PGMult where the initialization with LDA must be done to achieve meaningful learning of the representation of the model, DM-DTM or any of its ablations does not require any special initialization, thereby bringing an additional advantage to the table.

As expected, being parametric models, LDA-VEM, Dir-PFA, DTM, and PGMult all suffer from severe overfitting as K is increased. In particular, with higher values of η, the overfitting is more prominent. With small values of η, especially with $\eta = 0.10$, the topics discovered are very sparse and hence the perplexity does not increase with increasing value of K for the parametric models. *Note that DM-DTM outperforms all the other models by a large margin.* The significant gap between DM-DTM and γNBP or Dir-PFA shows that the performance difference is not due to the adoption of Gibbs sampling for inference, but due to the congruence between the modeling assumptions and the statistical characteristics of the corpora. Similarly, the performance gap between γNBP or Dir-PFA and LDA-VEM illustrates that the adoption of Gibbs sampling, instead of variational methods, for inference makes a difference. The gap between LDA-VEM and DM-DTM clearly proves that the performance difference is both due to better modeling assumptions and better inference algorithm based on Gibbs sampling. The performance difference between DM-DTM and DM-DTM-r/DM-DTM-ϕ also justifies the use of both the chains – the gamma Markov chain and the Dirichlet Markov chain. Indeed, the performance gap among DM-DTM, PGMult, and DTM justifies the modeling assumptions and the efficacy of the inference proposed. Please note that the performance gap between DM-DTM-r and DM-DTM-ϕ is mostly negligible, possibly because both chains are equally good in capturing the temporal dependencies in the data. Additionally, to illustrate the run-time complexity of DM-DTM, we present the variation of the total compute time, as measured on a MacBook with 2.5 GHz Intel Core i7 processors and 16 GB of RAM, as a function of the held-out fraction p for all the corpora in Fig. 8. Note that a higher fraction of held-out data implies a smaller training set and compute time.

5 CONCLUSIONS AND FUTURE WORK

This paper introduced DM-DTM, a novel nonparametric Bayesian dynamic topic model that allows the topic popularities and word-topic assignments to vary smoothly over time using a gamma Markov chain and a Dirichlet Markov chain, respectively. DM-DTM is equipped with a nonparametric Bayesian construction and a tractable inference mechanism. The experiments with several real-world corpora clearly demonstrate its supremacy over many of the existing baselines. In future, the inference can get further accelerated using the formulations of stochastic gradient Langevin dynamics [32, 33] and the sampling tricks proposed in Cong et al. [14, 15]. Additionally, the gamma Markov chain and the Dirichlet Markov chain can be used to model temporal evolution of other types of dyadic count data, for example, those prevalent in recommender systems [12, 25]. Interestingly, the models can be further enriched with the split-merge techniques [10] so that the genesis and termination of topics can be explicitly accounted for in the generative assumptions. Finally, the performance of the model can potentially be improved further using ideas from adversarial training [40] and advanced variational methods [50, 52].

REFERENCES

[1] A. Acharya, J. Ghosh, and M. Zhou. 2015. Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. In *Proc. of AISTATS*. 1–9.

[2] A. Acharya, A. Saha, M. Zhou, D. Teffer, and J. Ghosh. 2015. Nonparametric Dynamic Network Modeling. In *KDD Workshop on Mining and Learning from Time Series*.

[3] A. Acharya, D. Teffer, J. Henderson, M. Tyler, M. Zhou, and J. Ghosh. 2015. Gamma Process Poisson Factorization for Joint Modeling of Network and Documents. In *Proc. of ECML*. 283–299.

[4] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. Smola, and E. Xing. 2011. Online Inference for the Infinite Topic-Cluster Model: Storylines from Streaming Text. In *Proc. of AISTATS*. 101–109.

[5] A. Ahmed and E. Xing. 2008. Dynamic Non-Parametric Mixture Models and The Recurrent Chinese Restaurant Process : with Applications to Evolutionary Clustering. *Proc. of SDM* (2008).

[6] A. Ahmed and E. Xing. 2010. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In *Proc. of UAI*.

[7] A. Bhadury, J. Chen, J. Zhu, and S. Liu. 2016. Scaling Up Dynamic Topic Models. In *Proc. of WWW*. 381–390.

[8] D. M. Blei and J. D. Lafferty. 2006. Dynamic topic models. In *Proc. of ICML*. 113–120.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *JMLR* 3 (2003), 993–1022.

[10] M. Bryant and E. B. Sudderth. 2012. Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes. In *NIPS*. 2699–2707.

[11] A. T. Cemgil. 2009. Bayesian inference for nonnegative matrix factorisation models. *Intell. Neuroscience* (2009).

[12] L. Charlin, R. Ranganath, J. McInerney, and D.M. Blei. 2015. Dynamic Poisson Factorization. In *Proc. of RecSys*. 155–162.

[13] K. Christakopoulou and A. Banerjee. 2015. Collaborative Ranking with a Push at the Top. In *Proc. of WWW*. 205–215.

[14] Y. Cong, B. Chen, H. Liu, and M. Zhou. 2017. Deep Latent Dirichlet Allocation with Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC. In *Proc. of ICML*. 864–873.

[15] Y. Cong, B. Chen, and M. Zhou. 2017. Fast Simulation of Hyperplane-Truncated Multivariate Normal Distributions. *Bayesian Analysis* (2017).

[16] M. Deodhar and J. Ghosh. 2009. Mining for Most Certain Predictions from Dyadic Data. In *Proc. of KDD*.

[17] N. Du, M. Farajtabar, A. Ahmed, A.J. Smola, and L. Song. 2015. Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams. In *Proc. of KDD (to appear)*.

[18] A. Dubey, A. Hefny, S. Williamson, and E. P. Xing. 2013. A nonparametric mixture model for topic modeling over time. In *Proc. of SDM*. 530–538.

[19] H. Elibol, V. Nguyen, S. Linderman, M. Johnson, A. Hashmi, and F. Doshi-Velez. 2016. Cross-corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders. *JMLR* 17, 1 (2016), 4597–4634.

[20] T. S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* (1973).

[21] C. Févotte, J. L. Roux, and J. R. Hershey. 2013. Non-negative dynamical system with application to speech and audio. In *Proc. of ICASSP*. 3158–3162.

[22] Q. Ho, L. Song, and E.P. Xing. 2011. Evolving Cluster Mixed-Membership Blockmodel for Time-Varying Networks. In *Proc. of AISTATS*.

[23] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. 2011. A Time-dependent Topic Model for Multiple Text Streams. In *Proc. of KDD*. 832–840.

[24] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. 2010. Online Multiscale Dynamic Topic Models. In *Proc. of KDD*. 663–672.

[25] G. Jerfel, M. Basbug, and B. Engelhardt. 2017. Dynamic Collaborative Filtering With Compound Poisson Factorization. In *Proc. of AISTATS*. 738–747.

[26] N. L. Johnson, A. W. Kemp, and S. Kotz. 2005. *Univariate Discrete Distributions*. John Wiley & Sons.

[27] M. Kim and J. Leskovec. 2013. Nonparametric Multi-group Membership Model for Dynamic Networks. In *Proc. of NIPS*. 1385–1393.

[28] Y. Koren. 2009. Collaborative Filtering with Temporal Dynamics. In *Proc. of KDD*. 447–456.

[29] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* (2009).

[30] D. D. Lee and H. S. Seung. 2001. Algorithms for Non-negative Matrix Factorization. In *NIPS*.

[31] S. Linderman, M. Johnson, and R. P. Adams. 2015. Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation. In *Proc. of NIPS*. 3438–3446.

[32] Y. Ma, T. Chen, and E. B. Fox. 2015. A Complete Recipe for Stochastic Gradient MCMC. In *Proc. of NIPS*. 2917–2925.

[33] Y. Ma, N. J. Foti, and E. B. Fox. 2017. Stochastic Gradient MCMC Methods for Hidden Markov Models. In *Proc. of ICML*. 2265–2274.

[34] S.N. MacEachern. 2000. *Dependent Dirichlet Process*. Technical Report. Department of Statistics, The Ohio State University.

[35] R.M. Nallapati, S. Ditmore, J.D. Lafferty, and K. Ung. 2007. Multiscale Topic Tomography. In *Proc. of KDD*. 520–529.

[36] N. Natarajan and I.S. Dhillon. 2014. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 30, 12 (2014), 60–68.

[37] N. G. Polson, J. G. Scott, and J. Windle. 2013. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *J. Amer. Statist. Assoc.* 108, 504 (2013), 1339–1349.

[38] S. Raghavan, S. Gunasekar, and J. Ghosh. 2012. Review Quality Aware Collaborative Filtering. In *Proc. of RecSys*. 123–130.

[39] A. Schein, H. Wallach, and M. Zhou. 2016. Poisson-Gamma dynamical systems. In *Proc. of NIPS*. 5005–5013.

[40] J. Song, S. Zhao, and S. Ermon. 2017. A-NICE-MC: Adversarial Training for MCMC. In *Proc. of NIPS*. 5146–5156.

[41] N. Srebro and S. Roweis. 2005. Time-Varying Topic Models using Dependent Dirichlet Processes.

[42] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* 101 (December 2006), 1566–1581.

[43] T. Virtanen, A.T. Cemgil, and S. Godsill. 2008. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. of ICASSP*. 1825–1828.

[44] D. D. Walker, K. Seppi, and E. K. Ringger. 2012. Topics over Nonparametric Time: A Supervised Topic Model Using Bayesian Nonparametric Density Estimation. In *Proc. of UAI*. 74–83.

[45] C. Wang, D. Blei, and D. Heckerman. 2008. Continuous Time Dynamic Topic Models. In *Proc. of UAI*.

[46] P. Wang, P. Zhang, C. Zhou, Z. Li, and H. Yang. 2017. Hierarchical Evolving Dirichlet Processes for Modeling Nonlinear Evolutionary Traces in Temporal Data. *Data Min. Knowl. Discov.* 31, 1 (Jan. 2017), 32–64.

[47] X. Wang and A. McCallum. 2006. Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In *Proc. of KDD*. 424–433.

[48] L. Xiong, X. Chen, T. Huang, J. Schneider, and J. G. Carbonell. 2010. Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization. In *Proc. of SDM*.

[49] K.S. Xu and A.O. Hero. 2014. Dynamic Stochastic Blockmodels for Time-Evolving Social Networks. *J. Sel. Topics Signal Processing* 8, 4 (2014), 552–562.

[50] M. Yin and M. Zhou. 2018. Semi-implicit variational inference. In *Proc. of ICML*.

[51] K. Zhai and J.L. Boyd-graber. 2013. Online Latent Dirichlet Allocation with Infinite Vocabulary. In *Proc. of ICML*. 561–569.

[52] H. Zhang, D. Guo, B. Chen, and M. Zhou. 2018. WHAI: Weibull Hybrid Autoencoding Inference for Deep Topic Modeling. In *Proc. of ICLR*. to appear.

[53] M. Zhou. 2016. Nonparametric Bayesian Negative Binomial Factor Analysis. (Oct 2016).

[54] M. Zhou and L. Carin. 2012. Augment-and-Conquer Negative Binomial Processes. In *Proc. of NIPS*.

[55] M. Zhou and L. Carin. 2015. Negative Binomial Process Count and Mixture Modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence* (2015).

[56] M. Zhou, L. Hannah, D. Dunson, and L. Carin. 2012. Beta-Negative Binomial Process and Poisson Factor Analysis. In *Proc. of AISTATS*. 1462–1471.