Extraction of Topic Evolutions from References in Scientific Articles and Its GPU Acceleration

Tomonari Masada
Nagasaki University
1-14 Bunkyo-machi, Nagasaki-shi
Nagasaki, 852-8521 Japan
masada@nagasaki-u.ac.jp

Atsuhiro Takasu
National institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, 101-8430 Japan
takasu@nii.ac.jp

ABSTRACT

This paper provides a topic model for extracting topic evolutions as a corpus-wide transition matrix among latent topics.

Recent trends in text mining point to high demand for exploiting metadata, i.e., data about data. Especially, exploitation of reference relationships among documents induced by e.g. hyperlinking Web pages, citing scientific articles, tumblring blog posts, retweeting tweets, etc, is put in the foreground of the effort to achieve an effective mining. We focus on scholarly activities and propose a new topic model for obtaining a corpus-wide view on how research topics evolve along reference relationships among scientific articles. Our topic model, called TERESA, extends latent Dirichlet allocation (LDA) by introducing a corpus-wide topic transition probability matrix, with which we model reference relationships among articles as transitions between latent topics. Our approximated variational inference updates LDA posteriors and topic transition posteriors alternately. The main issue is execution time, because the time complexity is proportional to MK^2, where K is the number of topics and M is the number of reference relationships (i.e., links in citation network). Therefore, we accelerate the inference with Nvidia CUDA. We prove the effectiveness of TERESA by introducing a new evaluation measure called diversity plus focusedness (D+F). We also present examples of topic evolutions extracted by our method.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Topic modeling, Citation analysis, GPU

1. INTRODUCTION

Intensive analysis of reference relationships among documents is a key factor to success in today's text mining. In fact, this type of relationship prevails. For example, hyperlinking Web pages, citing academic articles, tumblring blog posts, retweeting tweets, etc. induce such relationships among documents. Classic link analysis [17, 9] puts less emphasis on text data found in each document. In contrast, we adopt LDA [4], a promising framework for effective text mining, and extend it for exploiting reference relationships.

In this paper, we focus on academic activities and propose a new LDA-like topic model for analyzing reference relationships among scientific articles. Since our model extracts *T*opic *E*volutions from *RE*ferences in *S*cientific *A*rticles, we call it TERESA by concatenating the italicized capital letters. The most important outcome obtained by TERESA is a corpus-wide topic transition matrix whose (i, j)th entry gives the posterior transition probability from topic t_j to topic t_j. Roughly, when the articles containing topic t_i with a high probability are often cited by the articles containing topic t_j with a high probability, the (i, j)th entry of this matrix gives a high transition probability. We can interpret this matrix as a summary of frequent topic evolutions latent in an article set. We call a task to extract a set of corpus-wide topic transition patterns *topic evolution analysis*.

The main contributions of this paper are as follows:

1. This paper proposes a method for topic evolution analysis, where we can obtain a corpus-wide and directed relationships among topics. First, TERESA provides a *corpus-wide* view on how research topics evolve along reference relationships. Our aim is not to obtain an accumulation of local views, e.g. a series of topic transition patterns between two adjacent time points along the time axis [3, 18]. This type of accumulation of local views makes it difficult to get a global view on how topics interrelate. Second, TERESA extracts *directed* relationships among topics. Therefore, we can respect the direction intrinsic to every citation, though some preceding works neglect the direction by regarding each relationship as symmetric [15, 5]. For evaluation, we devise a new measure called *diversity plus focusedness* (D+F) and compare TERESA with LDA.

2. This paper proposes an acceleration of inference for TERESA. We employ variational Bayesian inference (VB) for obtaining an estimation of the posterior parameters. In our VB, a large number of posterior parameters can be updated independently. Therefore,

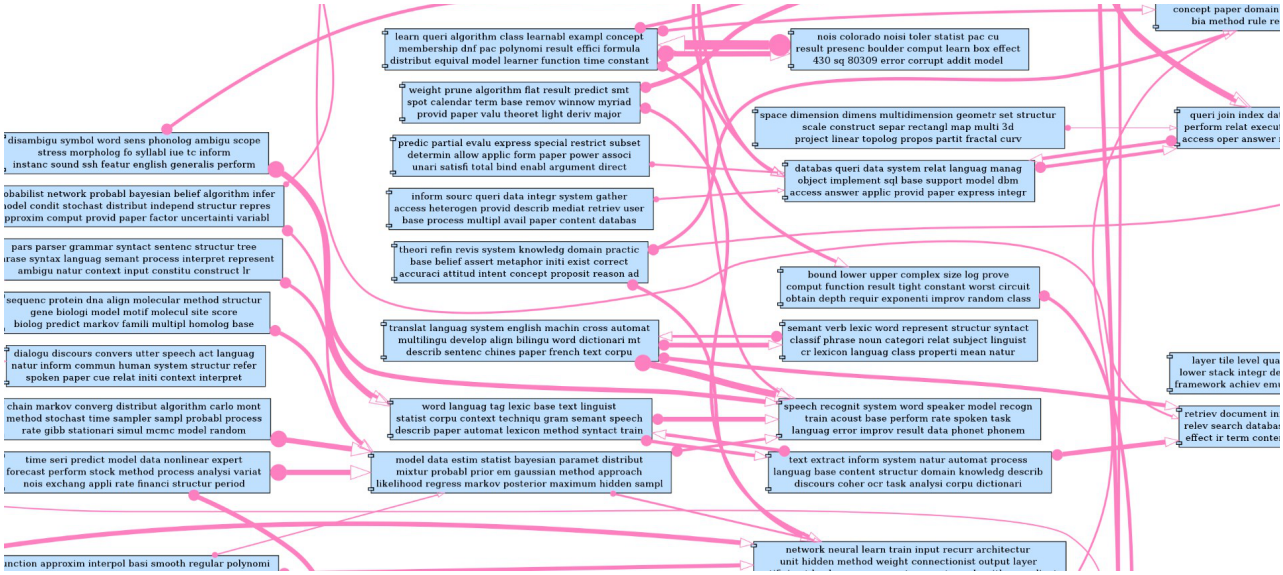

Figure 1: A trimmed portion of a topic evolution network extracted from the CORA data set. The number of topics is 300. Each topic is presented by the most frequent 21 words. Thicker arrows correspond to topic transitions of larger probability. We can find interesting relationships among the topics related to machine learning, natural language processing, data mining, etc. This global view on how topics evolve in the CORA data set was obtained as a result of variational Bayesian inference executed on Nvidia GTX580.

an efficient parallelization exists. To implement such a parallelization, we use Nvidia CUDA compatible devices, on which hundreds of threads can run in parallel.

Figure 1 is an example of the result of topic evolution analysis by TERESA. Each topic is presented by the 21 most frequent words placed in the same box, and each topic evolution is depicted as a directed arrow. This result is obtained by VB accelerated with Nvidia GTX580. We call the diagrams given by TERESA *topic evolution network*.

This paper is organized as follows. Section 2 gives preceding works related to our approach. Section 3 presents the detailed description of TERESA as well as a parallelization method with Nvidia CUDA compatible devices. Section 4 contains the definition of our new evaluation measure and the results of evaluation experiment. Section 5 concludes the paper with discussions and future works. For the rest of the paper, we will use the symbols defined in Table 1.

2. RELATED WORKS

In social media, we meet a wide variety of interactions among entities, e.g., persons, documents, products, places, images, movies, etc. Compelling data mining methods are expected to exploit such relationships observable among various types of entities. This paper focuses on reference relationships among scientific articles, because this type of relationship may help us to find papers we can cite and to predict what kind of research will follow ours.

Many methods are proposed to use reference relationships among scientific articles for topic extraction. However, our method is different from them as is discussed below.

2.1 Model Complexity

Nallapati et al. [14] propose a method for modeling topic transitions among documents. While we can find the works that do not give any direct modelings of relationships of topic distributions between cited and citing documents [13, 7, 11], Nallapati et al. explicitly model transitions of topic distributions among documents as "flows" of topics. However, their probabilistic model has K flow parameters at every link in citation network. Therefore, the complexity of the parameters for modeling citation links amounts to $O(MK)$, where K is the number of topics and M is the number of references, i.e., the number of links in citation network. It can be said that this model aims at giving an accumulation of local transition patterns, and thus that it is difficult to obtain a global view on how topics are related to each other. We may repeat the same discussion with respect to [19], where relationships of topic distributions among documents are modeled as Markov random fields. Further, Ren et al. [18] modifies HDP [20] so that we can distill topic transitions between adjacent time points along the time axis. That is, they consider topic transitions not between linked documents as in [14] but between adjacent time points. Therefore, the complexity of the parameter space is reduced. However, this proposal also makes us uncomfortable with respect to the issue how we can integrate a pile of time-dependent topic transitions to obtain a unified view on how topics interrelate. A similar discussion may also be valid for [3, 22].

In contrast, TERESA extracts a single $K \times K$ matrix of topic transition probabilities and thus has a smaller number of parameters for modeling citation network of scientific articles. In this manner, we obtain a single corpus-wide view of topic interaction as a digraph whose vertices are latent topics and whose weighted arcs are transitions among topics accompanied with their probabilities (Figure 1). Therefore, the result of our topic evolution analysis is easy to grasp.

2.2 Computation Acceleration

Table 1: Definitions of Symbols

$\mathcal{D} = \{d_1, \ldots, d_J\}$	set of documents
$\mathcal{V} = \{v_1, \ldots, v_W\}$	set of words
$\mathcal{T} = \{t_1, \ldots, t_K\}$	set of latent topics
$\mathcal{C}_j = \{d_{j_1}, \ldots, d_{j_{C_j}}\}$	set of documents cited by document d_j
$\boldsymbol{x}_j = \{x_{j1}, \ldots, x_{jn_j}\}$	word tokens in document d_j
$\boldsymbol{z}_j = \{z_{j1}, \ldots, z_{jn_j}\}$	topic assignments in document d_j
$\theta_j = (\theta_{j1}, \ldots, \theta_{jK})$	θ_{jk}, probability that a word token in document d_j is assigned to topic t_k.
$\phi_k = (\phi_{k1}, \ldots, \phi_{kW})$	ϕ_{kw}, probability that a token of word v_w is assigned to topic t_k.
$\boldsymbol{r}_k = (r_{k1}, \ldots, r_{kK})$	$r_{kk'}$, transition probability from topic t_k to topic $t_{k'}$.
$\alpha = (\alpha_1, \ldots, \alpha_K)$	parameters of the Dirichlet prior for the topic multinomial parameters $\theta_1, \ldots, \theta_J$
$\beta = (\beta_1, \ldots, \beta_W)$	parameters of the Dirichlet prior for the word multinomial parameters $\phi_1, \ldots, \phi_K$
$\gamma = (\gamma_1, \ldots, \gamma_K)$	parameters of the Dirichlet prior for topic transition multinomial parameters $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K$
$\pi_{jw} = (\pi_{jw1}, \ldots, \pi_{jwK})$	variational probabilties that a token of word v_w in document d_j is assigned to each topic
$\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jK})$	variational Dirichlet posterior parameters corresponding to θ_j
$\mu_k = (\mu_{k1}, \ldots, \mu_{kW})$	variational Dirichlet posterior parameters corresponding to ϕ_k
$\nu_k = (\nu_{k1}, \ldots, \nu_{kK})$	variational Dirichlet posterior parameters corresponding to $\boldsymbol{r}_k$
τ	the mixing proportion of the topic multinomial parameters of the cited documents
$\omega_{jj'}$	the weight of document $d_{j'}$ with respect to document d_j

Existing topic models provide an analysis of reference relationships mainly with the following two approaches: establishing links among topic multinomial probabilities via Dirichlet processes [18, 24] or via logistic normal distributions[14, 23]. The inference for the former approach is complicated due to its nonparametric nature. This type of approach is not discussed any further in this paper. The latter approach is also intricate, because the model has a part of its parameters in $K \times K$ dense covariance matrices, which need to be inverted. Therefore, covariance matrices used for logistic normal are often assumed to be diagonal [3, 1, 8]. One important reason this approach avoids dense matrices is that it considers topic transitions at many different places in observed data, e.g. at all citation links or at all pairs of adjacent time points along the time axis, as is discussed in Section 2.1. Consequently, many matrices should be taken into consideration in modeling, and thus the assumption of sparseness is required to accelerate inference.

In contrast, TERESA has a single $K \times K$ matrix to model topic transition probabilities. Therefore, we have no reason to make this matrix sparse. However, our VB inference has a time complexity proportional to MK^2, where M is the number of citation links. Therefore, we have decided to use GPU to accelerate the inference.

3. PROPOSAL

3.1 Generative Description

First, we describe how documents are generated by our topic model, TERESA. We extend LDA by introducing a corpus-wide topic transition matrix $\boldsymbol{R}$. The (k, k')th entry $r_{kk'}$ of $\boldsymbol{R}$ means the probability of the transition from topic t_k to topic $t_{k'}$. TERESA generates documents as follows:

1. For each topic $t_k \in \mathcal{T}$, $k = 1, \ldots, K$, draw a word multinomial distribution Multi(ϕ_k) from the corpus-wide word Dirichlet prior distribution Dir(β). The wth entry ϕ_{kw} of ϕ_k is the probability that word $v_w \in \mathcal{V}$ is used to express topic t_k. The wth hyperparameter β_w of Dir(β) corresponds to word v_w.
2. For each topic $t_k \in \mathcal{T}$, $k = 1, \ldots, K$, draw a topic transition multinomial distribution Multi($\boldsymbol{r}_k$) from the corpus-wide topic transition Dirichlet prior distribution Dir(γ). The k'th entry $r_{kk'}$ of $\boldsymbol{r}_k$ is the transition probability from topic t_k to topic $t_{k'}$. The k'th hyperparameter $\gamma_{k'}$ of Dir(γ) corresponds to topic $t_{k'}$ appearing in the citing documents, not in the cited documents. The corpus-wide topic transition probability matrix $\boldsymbol{R}$ is obtained by setting the kth row of $\boldsymbol{R}$ to $\boldsymbol{r}_k^T$, i.e., $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K)^T$.
3. For each document $d_j \in \mathcal{D}$, $j = 1, \ldots, J$,
   (a) Draw a topic multinomial distribution Multi(θ_j) from the corpus-wide topic Dirichlet prior distribution Dir(α). The kth entry θ_{jk} of θ_j is the probability that any word in document d_j is used to express topic t_k. The kth hyperparameter α_k of Dir(α) corresponds to topic t_k.
   (b) Let $\mathcal{C}_j \subseteq \mathcal{D}$ be the set of the documents cited by document d_j. We obtain a new topic multinomial distribution with parameters $(\eta_{j1}, \ldots, \eta_{jK})$ by combining topic multinomial parameters linearly as follows:
   $$\eta_j \equiv (1 - \tau)\theta_j + \tau \sum_{j'} \omega_{jj'} \boldsymbol{R}^T \theta_{j'} \qquad (1)$$
   where $\omega_{jj'}$ is the weight of document $d_{j'}$ with respect to document d_j. $\sum_{j'} \omega_{jj'} = 1$ holds for every j, where $\omega_{jj'} = 0$ for $d_{j'} \notin \mathcal{C}_j$. For simplicity, we set $\omega_{jj'} = 1/C_j$ when $d_{j'} \in \mathcal{C}_j$. Obviously, TERESA is reduced to LDA when $\tau = 0$. In the evaluation experiment, we tested the values from 0.1 to 0.9 with step size 0.1 for τ.
   (c) Let n_j be the length of document d_j. For each word token x_{ji}, $i = 1, \ldots, n_j$, draw a topic from Multi(η_j) and set the drawn topic as the value of the latent variable z_{ji}. Then, draw a word from Multi($\phi_{z_{ji}}$) and set the drawn word as the value of the observed variable x_{ji}.

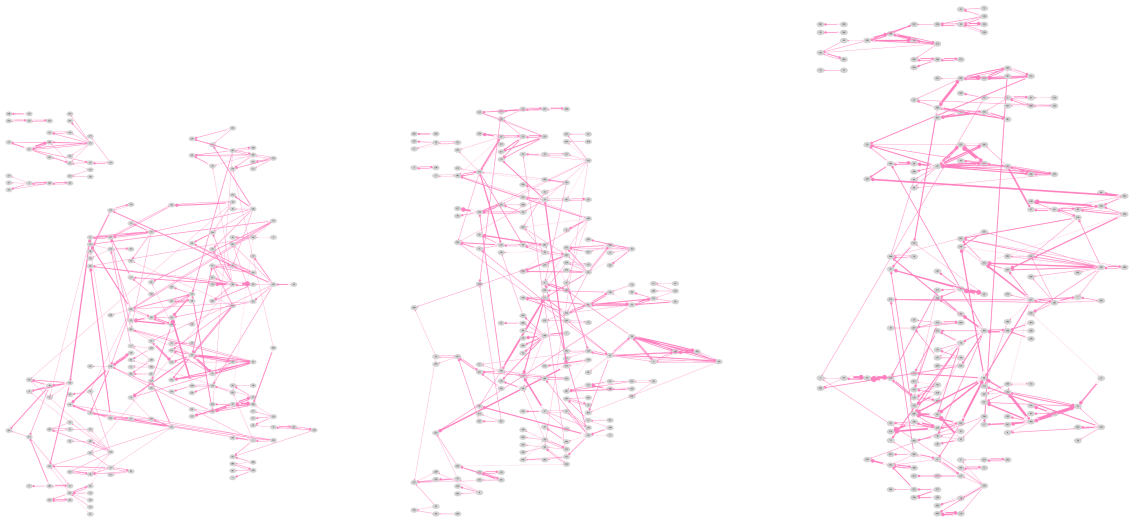

Figure 2: Three examples of topic evolution networks extracted by TERESA. In this figure, topics are depicted just as a gray oval. All examples are obtained for the HEP-PH data set with the same settings of experiment ($K = 300, \tau = 0.2$). Each instance of our VB inference is initialized by randomly assigning each word tokens to a topic. Therefore, different results are obtained for different instances of VB inference as in case of LDA. Practically, TERESA can be applied to the same data set repeatedly, and then users can use any of the results to discover interesting transition patterns.

3.2 Inference and Its Parallelization

Second, we describe how the posterior distributions of TERESA can be inferred. TERESA extends LDA by introducing a topic transition matrix $\boldsymbol{R}$. However, this matrix $\boldsymbol{R}$ is of size $O(K^2)$ and increases the computational burden of inference. While the time complexity of the inference for LDA is only proportional to K, that for TERESA is proportional to K^2. Therefore, we need to accelerate our inference.

We employ variational Bayesian inference (VB) [4] for an approximated estimation of posterior parameters, because VB is suitable for an efficient parallelization. Zhai et al. subtly discuss why VB can be efficiently parallelized [26]. However, we do not consider a parallelization with OpenMP and/or MPI, which is considered by Zhai et al. We consider a parallelization with Nvidia CUDA compatible devices, because our VB frequently performs the same operation on hundreds of different data simultaneously and favors SIMD architecture, which is implemented by Nvidia GPUs. In this paper, we assume that the input data can be stored on the device memory of GPU without being split into many subsets. Therefore, VB is not seriously affected by the latency during a data transfer between CPU and GPU, because such transfer occurs only before and after the computation on GPU, not in the course of the computation.

With respect to VB for LDA, the inaugural work by Blei et al. [4] contains detailed descriptions. Therefore, most details can be referred to [4]. Here we show the update formulae for the posterior parameters without derivation. Appendix A only gives derivations required for TERESA. Our approximated variational inference updates LDA posteriors and topic transition posteriors alternately.

Let π_{jwk} be the variational posterior probability that word v_w expresses topic t_k in document d_j. Note that $\sum_k \pi_{jwk} = 1$ for every pair of j and w. Let $\mathrm{Dir}(\lambda_j)$ be the variational Dirichlet posterior distribution for the topic multinomial distribution $\mathrm{Multi}(\theta_j)$. Further, let $\mathrm{Dir}(\mu_k)$ be the variational Dirichlet posterior distribution for the word multinomial distribution $\mathrm{Multi}(\phi_k)$. Then, based on the discussion in [4],

$$\pi_{jwk} \propto \exp \Psi(\lambda_{jk}) \cdot \frac{\exp \Psi(\mu_{kw})}{\exp \Psi(\mu_{k0})}$$
$$\mu_{kw} = \beta_w + \sum_j n_{jw} \pi_{jwk} \tag{2}$$

where $\Psi(\cdot)$ represents digamma function, μ_{k0} is defined to be $\sum_w \mu_{kw}$, and n_{jw} is the number of the occurrences of word v_w in document d_j. Obviously, $(\pi_{jw1}, \ldots, \pi_{jwK})$ can be updated in parallel for different pairs of j and w. Further, μ_{kw} can be updated independently for different pairs of k and w. Therefore, the updates in Eq. (2) are efficiently parallelized by using GPU.

With respect to λ_{jk}, we use the following update for LDA:

$$\lambda_{jk} = \alpha_k + \langle n_{jk} \rangle, \tag{3}$$

where $\langle n_{jk} \rangle \equiv \sum_w n_{jw} \pi_{jwk}$. However, in TERESA, the generation of word tokens in document d_j is affected not only by the topic distribution of d_j but also by those of the documents cited by d_j. Therefore, we obtain a completely different update formula. Let $\mathrm{Dir}(\nu_k)$ be the variational Dirichlet posterior for the topic transition multinomial $\mathrm{Multi}(\boldsymbol{r}_k)$. Then, λ_{jk} can be updated by using the equation below:

$$\lambda_{jk} \approx \alpha_k + (1-\tau)\langle n_{jk} \rangle + \frac{1}{\Psi'(\lambda_{jk})} \cdot \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \big\{ \Psi(\nu_{kk'}) - \Psi(\nu_{k0}) \big\} \tag{4}$$

where $\nu_{k0} \equiv \sum_{k'} \nu_{kk'}$, $\lambda_{j0} \equiv \sum_k \lambda_{jk}$, and $\Psi'(\cdot)$ is trigamma function. Recall that $\omega_{j'j} = 0$ when document d_j is not cited by document $d_{j'}$. Further, note that λ_{j0} is equal to $\sum_k \alpha_k + n_j$ and thus does not depend on λ_{jk}. Eq. (4) is of the form $x = a + b/\Psi'(x)$, which can be solved by a

binary search for the constants a and b. Therefore, based on Eq. (4), we can obtain an updated value for λ_{jk}. Since the right hand side of Eq. (4) does not contain $\lambda_{jk'}$ for $k' \neq k$ and λ_{j0} is a constant when α_ks are fixed, we can update λ_{jk} in parallel for different pairs of j and k.

However, the approximation we introduce for obtaining Eq. (4) is not good when λ_{jk} is small (cf. Eq. (14)). Consequently, λ_{j0} becomes different from $\sum_k \alpha_k + n_j$ after updating each λ_{jk}. Therefore, we rescale $\lambda_{j1}, \ldots, \lambda_{jK}$ to make λ_{j0} equal to $\sum_k \alpha_k + n_j$ after an update of K parameters $\lambda_{j1}, \ldots, \lambda_{jK}$ for a fixed j. Inferences with less approximation can be achieved by a gradient-based method. Although we implemented an optimization using L-BFGS [16, 10], the running time was prohibitively long, because we should call L-BFGS once for each document in each iteration to solve a K-dimensional optimization problem. Therefore, it is reserved as future work to make gradient-based inferences practical.

The topic transition posterior parameters $\nu_{kk'}$ are updated by using the following formula:

$$\nu_{kk'} = \gamma_{k'} + \tau \sum_j \langle n_{jk'} \rangle \sum_{j'} \omega_{jj'} \frac{\lambda_{j'k}}{\lambda_{j'0}} \tag{5}$$

Since $\lambda_{j'k}/\lambda_{j'0}$ is the posterior probability of topic t_k in document $d_{j'}$, $\nu_{kk'}$ is updated by using linear combinations of topic probabilities along citation links. This is occasionally similar to the proposal in [14]. The derivation of the above update formulae for λ_{jk} and $\nu_{kk'}$ is included in Appendix A.

3.3 Implementation

We implemented the inference for TERESA as follows: 1) Initialize the assignment of each word token to a topic randomly selected from $\{t_1, \ldots, t_K\}$; 2) Run 500 iterations of collapsed Gibbs sampling (CGS) for LDA [6]; 3) Initialize the posteriors π_{jwk}, λ_{jk}, and μ_{kw} of TERESA by using the result of CGS, and initialize $\nu_{kk'}$ by using these initialized posteriors; and 4) Run 50 iterations of VB for TERESA. The Dirichlet hyperparameters are updated by Minka's method [12].

We have tested many settings for GPU acceleration of VB. We obtained various wall-clock running times as in Table 2. Each running time is the time required for one iteration of VB, which is obtained by dividing the total running time of 50 iterations of VB by 50. We used GPU only for VB, not for the initialization by CGS for LDA. Table 2 also contains per-iteration running times obtained for Intel Core i7 CPUs. Therefore, we can compare the efficiency achieved by multithreading on Core i7 CPUs with that achieved by multithreading on GeForce GPUs. We fixed the number of GPU threads to 512 for ease of comparison between different types of GPU.

While it is known that collapsed variational Bayesian inference (CVB) [21, 2] is also suitable for parallelization by GPUs [25], it is reserved as future work.

4. EXPERIMENT

4.1 Data Sets

We employed two data sets available from the Web in our experiment. Their specifications are presented in Table 3.

The one data set is the CORA data set[1], often used for the experiments related to citation analysis. The other is the HEP-PH data set, a set of tex documents along with a citation graph available at KDD Cup 2003 Web site[2]. Each tex document in the HEP-PH data set corresponds to a paper in the hep-ph portion of the arXiv. For both data sets, we removed stop words and applied a Porter's stemmer. We did not remove any tex commands from the HEP-PH data set, because we wanted to know what kind of special symbols (e.g. `\langle`, `\widehat`, `\dag`, etc) are likely to be used in a particular discipline of physics.

Table 2: Running Time per VB Iteration (sec).

CORA data set ($K = 300$)	
8 threads on i7-2600K @ 3.40GHz	3,000
8 threads on i7 950 @ 3.07GHz	2,200
8 threads on i7 X 990 @ 3.47GHz	1,500
12 threads on i7-3930K @ 3.20GHz	710
1 block of size 512 on GTX 580 @ 1.54GHz	170
1 block of size 512 on GTX 570 @ 1.56GHz	166
HEP-PH data set ($K = 300$)	
8 threads on i7 970 @ 3.20GHz	29,000
1 block of size 512 on GTX 580 @ 1.54GHz	840
1 block of size 512 on GTX 580 @ 1.59GHz	810

We checked the soundness of inference by calculating perplexity [4] for 10% randomly chosen word tokens, whose number is denoted by N_{test}. The perplexity is defined as

$$perplexity \equiv \exp\Big\{ -\sum_j \sum_i \log \sum_k \lambda_{jk} \mu_{kx_{ji}} \Big/ N_{test} \Big\}. \tag{6}$$

We had a perplexity around 570 for the CORA data set after CGS for LDA, and the perplexity was not significantly modified by VB for TERESA. For the HEP-PH data set, we had a perplexity around 710 after CGS for LDA, and the perplexity was also not significantly modified by VB for TERESA. These are the perplexities obtained when $K = 300$. Since TERESA did not significantly modify the perplexity obtained as a result of CGS for LDA, we could extract a corpus-wide view of topic transitions by TERESA without affecting the generalization power of LDA.

4.2 Evaluation Method

As a quantitative measure for evaluating the quality of topic evolution analysis, we devise a new measure called *diversity plus focusedness (D+F)* based on the posterior probabilities of topic transitions. Note that the posterior probability of the transition from topic t_k to topic $t_{k'}$ is $\nu_{kk'}/\nu_{k0}$.

D+F score is defined by combining the two measures, *transition diversity* and *expected focusedness*:

- *Transition diversity.* Let $P(t_k) = \sum_j \lambda_{jk} / \sum_{j,k} \lambda_{jk}$ and $P_{Tr}(t_{k'}|t_k) = \nu_{kk'}/\nu_{k0}$. $P(t_k)$ is the posterior probability of the occurrence of topic t_k. Then the posterior probability of the transition to topic t_k can be written as $P_{To}(t_k) = \sum_{k'} P(t_{k'}) P_{Tr}(t_k|t_{k'})$. By using $P_{To}(t_k)$, we define *transition diversity* as follows:

$$TrDiv \equiv -\sum_k P_{To}(t_k) \log P_{To}(t_k). \tag{7}$$

[1] `http://people.cs.umass.edu/~mccallum/data.html`

[2] `http://www.cs.cornell.edu/projects/kddcup/datasets.html`

Figure 3: Two trimmed portions of topic evolution networks extracted by TERESA from the CORA data set for $K = 300$. Each topic is represented by the most frequent 21 words. Thicker arrows correspond to topic transitions of larger probability. VB was run on Nvidia GTX580 @ 1.59GHz . The running time of the inference containing CGS for LDA was around 9,200 seconds with Intel Core i7-2600K @ 3.40GHz.

$TrDiv$ is an entropy measure. When $TrDiv$ is large, the transition to each topic occurs equally often. This means that the topic model under evaluation can extract topics so that every topic joins a corpus-wide topic evolution equally. In other words, no topics are neglected in topic evolution analysis.

- We define a measure called *focusedness* as follows:

$$Foc(t_k) \equiv \sum_{k'} P_{Tr}(t_{k'}|t_k) \log P_{Tr}(t_{k'}|t_k), \qquad (8)$$

which is a negative entropy measure and is thus less than or equal to 0. When $Foc(t_k)$ is close to 0, only a limited number of topics are frequently reached from topic t_k. Further, we obtain an expected focusedness as follows:

$$EFoc \equiv \sum_{k} P(t_k) Foc(t_k). \qquad (9)$$

When $EFoc$ is large, the probability distribution of the transitions from any topic is highly skewed. In other words, only a limited number of transition edges starting from each node have a large probability.

D+F is defined to be the sum of $TrDiv$ and $EFoc$. When D+F is large, all topics are equally visited, and, at the same time, the choice of transition destination is highly selective. We contend that topic evolution analyses that can give larger D+F are better. We give a rationale for this contention by considering three extreme cases.

First, assume that the transition probability matrix is the identity matrix. Then $TrDiv = -\sum_k P(t_k) \log P(t_k)$ and $EFoc = 0$. Therefore, D+F is equal to the entropy $-\sum_k P(t_k) \log P(t_k)$, which is larger when $P(t_k)$s show less differences. In this case, we extract the topics that are totally "independent" in the sense that each topic only transits to itself. Of course, articles in reality may cite articles from heterogenous research fields. However, it is desirable to extract relatively independent components where the transitions among different components rarely occur. Therefore, it can be said that a larger D+F is better.

Second, assume that the transition probability matrix is a matrix all of whose columns except one are zero vectors. Then $TrDiv = 0$ and $EFoc = 0$. Therefore, D+F is 0. In our experiment, we found that smaller values of D+F corresponded to an inference of poor quality. To be precise, it was observed that, when the inference was not successful, a large number of transitions pointed to a limited number of topics, which correspond to a vague content. Typically, such topics were represented by the words expressing vague concepts or by the words similar to stop words. In this case, the transition matrix had large values only in a limited number of columns. Therefore, it can be said that D+F should be substantially larger than 0.

Third, assume that all entries of the transition probability matrix are $1/K$. Then $TrDiv = \log K$ and $EFoc = -\log K$. Therefore, D+F is 0. This matrix means that all transitions are equally frequent, and thus the matrix corresponds to no meaningful topic evolution analyses. It is confirmed again that D+F should be substantially larger than 0.

The latter two types of topic transition probability matrix is not informative, and the first one is, in a sense, an unreachable ideal. Therefore, we consider that a larger D+F value

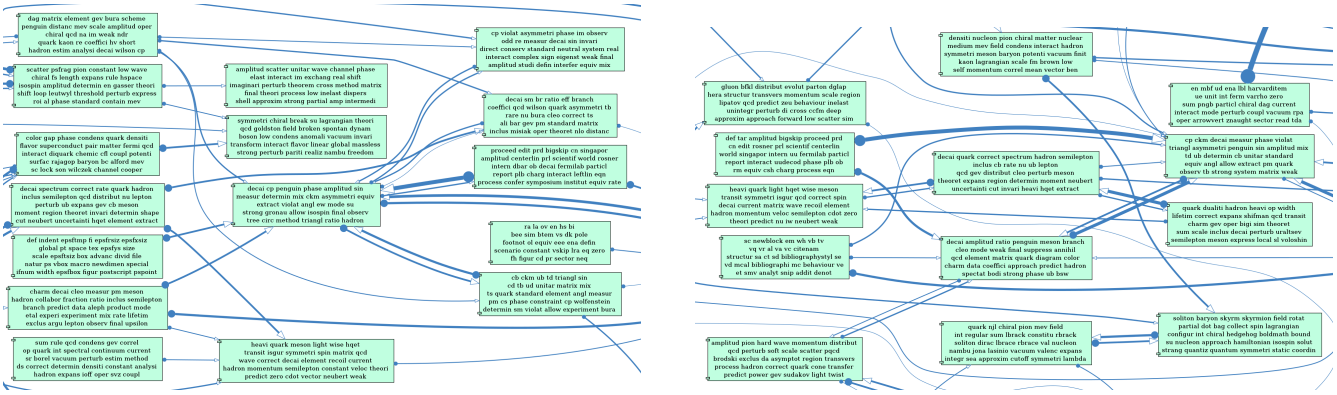

Figure 4: Two trimmed portions of topic relationships extracted by TERESA from the HEP-PH data set. Each topic is represented by the most frequent 30 words. Thicker arrows correspond to topic transitions of larger probability. VB was run on Nvidia GTX580 for $K = 300$. The running time was around 67,000 seconds.

corresponds to a better topic evolution analysis, though we cannot reach an extreme case where the transition matrix is the identity matrix. Each document set may have its own largest among the reachable D+F scores.

4.3 Evaluation Results

Figures 5 and 6 present the evaluation results for the CORA data set when $K = 300$ and $K = 200$, respectively, where we use the proposed measure D+F. Since these two figures makes no large difference, we only provide the results for $K = 300$ with respect to the HEP-PH data set in Figure 7. All line graphs contain the results when we choose the value of τ from 0.1 to 0.9 with step size 0.1. The horizontal axis represents τ. The vertical axis represents D+F score. We executed ten instances of VB inference, and the mean of the corresponding ten D+F scores is presented at each value of τ. When $\tau = 0$ and $\tau = 1$, D+F was 0, though not depicted in the graphs. The error bars on each graph show the width of plus/minus one standard deviation.

Solid and dashed lines show the results after 50 iterations and those after one iteration of VB inference, respectively. Therefore, dashed lines correspond to the cases where $\nu_{kk'}$ is obtained directly from the results of CGS for LDA by applying Eq. (5). In contrast, solid lines correspond to the cases where we alternately update LDA posteriors and topic transition posteriors 50 times. Namely, the difference between the solid line and the dashed line in each graph corresponds to the difference between TERESA and LDA. As is shown in each graph, TERESA can give a better topic evolution analysis for smaller values of τ. However, when τ is close to 1, TERESA seems to put too much emphasis on citation data and thus fails to achieve a balanced inference over text data and citation data. We can recommend 0.1 or 0.2 for the value of τ.

4.4 Examples of Topic Evolutions

We have already given an example of the results of topic evolution analysis in Figure 1, where we can find several interesting transitions. For example, the topic seemingly corresponding to machine translation (represented by *translat, languag, system, english, machin, cross, automat*, etc) and the topic seemingly corresponding to speech recognition (represented by *speech, recognit, system, word, speaker, model, recogn*, etc) point to each other with a thick arrow. That is, we can reveal that machine translation has a close bi-directional relationship with speech recognition based on the result given by TERESA.

More examples are presented in Figures 3 and 4. Figure 3 contains the examples of topic evolutions extracted from the CORA data set. On the top panel of Figure 3, the word *network* appears in several boxes. However, the boxes representing the topics related to network protocols (i.e., the boxes containing the words *network, protocol, internet, servic, multicast*, etc) and the boxes representing the topics related to neural networks (i.e., the boxes containing the words *network, neural, learn, train*, etc) are never connected. This conforms to our intuition. On the bottom panel of Figure 3, the topic seemingly related to Bayesian statistical analysis (represented by *model, estim, data, distribut, statist, bayesian, paramet*, etc), has incoming arcs both from the topic seemingly related to MCMC inferences (represented by *chain, markov, carlo, mont, converg*, etc), and from the topic seemingly related to bioinformatics (represented by *sequence, protein, dna, molecular, align*, etc). We can guess that some papers on Bayesian methods were written based on the preceding works related to MCMC inferences and also on the works related to bioinformatics.

We present the examples of topic evolutions extracted from the HEP-PH data set in Figure 4. On the left panel, the topic at the top left corner has the word *dag* as the most frequent word, which is the mathematical symbol "†" often used in expressing the unitarity of matrices. This topic and the topics that have incoming arcs from it are seemingly related to penguin diagram (represented by *penguin, decai, ckm, asymmetri*, etc). On the right panel, we can find a similar topic (represented by *penguin, decai, asymmetri*, etc) at the left center position. This topic has a thin arc coming from the topic represented by *decai, quark, hadron, semilepton*, etc, and this topic has bidirectional relationships with the topics that are also represented by *semilepton, hadron*, etc. However, the word *semilepton* does not appear on the left panel of Figure 4 at the position close to the topic related to penguin diagram. Therefore, we can guess that the papers discussing penguin diagram cite the papers on semileptonic decay only in a limited frequency, and thus that semileptonic decay is not a main topic of the papers cited by the papers on penguin diagram. In this manner, we can com-

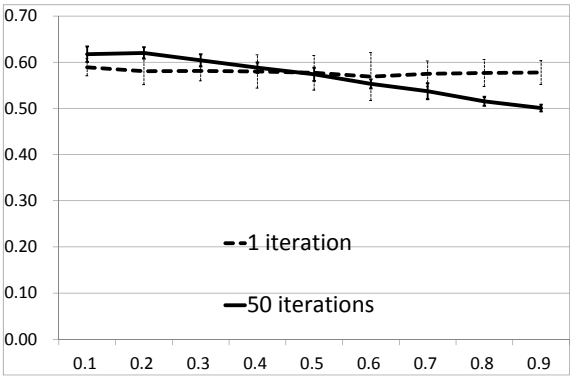

Figure 5: D+F scores for CORA ($K = 300$).

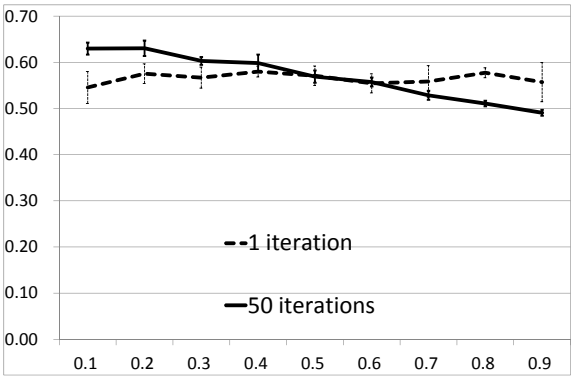

Figure 6: D+F scores for CORA ($K = 200$).

pare topic evolution networks given by different executions of VB on the same data set to obtain a more robust and comprehensive observation.

4.5 Usage

As in case of LDA, TERESA gives different inference results for different random initialization. Figure 2 presents three different topic transition networks extracted from the HEP-PH data set by TERESA under completely the same setting, i.e., $K = 300$ and $\tau = 0.2$. However, TERESA can provide a corpus-wide view on how topics interrelate. Therefore, users of our method can run the inference several times on the same data set under the same setting and inspect the obtained topic evolution networks to crop interesting portions from each result and then to compare those cropped portions. In contrast, when we use the methods that give a pile of local topic transition patterns, as are enumerated in Section 2, this type of usage is hardly realizable, because it is difficult, in the first place, to integrate local views for obtaining a unified view on topic evolutions.

5. CONCLUSION

This paper provides an LDA-like topic model for extracting a corpus-wide view on how latent topics interrelate in a given set of scientific articles. While many existing approaches provide an accumulation of a large number of local views, our approach, TERESA, can give a single global view.

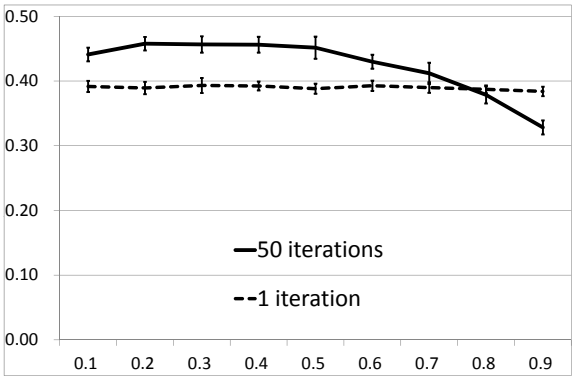

Figure 7: D+F scores for HEP-PH ($K = 300$).

Since topic modeling is primarily a method of dimensionality reduction, TERESA gives a "projection" of complicated semantical relationships among documents as a single transition matrix. Practically, we may run our method repeatedly on the same data set and obtain multiple "projections" to find interesting topic evolutions, as we have done in Section 4.4 with respect to the figures presented in this paper.

The only weak point of our method is the approximation introduced to obtain a non-gradient-based inference (cf. Eq. (14)). Therefore, the most important future work is to devise an inference with less approximation.

6. REFERENCES

[1] A. Agovic and A. Banerjee. Gaussian process topic models. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 10–19, Corvallis, Oregon, 2010. AUAI Press.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34, Arlington, Virginia, United States, 2009. AUAI Press.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

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

[5] J. Chang and D. M. Blei. Relational topic models for document networks. *Journal of Machine Learning Research - Proceedings Track*, 5:81–88, 2009.

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

[7] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 957–966, New York, NY, USA, 2009. ACM.

Table 3: Specifications of Data Sets

data set	# docs	# words	# unique doc-word pairs	# references	# word tokens for training (for test)
CORA	36,183	8,542	1,523,768	129,492	2,127,005 (235,566)
HEP-PH	34,546	86,964	21,399,454	416,146	90,796,243 (10,084,293)

[8] P. Hennig, D. H. Stern, R. Herbrich, and T. Graepel. Kernel topic models. *CoRR*, abs/1110.4713, 2011.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

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

[11] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 665–672, New York, NY, USA, 2009. ACM.

[12] T. P. Minka. Estimating a dirichlet distribution, 2000. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`.

[13] R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. Association for the Advancement of Artificial Intelligence, 2008.

[14] R. Nallapati, D. A. McFarland, and C. D. Manning. TopicFlow model: Unsupervised learning of topic-specific influences of hyperlinked documents. *Journal of Machine Learning Research - Proceedings Track*, 15:543–551, 2011.

[15] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[16] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35:773–782, 1980.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[18] L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 824–831, New York, NY, USA, 2008. ACM.

[19] Y. Sun, J. Han, J. Gao, and Y. Yu. iTopicModel: Information network-integrated topic modeling. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 493–502, Washington, DC, USA, 2009. IEEE Computer Society.

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[21] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[22] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *UAI*, pages 579–586, 2008.

[23] W. Y. Wang, E. Mayfield, S. Naidu, and J. Dittmar. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju Island, Korea, July 2012. ACL.

[24] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[25] F. Yan, N. Xu, and A. Qi. Parallel inference for latent Dirichlet allocation on graphics processing units. *Advances in Neural Information Processing Systems*, 22:2134–2142, 2009.

[26] K. Zhai, J. Boyd-Graber, N. Asadi, and M. Alkhouja. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *ACM International Conference on World Wide Web*, 2012.

APPENDIX

A. DERIVATION

We have the full joint distribution of TERESA as follows:

$$p(\boldsymbol{x}, \boldsymbol{z}, \theta, \phi, \boldsymbol{R}|\alpha, \beta, \gamma) = \prod_j p(\theta_j|\alpha) \cdot \prod_k p(\phi_k|\beta) \cdot \prod_k p(\boldsymbol{r}_k|\gamma) \cdot \prod_j \prod_i p(z_{ji}|\theta, \boldsymbol{R}) p(x_{ji}|\phi_{z_{ji}}) \quad (10)$$

where we omit free parameters τ and ω_{ji}. We apply a standard procedure of variational approximation as is provided in [4] by assuming that the variational posterior is fully factorized. The lower bound of the log evidence can be obtained by applying Jensen's inequality as follows:

$$\log p(\boldsymbol{x}|\alpha, \beta, \gamma) \geq \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}|\pi) q(\theta|\lambda) q(\phi|\mu) q(\boldsymbol{R}|\nu) \cdot \log \frac{p(\boldsymbol{x}|\boldsymbol{z}, \phi) p(\boldsymbol{z}|\theta, \boldsymbol{R}) p(\theta|\alpha) p(\phi|\beta) p(\boldsymbol{R}|\gamma)}{q(\boldsymbol{z}|\pi) q(\theta|\lambda) q(\phi|\mu) q(\boldsymbol{R}|\nu)} d\theta d\phi d\boldsymbol{R} \quad (11)$$

We denote the left hand side of Eq. (11) simply by $\mathcal{L}$. For convenience, we use the following notations: $\lambda_{j0} \equiv \sum_k \lambda_{jk}$, $\mu_{k0} \equiv \sum_w \mu_{kw}$, $\nu_{k0} \equiv \sum_{k'} \nu_{kk'}$, $\langle n_{kw} \rangle \equiv \sum_j \pi_{jwk} n_{jw}$ and

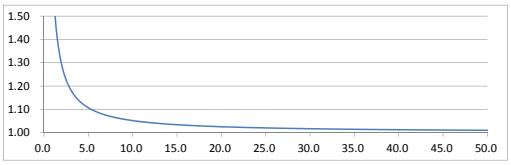

Figure 8: The function $f(x) = x\Psi'(x)$.

$\langle n_{jk}\rangle \equiv \sum_w \pi_{jwk} n_{jw}$. Then $\mathcal{L}$ can be expanded as follows:

$$\begin{aligned}
\mathcal{L} &= \sum_{k,w} \langle n_{kw}\rangle \big\{\Psi(\mu_{kw}) - \Psi(\mu_{k0})\big\} \\
&+ (1-\tau)\sum_{j,k} \langle n_{jk}\rangle \big\{\Psi(\lambda_{jk}) - \Psi(\lambda_{k0})\big\} \\
&+ \tau \sum_{j,k} \langle n_{jk}\rangle \sum_{j'} \omega_{jj'} \sum_{k'} \frac{\lambda_{j'k'}}{\lambda_{j'0}} \big\{\Psi(\nu_{k'k}) - \Psi(\nu_{k'0})\big\} \\
&+ J\log\Gamma(\alpha_0) - J\sum_k \log\Gamma(\alpha_k) - \sum_j \log\Gamma(\lambda_{j0}) \\
&+ \sum_{j,k} \log\Gamma(\lambda_{jk}) + \sum_{j,k} (\alpha_k - \lambda_{jk})\big\{\Psi(\lambda_{jk}) - \Psi(\lambda_{j0})\big\} \\
&+ K\log\Gamma(\beta_0) - K\sum_w \log\Gamma(\beta_w) - \sum_k \log\Gamma(\mu_{k0}) \\
&+ \sum_{k,w} \log\Gamma(\mu_{kw}) + \sum_{k,w} (\beta_w - \mu_{kw})\big\{\Psi(\mu_{kw}) - \Psi(\mu_{k0})\big\} \\
&+ K\log\Gamma(\gamma_0) - K\sum_{k'} \log\Gamma(\gamma_{k'}) - \sum_k \log\Gamma(\nu_{k0}) \\
&+ \sum_{k,k'} \log\Gamma(\nu_{kk'}) + \sum_{k,k'} (\gamma_{k'} - \nu_{kk'})\big\{\Psi(\nu_{kk'}) - \Psi(\nu_{k0})\big\} \\
&- \sum_{j,w} n_{jw} \sum_k \pi_{jwk} \log \pi_{jwk}. \qquad (12)
\end{aligned}$$

where the third term is obtained by an additional application of Jensen's inequality to the log of the linear combination of topic multinomial parameters in Eq. (1).

It is a complicated process to optimize $\mathcal{L}$ with respect to λ_{jk}. By differentiating $\mathcal{L}$ with respect to λ_{jk}, we obtain:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \lambda_{jk}} &= \big\{(1-\tau)\langle n_{jk}\rangle - (\lambda_{jk} - \alpha_k)\big\}\Psi'(\lambda_{jk}) \\
&- \big\{(1-\tau) n_j - (\lambda_{j0} - \alpha_0)\big\}\Psi'(\lambda_{j0}) \\
&+ \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'}\rangle \big\{\Psi(\nu_{kk'}) - \Psi(\nu_{k0})\big\} \\
&- \frac{\tau}{\lambda_{j0}^2} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'}\rangle \sum_{\tilde{k}} \lambda_{j\tilde{k}} \big\{\Psi(\nu_{\tilde{k}k'}) - \Psi(\nu_{\tilde{k}0})\big\} \qquad (13)
\end{aligned}$$

We introduce an approximation to Eq. (13) by assuming

$$\lambda_{jk}\Psi'(\lambda_{jk}) \approx \lambda_{j0}\Psi'(\lambda_{j0}) \qquad (14)$$

Figure 8 is a plot of the function $f(x) = x\Psi'(x)$. As this graph shows, $f(x)$ is almost equal to 1 for large values of x. Therefore, our approximation works for larger values of λ_{jk} and λ_{j0}. Consequently, we obtain the following derivative:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \lambda_{jk}} &\approx \big\{(1-\tau)\langle n_{jk}\rangle - (\lambda_{jk} - \alpha_k)\big\}\Psi'(\lambda_{jk}) \\
&- \big\{(1-\tau) n_j - (\lambda_{j0} - \alpha_0)\big\}\Psi'(\lambda_{j0}) \\
&+ \frac{\tau\Psi'(\lambda_{jk})}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'}\rangle \frac{\Psi(\nu_{kk'}) - \Psi(\nu_{k0})}{\Psi'(\lambda_{jk})} \\
&- \frac{\tau\Psi'(\lambda_{j0})}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'}\rangle \sum_{\tilde{k}} \frac{\Psi(\nu_{\tilde{k}k'}) - \Psi(\nu_{\tilde{k}0})}{\Psi'(\lambda_{j\tilde{k}})}. \qquad (15)
\end{aligned}$$

Eq. (15) tells that $\partial\mathcal{L}/\partial\lambda_{jk} = 0$ holds when

$$\begin{aligned}
\lambda_{jk} &\approx \alpha_k + (1-\tau)\langle n_{jk}\rangle \\
&+ \frac{\tau}{\lambda_{j0}\Psi'(\lambda_{jk})} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'}\rangle \big\{\Psi(\nu_{kk'}) - \Psi(\nu_{k0})\big\}. \qquad (16)
\end{aligned}$$

This equation can be solved by a binary search with respect to λ_{jk} as is discussed in Section 3.2.

It is an important future work to propose an inference avoiding the approximation we introduce here, because we apply this approximation as a last resort to achieve a non-gradient-based and thus efficient inference. Although a gradient-based inference was once implemented, the running time was unacceptably long, because K dimensional optimization problem should be solved for each of the J documents.

Further, we differentiate $\mathcal{L}$ with respect to $\nu_{kk'}$ and obtain:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \nu_{kk'}} &= \tau\Psi'(\nu_{kk'}) \sum_j \langle n_{jk'}\rangle \sum_{j'} \omega_{jj'} \frac{\lambda_{j'k}}{\lambda_{j'0}} \\
&- \tau\Psi'(\nu_{k0}) \sum_{k'} \sum_j \langle n_{jk'}\rangle \sum_{j'} \omega_{jj'} \frac{\lambda_{j'k}}{\lambda_{j'0}} \\
&+ (\gamma_{k'} - \nu_{kk'})\Psi'(\nu_{kk'}) - \Psi'(\nu_{k0}) \sum_{\tilde{k}} (\gamma_{\tilde{k}} - \nu_{k\tilde{k}}) \qquad (17)
\end{aligned}$$

Therefore, $\nu_{kk'}$ can be updated as:

$$\nu_{kk'} = \gamma_{k'} + \tau \sum_j \langle n_{jk'}\rangle \sum_{j'} \omega_{jj'} \frac{\lambda_{j'k}}{\lambda_{j'0}} \qquad (18)$$