The Contextual Focused Topic Model

Xu Chen
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
xu.chen@duke.edu

Mingyuan Zhou
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
mz1@ee.duke.edu

Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
lcarin@ee.duke.edu

ABSTRACT

A nonparametric Bayesian contextual focused topic model (cFTM) is proposed. The cFTM infers a sparse ("focused") set of topics for each document, while also leveraging contextual information about the author(s) and document venue. The hierarchical beta process, coupled with a Bernoulli process, is employed to infer the focused set of topics associated with each author and venue; the same construction is also employed to infer those topics associated with a given document that are unusual (termed "random effects"), relative to topics that are inferred as probable for the associated author(s) and venue. To leverage statistical strength and infer latent interrelationships between authors and venues, the Dirichlet process is utilized to cluster authors and venues. The cFTM automatically infers the number of topics needed to represent the corpus, the number of author and venue clusters, and the probabilistic importance of the author, venue and random-effect information on word assignment for a given document. Efficient MCMC inference is presented. Example results and interpretations are presented for two real datasets, demonstrating promising performance, with comparison to other state-of-the-art methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Clustering; H.2.8 [**Information Systems Applications**]: Database Applications-Data mining

General Terms

Algorithm, Experimentation

Keywords

Topic modeling, Bayesian nonparametric, clustering, Dirichlet process, hierarchical beta process

1. INTRODUCTION

With the popularization of Web applications and other digital media, frequently one is interested in analyzing a large corpus [14, 1, 11, 13, 26], and it is desirable to place the analysis of such data within the context of other readily available associated information [22, 19, 6, 18, 5]. For example, with a large corpus, many documents may be written by the same authors or groups of authors, and it is desirable to account for this information when analyzing the documents [22]. It is likely that a given author may concentrate on a subset of topics, and utilization of this information may help infer the topics associated with each of the documents in the corpus. Further, each document is typically published in a venue (*e.g.*, magazine, newspaper, website, conference proceedings, etc.), and the network of venue information also carries significant information [6] (the inference of topics associated with any single document is influenced by other documents published at the same or similar venues). It is therefore of interest to learn interrelationships between venues, and between the authors, to allow an appropriate sharing of information from multiple documents.

In the specific examples considered in this paper, the corpus of documents consists of technical papers and proposals, and we leverage information from the associated author names and publication venues. We propose a novel nonparametric Bayesian approach that extends the recently developed focused topic model (FTM) [28]. We refer to the proposed model as a *contextual* FTM, or cFTM, as it is capable of accounting for an arbitrary set of relational contextual information; while here we focus on two forms of context, author names and publication venues, the basic approach may be extended to arbitrary types and numbers of context. The model is nonparametric in the sense that it infers the number of topics characteristic of the corpus automatically, and it also infers a clustering of the authors and venues (such that documents from similar authors and venues influence the topics associated with any given document). In addition to increasing the power of the model by appropriate sharing of data, the clustering of venues and authors is of interest in its own right, for inferring relational information. The Dirichlet process [9, 10, 21, 25] is employed to perform the clustering nonparametrically (the number of clusters required of the data is inferred, and additional clusters may be added as new data warrants).

The proposed cFTM has three principal advantages compared to related models: (1) It automatically infers the number of topics by combining properties from the Dirichlet process [25] and hierarchical beta process [27], allowing an unbounded number of topics for the entire corpus, while

KDD'12, August 12–16, 2012, Beijing, China.

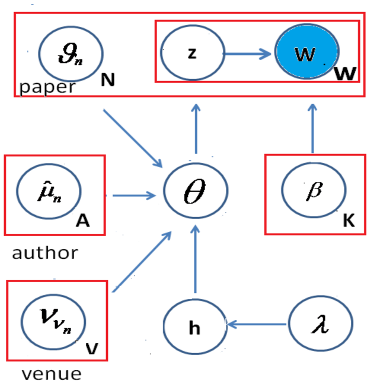


Figure 1: The graphical model for the proposed contextual focused topic model (cFTM), where A is the total set of authors, V is the total set of venues. The shaded variable is the observable variable and the hyper-parameters are omitted.

inferring a focused (sparse) set of topics for each individual document. (2) The cFTM nonparametrically clusters the authors and venues, thereby increasing statistical strength while also inferring useful relational information. (3) Instead of pre-specifying the importance of author/venue information (as was done in [6]), the cFTM automatically infers the document-dependent, probabilistic importance of the author/venue information on word assignment.

2. CONTEXTUAL FOCUSED TOPIC MODEL

Consider a corpus of N documents with W unique words in the vocabulary, A unique authors, and V unique venues. The corpus is represented as $\{\boldsymbol{d}_n, \boldsymbol{a}_n, v_n\}_{n=1,\ldots,N}$, where $\boldsymbol{d}_n$ denotes the set of words in document n (the order of the words is exchangeable, *i.e.*, a bag-of-words model), $\boldsymbol{a}_n$ consists of a subset of authors from the set $\{1, \cdots, A\}$, and $v_n \in \{1, \cdots, V\}$ is the venue index. Each document appears in one venue and has one or multiple authors.

2.1 Word Assignment

In a topic model, document n is typically characterized by a distribution over topics, $\boldsymbol{\theta}_n$, and topic k by a distribution over words, $\boldsymbol{\beta}_k$. Let $w_{ni} \in \{1, \ldots, V\}$ denote word i in document n. Each w_{ni} is assumed constituted by first drawing a single topic index $z_{ni} \in \{1, \ldots, K\}$ from a multinomial distribution with probability vector $\boldsymbol{\theta}_n$; then w_{ni} is drawn from a multinomial distribution with probability vector $\boldsymbol{\beta}_{z_{ni}}$. Different topic models are distinguished by how $\{\boldsymbol{\theta}_n\}_{n=1,N}$ are constituted. For example, in latent Dirichlet allocation (LDA) [1], an early topic model, the words are generated as

$$w_{ni} \sim \text{Discrete}(\boldsymbol{\beta}_{z_{ni}}),\ z_{ni} \sim \text{Discrete}(\boldsymbol{\theta}_n) \quad (1)$$

$$\boldsymbol{\beta}_k \sim \text{Dir}(\eta, \cdots, \eta),\ \boldsymbol{\theta}_n \sim \text{Dir}(\alpha_\theta, \cdots, \alpha_\theta) \quad (2)$$

where Discrete$(\boldsymbol{\beta}_{z_{ni}})$ is a distribution over indices, from which a *single* index is drawn from a multinomial distribution, with probability vector $\boldsymbol{\beta}_{z_{ni}}$ defining the probabilities of selecting the indices; η and α_θ are hyperparameters. In this simple model the $\boldsymbol{\theta}_n$ are drawn *independently*, not leveraging information from across the corpus, and not leveraging contextual information. We first discuss how this is generalized such that the distribution over topics for document n accounts for author and venue information, while also allowing a document-dependent "random effect" that accounts for idiosyncratic characteristics of a given document (*e.g.*, an outlier among papers by particular authors, published in a particular venue).

Let $\boldsymbol{\mu}_j$ represent a distribution over topics for author $j \in \{1, \ldots, A\}$, and let $\boldsymbol{\mu}_v$ represent a distribution over topics for venue $v \in \{1, \ldots, V\}$. Finally, let $\boldsymbol{\vartheta}_n$ represent a distribution over topics for document n, with this term constituting a "random effect" term, meaning that it captures unique aspects of document n that are not captured by the distribution over topics associated with the corresponding author(s) and venue. Recalling that $\boldsymbol{a}_n$ denotes the authors associated with document n, the averaged author distribution over topics for document n is

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{|\boldsymbol{a}_n|} \sum_{j \in \boldsymbol{a}_n} \boldsymbol{\mu}_j \quad (3)$$

where $|\boldsymbol{a}_n|$ is the number of authors in document n. In (3) each of the $|\boldsymbol{a}_n|$ authors is assumed to contribute topics equally (uniformly) to document n; the model may be extended to infer the relative importance of the individual authors to multi-author documents (this may be done in a manner analogous to that discussed below, where we infer the importance of the author(s), venue and random effects to topic generation).

In the proposed model, word w_{ni} is drawn either from $\boldsymbol{\vartheta}_n$, $\hat{\boldsymbol{\mu}}_n$ or $\boldsymbol{\nu}_{v_n}$, where v_n is the venue of document n. The probability of selecting from these is respectively λ_{n1}, λ_{n2} and λ_{n3}, with $\sum_{j=1}^{3} \lambda_{nj} = 1$ and $\lambda_{nj} \geq 0$. Assuming $\{\boldsymbol{\vartheta}_n\}$, $\{\boldsymbol{\mu}_j\}$ and $\{\boldsymbol{\nu}_v\}$ are given/specified, the generative model is

$$w_{ni} \sim \text{Discrete}(\boldsymbol{\beta}_{z_{ni}}), z_{ni} \sim \text{Discrete}(\boldsymbol{\theta}_{ni}) \quad (4)$$

$$\boldsymbol{\beta}_k \sim \text{Dir}(\eta, \cdots, \eta) \quad (5)$$

$$\boldsymbol{\theta}_{ni} = h_{ni1}\boldsymbol{\vartheta}_n + h_{ni2}\hat{\boldsymbol{\mu}}_n + h_{ni3}\boldsymbol{\nu}_{v_n} \quad (6)$$

$$(h_{ni1}, h_{ni2}, h_{ni3}) \sim \text{Mult}(\lambda_{n1}, \lambda_{n2}, \lambda_{n3}) \quad (7)$$

$$(\lambda_{n1}, \lambda_{n2}, \lambda_{n3}) \sim \text{Dir}(\alpha, \alpha, \alpha) \quad (8)$$

where $(h_{ni1}, h_{ni2}, h_{ni3}) \in \{(1,0,0), (0,1,0), (0,0,1)\}$ is a three dimensional binary vector indicating which of the three terms the word w_{ni} belongs to.

The construction in (4)-(8) will prove convenient for inference; however, to connect it to more-conventional models like (1), (4)-(8) is equivalent to specifying $\boldsymbol{\theta}_n = \lambda_{n1}\boldsymbol{\vartheta}_n + \lambda_{n2}\hat{\boldsymbol{\mu}}_n + \lambda_{n3}\boldsymbol{\nu}_{v_n}$, where λ_{n1} represents the probability that w_{ni} is drawn from a topic unanticipated from the authors and venue, λ_{n2} represents the probability that the associated topic is characteristic of the author(s), and λ_{n3} quantifies the probability that the topic is characteristic of the venue.

This decomposition suggests developing *focused* (sparse) topic distributions for $\{\boldsymbol{\vartheta}_n\}$, $\{\boldsymbol{\mu}_j\}$ and $\{\boldsymbol{\nu}_v\}$, such that each of these distributions over topics focuses on the characteristics of authors and venues, while also identifying random-effect topics not anticipated by either; the construction of focused distributions for $\{\boldsymbol{\vartheta}_n\}$, $\{\boldsymbol{\mu}_j\}$ and $\{\boldsymbol{\nu}_v\}$ is detailed in Section 2.3. We first discuss how we may cluster the authors and venues, with the clustering manifested in terms of the probabilities over topics, $\{\boldsymbol{\mu}_j\}$ and $\{\boldsymbol{\nu}_v\}$; within each author/venue cluster, topic usage is similar.

2.2 Author and Venue Clustering

It is expected that authors working in the same area tend to write documents addressing similar topics. It is also anticipated that publication venues that are closely related to one other (*e.g.*, KDD and SDM), tend to publish documents on similar topics. So motivated, we seek to cluster authors and venues based on their usage of topics. This clustering is performed nonparametrically through use of the Dirichlet process (DP) [10].

The probability vectors $\{\boldsymbol{\mu}_j\}$ are drawn from a DP as $\boldsymbol{\mu}_j \sim \hat{G}_\mu$ with $\hat{G}_\mu \sim \mathrm{DP}(\lambda_\mu, G_\mu)$, and a stick-breaking construction is employed [21]. Hence, $\boldsymbol{\mu}_j$ is drawn

$$\boldsymbol{\mu}_j \sim \sum_{m=1}^{\infty} c_m \delta_{\boldsymbol{\mu}^*_m} \tag{9}$$

$$c_m = c'_m \prod_{l<m}(1-c'_l),\ c'_l \sim \mathrm{Beta}(1, \lambda_\mu) \tag{10}$$

$$\lambda_\mu \sim \mathrm{Gamma}(g, h), \tag{11}$$

where each $\boldsymbol{\mu}^*_m$ is drawn i.i.d. from the "base" probability measure G_μ, and $\delta_{\boldsymbol{\mu}^*_m}$ is a unit point measure concentrated at $\boldsymbol{\mu}^*_m$. The form of G_μ is discussed in Section 2.3. Letting $\boldsymbol{c}$ represent the vector of probabilities $\boldsymbol{c} = (c_1, \dots)^T$, we denote the above process as $\boldsymbol{c} \sim \mathrm{Stick}(\lambda_\mu)$. In practice, the number of "sticks" is truncated as $\boldsymbol{\mu}_j \sim \sum_{m=1}^{M} c_m \delta_{\boldsymbol{\mu}^*_m}$, with $c'_M = 1$. We similarly draw the $\{\boldsymbol{\nu}_v\}$ as $\boldsymbol{\nu}_v \sim \hat{G}_\nu$ with $\hat{G}_\nu \sim \mathrm{DP}(\lambda_\nu, G_\nu)$.

In (9), c_m represents the probability that a given $\boldsymbol{\mu}_j$ is associated with cluster m; cluster m has a corresponding probability vector over topics defined by $\boldsymbol{\mu}^*_m$. Therefore, the number of components in the vector $\boldsymbol{c}$ with significant weight plays an important role in defining the number of clusters the vectors $\{\boldsymbol{\mu}_j\}$ are associated with. In this context parameter λ_μ is important, with only one cluster manifested as $\lambda_\mu \to 0$ (in this case all probability vectors in the set $\{\boldsymbol{\mu}_j\}$ are the same), and when $\lambda_\mu \to \infty$ all of the components in $\boldsymbol{c}$ have infinitesimal weight (in this case all probability vectors in the set $\{\boldsymbol{\mu}_j\}$ are unique, with probability one). Since λ_μ and λ_ν are important parameters, each is inferred, with gamma priors placed on each.

What remains is to define G_μ and G_ν, as well as the distribution on $\{\boldsymbol{\vartheta}_n\}$. These probability distributions are defined in a hierarchical manner, through a generalization of the focused topic model [28].

2.3 Hierarchical Beta Process

A direct extension of the LDA framework in (1) is to let the document-dependent random effects be drawn $\boldsymbol{\vartheta}_n \sim \mathrm{Dir}(\alpha_\vartheta, \cdots, \alpha_\vartheta)$, and similarly to define $G_\mu = \mathrm{Dir}(\alpha_\mu, \cdots, \alpha_\mu)$ and $G_\nu = \mathrm{Dir}(\alpha_\nu, \cdots, \alpha_\nu)$. However, in this setting the number of topics needs to be set *a priori*. Further, by the construction in (6) it is desirable that the probability vector $\boldsymbol{\mu}_j$ "focus" on the topics typically associated with author j, with $\boldsymbol{\nu}_v$ focusing on the topics typically associated with venue v. The probability vector $\boldsymbol{\vartheta}_n$ may then focus on those topics associated with document n that are unique to that document, and not addressed by the probabilities over topics associated with the corresponding author(s) and venue. These objectives motivate extending the focused topic model [28] for the purposes of the proposed model. In [28, 32] a beta-Bernoulli construction was employed to infer focused topics. Here we extend this setting to a *hierarchical* beta process (HBP) setting, coupled with the Bernoulli process; the hierarchical construction is motivated by our use of relational information, which was not considered in [28, 32].

The hierarchical beta process (HBP) was first studied in [27] and applied to dictionary learning for image reconstruction [33]. Reviewing, a draw $B \sim \mathrm{BP}(c_0, B_0)$ is defined by a real constant $c_0 > 0$ and a probability measure B_0. The measure B may be expressed in the form

$$B = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\beta}_k} \tag{12}$$

where each $\boldsymbol{\beta}_k$ is drawn i.i.d. from B_0. If B_0 is a continuous (non-atomic) probability measure, then each $\pi_k \in [0, 1]$ is drawn i.i.d. from the "degenerate" beta distribution $c_0\pi^{-1}(1-\pi)^{c_0-1}$, which has a singularity at $\pi = 0$, encouraging that only a minority of the $\{\pi_k\}$ will have significant values, away from $\pi = 0$. For an atomic base measure $B_0 = \sum_k q_k \delta_{\omega_k}$, with $q_k \in [0, 1]$, the draw B is of the same form as in (12), but now $\pi_k \sim \mathrm{Beta}(c_0 q_k, c_0(1-q_k))$. A draw B from a beta process is often linked with a Bernoulli process, where $X \sim \mathrm{Bernoulli}(B)$ is of the form $X = \sum_{k=1}^{\infty} b_k \delta_{\omega_k}$, with $b_k \in \{0, 1\}$ and $b_k \sim \mathrm{Bernoulli}(\pi_k)$.

While the form of B_0 is general, here $B_0 = \mathrm{Dir}(\eta, \dots, \eta)$, and therefore each draw $\boldsymbol{\beta}_k \sim B_0$ corresponds to a topic, reflected in a distribution over the W words in the vocabulary; this is the same $\boldsymbol{\beta}_k$ as considered in the above discussion (*e.g.*, in (1)), but now the π_k in $B = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\beta}_k}$ reflects (through the Bernoulli process) the probability of whether topic k is utilized. The proposed HBP construction is defined by drawing $B \sim \mathrm{BP}(c_0, B_0)$, and

$$B^{(\vartheta)} \sim \mathrm{BP}(c_1, B), \quad B^{(\mu)} \sim \mathrm{BP}(c_1, B), \quad B^{(\nu)} \sim \mathrm{BP}(c_1, B)$$

The measure $B^{(\mu)} = \sum_{k=1}^{\infty} \pi^{\mu}_k \delta_{\boldsymbol{\beta}_k}$ defines the probability π^{μ}_k that topic k is utilized by the authors, while similarly $B^{(\nu)} = \sum_{k=1}^{\infty} \pi^{\nu}_k \delta_{\boldsymbol{\beta}_k}$ defines the probability π^{ν}_k that topic k is utilized across the venues, and $B^{(\theta)} = \sum_{k=1}^{\infty} \pi^{\theta}_k \delta_{\boldsymbol{\beta}_k}$ defines the probability π^{θ}_k that topic k is utilized across the document-dependent random effects. Since B is shared across the draws for $B^{(\vartheta)}$, $B^{(\mu)}$ and $B^{(\nu)}$, the topics reflected by $\{\boldsymbol{\beta}_k\}$ are also shared, but each has a unique set of probabilities $\{\pi^{\theta}_k\}$, $\{\pi^{\mu}_k\}$ and $\{\pi^{\nu}_k\}$ that the topics are utilized.

Note that for convenient implementation, one often employs a finite approximation for BP draws [32], with a truncation to $K_{\max}$ topics. In this setting we have

$$\pi_k \sim \mathrm{Beta}(c_0\epsilon, c_0(1-\epsilon)), \pi^{\vartheta}_k \sim \mathrm{Beta}(c_1\pi_k, c_1(1-\pi_k))$$
$$\pi^{\mu}_k \sim \mathrm{Beta}(c_1\pi_k, c_1(1-\pi_k)), \pi^{\nu}_k \sim \mathrm{Beta}(c_1\pi_k, c_1(1-\pi_k))$$

where $\epsilon = 1/K_{\max}$.

For document n, cluster m for the authors, and cluster m' for the venues, we employ the Bernoulli process (BeP), yielding $X^{(\vartheta)}_n \sim \mathrm{BeP}(B^{(\vartheta)})$, $X^{(\mu)}_m \sim \mathrm{BeP}(B^{(\mu)})$ and $X^{(\nu)}_{m'} \sim \mathrm{BeP}(B^{(\nu)})$, which can be expressed as

$$X^{(\vartheta)}_n = \sum_{k=1}^{K_{\max}} b^{(\vartheta)}_{nk} \delta_{\boldsymbol{\beta}_k},\ b^{(\vartheta)}_{nk} \sim \mathrm{Bernoulli}(\pi^{(\vartheta)}_k) \tag{13}$$

$$X^{(\mu)}_m = \sum_{k=1}^{K_{\max}} b^{(\mu)}_{mk} \delta_{\boldsymbol{\beta}_k},\ b^{(\mu)}_{mk} \sim \mathrm{Bernoulli}(\pi^{(\mu)}_k) \tag{14}$$

$$X^{(\nu)}_{m'} = \sum_{k=1}^{K_{\max}} b^{(\nu)}_{m'k} \delta_{\boldsymbol{\beta}_k},\ b^{(\nu)}_{m'k} \sim \mathrm{Bernoulli}(\pi^{(\nu)}_k). \tag{15}$$

The binary vector $\boldsymbol{b}_n^{(\vartheta)} = (b_{n1}^{(\vartheta)}, \ldots, b_{nK_{\max}}^{(\vartheta)})^T$ defines which of the $K_{\max}$ topics are "on" (those for which $b_{nk} = 1$). Since the beta-Bernoulli process leads to a sparse set of non-zero b_{nk}^{ϑ}, the vector $\boldsymbol{b}_n^{(\vartheta)}$ defines which topics the random effects associated with document n focuses on. The binary vectors $\boldsymbol{b}_m^{(\mu)}$ and $\boldsymbol{b}_{m'}^{(\nu)}$ similarly define which topics are focused on by cluster m of the authors and cluster m' of the venues respectively. Completing the model, we have

$$\boldsymbol{\vartheta}_n \sim \mathrm{Dir}(\boldsymbol{b}_n^{(\vartheta)} \circ \boldsymbol{r}^{(\vartheta)}),\ r_k^{(\vartheta)} \sim \mathrm{Gamma}(\gamma_1, 1) \tag{16}$$

$$\boldsymbol{\mu}_m^* \sim \mathrm{Dir}(\boldsymbol{b}_m^{(\mu)} \circ \boldsymbol{r}^{(\mu)}),\ r_k^{(\mu)} \sim \mathrm{Gamma}(\gamma_2, 1) \tag{17}$$

$$\boldsymbol{\nu}_m^* \sim \mathrm{Dir}(\boldsymbol{b}_m^{(\nu)} \circ \boldsymbol{r}^{(\nu)}),\ r_k^{(\nu)} \sim \mathrm{Gamma}(\gamma_3, 1) \tag{18}$$

where γ_1, γ_2 and γ_3 are hyperparameters and $\circ$ represents the Hadamard (element-wise) product. Note, for example, that $\boldsymbol{\vartheta}_n$ will only have non-zero components for the topic indices k for which the components of $\boldsymbol{b}_n^{(\vartheta)}$ are non-zero, yielding the desired focused set of topics.

Relating the above discussion to the DP base measures G_μ and G_ν discussed in Section 2.2, G_μ is defined by the hierarchical combination of the HBP, Bernoulli process, and the construction in (17), with G_ν defined similarly. Further, the $\boldsymbol{\vartheta}_n$ in (6) is defined by the hierarchy of HBP, the Bernoulli process, and the construction in (16).

Denote $\Omega_{nk}^{(\vartheta)}$ as the number of words associated with the kth topic and assigned to the nth document's random effect term $\boldsymbol{\vartheta}_n$, denote $\Omega_{mk}^{(\mu)}$ as the number of words associated with the kth topic and assigned to the mth author cluster $\boldsymbol{\mu}_m^*$, and denote $\Omega_{mk}^{(\nu)}$ as the number of words associated with the kth topic and assigned to the mth venue cluster $\boldsymbol{\nu}_m^*$. Following [28] and the derivations in [32],

$$\Omega_{nk}^{(\vartheta)} \sim \mathrm{NB}(b_{nk}^{(\vartheta)} r_k^{(\vartheta)}, 0.5) \tag{19}$$

$$\Omega_{mk}^{(\mu)} \sim \mathrm{NB}(b_{mk}^{(\mu)} r_k^{(\mu)}, 0.5) \tag{20}$$

$$\Omega_{mk}^{(\nu)} \sim \mathrm{NB}(b_{mk}^{(\nu)} r_k^{(\nu)}, 0.5) \tag{21}$$

where NB denotes the negative binomial distribution. The above equations together with (16)-(18) can be used to infer the posterior distributions of $r_k^{(\vartheta)}$, $r_k^{(\mu)}$ and $r_k^{(\nu)}$.

3. MCMC INFERENCE

We utilize MCMC inference to sample latent variables from their conditional posterior distributions. The inputs for cFTM include the text information $\{\boldsymbol{d}_n\}_{n=1,N}$, author information $\{\boldsymbol{a}_n\}_{n=1,N}$, venue information $\{v_n\}_{n=1,N}$, the hyperparameters $\eta, \alpha, \gamma_1, \gamma_2, \gamma_3, g, h$ and the number of sticks M (the setting of these parameters is discussed when presenting results).

I). For $k = 1, 2, \ldots, K$,

(a) Sample π_k, $\pi_k^{(\vartheta)}, \pi_k^{(\mu)}, \pi_k^{(\nu)}$. We sample π_k with slice sampling utilized in [33]. The rejection sampling proposed in [27] may also be used to sample π_k.

$$p(\pi_k^{(\vartheta)}|-) \sim \mathrm{Beta}(c_1\pi_k + \sum_{n=1}^{N} b_{nk}^{(\vartheta)}, N + c_1(1-\pi_k) - \sum_{n=1}^{N} b_{nk}^{(\vartheta)})$$

$$p(\pi_k^{(\mu)}|-) \sim \mathrm{Beta}(c_1\pi_k + \sum_{m=1}^{M} b_{mk}^{(\mu)}, M + c_1(1-\pi_k) - \sum_{m=1}^{M} b_{mk}^{(\mu)})$$

$$p(\pi_k^{(\nu)}|-) \sim \mathrm{Beta}(c_1\pi_k + \sum_{m=1}^{V} b_{mk}^{(\nu)}, V + c_1(1-\pi_k) - \sum_{m=1}^{V} b_{mk}^{(\nu)}).$$

(b) Sample $r_k^{(\vartheta)}$, $r_k^{(\mu)}$, $r_k^{(\nu)}$,γ_1, γ_2, γ_3.

$$p(r_k^{(\vartheta)}|-) \propto \mathrm{Gamma}(r_k^{(\vartheta)}; \gamma_1, 1) \prod_{n: b_{nk}^{(\vartheta)}=1} \mathrm{NB}\left(\Omega_{nk}^{(\vartheta)}; r_k^{(\vartheta)}, 0.5\right)$$

$$p(r_k^{(\mu)}|-) \propto \mathrm{Gamma}(r_k^{(\mu)}; \gamma_2, 1) \prod_{m: b_{mk}^{(\mu)}=1} \mathrm{NB}\left(\Omega_{mk}^{(\mu)}; r_k^{(\mu)}, 0.5\right)$$

$$p(r_k^{(\nu)}|-) \propto \mathrm{Gamma}(r_k^{(\nu)}; \gamma_3, 1) \prod_{m: b_{mk}^{(\nu)}=1} \mathrm{NB}\left(\Omega_{mk}^{(\nu)}; r_k^{(\nu)}, 0.5\right).$$

The above equations are log differentiable with respect to $r_k^{(\vartheta)}$, $r_k^{(\mu)}$, $r_k^{(\nu)}$, γ_1, γ_2 and γ_3. Therefore, the Hybrid Monte Carlo [17, 28] is utilized to sample them from their conditional posteriors. We may also use the Metropolis-Hastings algorithm to sample these values [32].

II). For $m = 1, \ldots, M$, $n = 1, \ldots, N$, sample the binary vector for the documents $(b_{n1}^{(\vartheta)}, \ldots, b_{nK_{\max}}^{(\vartheta)})$, the author $(b_{m1}^{(\mu)}, \ldots, b_{mK_{\max}}^{(\mu)})$ and the venue $(b_{m1}^{(\nu)}, \ldots, b_{mK_{\max}}^{(\nu)})$. When $\Omega_{nk}^{(\vartheta)} > 0$, we have $b_{nk}^{(\vartheta)} \equiv 1$. When $\Omega_{nk}^{(\vartheta)} = 0$, we have

$$p(b_{nk}^{(\vartheta)} = 1|-) \propto \pi_k^{(\vartheta)} \mathrm{NB}(0; r_k^{(\vartheta)}, 0.5) = \frac{\pi_k^{(\vartheta)}}{2^{r_k^{(\vartheta)}}} \tag{22}$$

$$p(b_{nk}^{(\vartheta)} = 0|-) \propto (1 - \pi_k^{(\vartheta)}). \tag{23}$$

Similar formulations can be derived for b_{mk}^{μ} and b_{mk}^{ν}.

III). Sample the variables z: when $h_{ni1} = 1$

$$p(z_{ni} = k|\gamma_{n1}, \boldsymbol{\vartheta}_{nk}, w, \boldsymbol{z}_{-ni}, \gamma_1, \pi_k^{(\vartheta)}, r_k^{(\vartheta)}) \propto p(w_{ni}|\boldsymbol{\beta}_k)$$
$$\int d\boldsymbol{\vartheta}_{nk} p(z_{ni}|\boldsymbol{\vartheta}_{nk}) p(\boldsymbol{\vartheta}_{nk}|\boldsymbol{z}_{-ni}, \gamma_1, \pi_k^{(\vartheta)}, r_k^{(\vartheta)})\ , \tag{24}$$

where $\boldsymbol{z}_{-ni}$ defines all $\boldsymbol{z}$ but z_{ni} conditioned on the sparse binary vector $\boldsymbol{b}_n^{(\vartheta)}$ and the gamma random variables $\boldsymbol{r}^{(\vartheta)}$, the topic proportion vector $\boldsymbol{\vartheta}_n$ is distributed according to a Dirichlet distribution. The sparse vector $\boldsymbol{b}_n^{(\vartheta)}$ determines the subset of topics over which the Dirichlet distribution is defined and $\boldsymbol{r}^{(\vartheta)}$ determines the values of the Dirichlet parameters at these points.

$$p(\boldsymbol{\vartheta}_n|\boldsymbol{z}_{-ni}, \pi_k^{(\vartheta)}, \gamma_1, r_k^{(\vartheta)}) \propto \int dr_k^{(\vartheta)}$$
$$\sum_{b_n^{(\vartheta)}} \mathrm{Dir}(\boldsymbol{\vartheta}_n|(Q_{-i}^n + r^{(\vartheta)}) \circ \mathbf{b}_n^{(\vartheta)}) p(\boldsymbol{b}_n^{(\vartheta)}, r_k^{(\vartheta)}|\gamma_1, \pi_k^{(\vartheta)})\ ,$$

where Q_{-i}^n is the topic assignment statistic excluding word w_{ni}. When $h_{ni2} = 1$ or $h_{ni3} = 1$, similar derivations are constituted.

IV). Sample λ_{n1}, λ_{n2}, λ_{n3}:

$$p((\lambda_{n1}, \lambda_{n2}, \lambda_{n3})|-) \sim$$
$$\mathrm{Dir}\left(\alpha + \sum_{i=1}^{|\boldsymbol{d}_n|} h_{ni1}, \alpha + \sum_{i=1}^{|\boldsymbol{d}_n|} h_{ni2}, \alpha + \sum_{i=1}^{|\boldsymbol{d}_n|} h_{ni3}\right). \tag{25}$$

V). Sample $h_{ni1}, h_{ni2}, h_{ni3}$:

$$p(h_{ni1}|-) \propto \lambda_1 \mathrm{Discrete}(w_{ni}; \boldsymbol{\Phi}\boldsymbol{\vartheta}_n) \tag{26}$$

$$p(h_{ni2}|-) \propto \lambda_2 \mathrm{Discrete}(w_{ni}, \boldsymbol{\Phi}\hat{\boldsymbol{\mu}}_n) \tag{27}$$

$$p(h_{ni3}|-) \propto \lambda_3 \mathrm{Discrete}(w_{ni}; \boldsymbol{\Phi}\boldsymbol{\nu}_{v_n}) \tag{28}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K] \in \mathbb{R}^{W \times K}$.

VI). Sample the stick lengths c'_l:

$$p(c'_l|-) \quad \sim \quad \text{Beta}(1 + N_l, \lambda_\mu + \sum_{m=l+1}^{M} N_m) \tag{29}$$

where N_m is the number of authors assigned to the mth stick. The parameter λ_μ is inferred as in [20] and we similarly sample the stick lengths for venue clustering.

4. RELATED MODELS

A topic model with biased propagation (TMBP) [6] was recently proposed to discover latent semantic topics, while leveraging contextual information such as the authors and venue. However, in TMBP the number of topics and the importance weights of the author and venue information on word assignment have to be predefined, with cross-validation often necessary.

A FTM using the Indian buffet process (IBP) compound DP prior [28] is proposed as a nonparametric Bayesian topic model to automatically determine the number of topics. The FTM employs an IBP [12] to place binary weights on the topics used within a given document, and therefore only a subset of topics are used within a given document. This imposes that the model focuses on representing a document in terms of a concise set of topics, which should be contrasted with previous hierarchical Dirichlet process (HDP) [26] models for a document corpus, in which each topic is manifested in general with non-zero probability within a given document. The FTM decouples the across-document popularity and within-document prevalence of topic usages, leading to improved performance compared to the HDP [26]. However, the FTM does not utilize contextual information such as the authors and venue.

A key contribution of this paper concerns extending the FTM to the class of problems considered by TMBP, thereby developing a novel nonparametric model for a document corpus that infers a focused set of topics, while also leveraging author/venue contextual information. Our cFTM model can readily handle multiple types of context in a nonparametric Bayesian framework. We demonstrated in the experiments that such flexibility provides significantly better performance in the analysis of documents.

Concerning other related work, the author topic model (ATM) [22] constitutes a representation in which topic distributions are tied to authors. The proposed cFTM extends ATM by considering venue information in addition to author identity, by inferring clusters of authors and venues, and by doing so with focused topics (via use of HBP). Additionally, the proposed cFTM has a "random effects" term $\boldsymbol{\vartheta}_n$ that accounts for situations in which a given document has topic usage that is inconsistent with the expectations of a given author or venue, increasing model flexibility, and allowing detection of outlier documents. Other related work includes NetPLSA [18], Laplacian PLSI [3], locally-consistent topic model [4], and citation and social network analysis [19][24][5] have been proposed for combining topic modelling and network structure. Graph-based semi-supervised learning [34] [30] [31] and link analysis [2][29][15] have also been applied to data mining [7][8]. However, these models did not explore contextual information.

5. EXPERIMENTAL RESULTS

We evaluate the proposed cFTM on the Digital Bibliography and Library Project (DBLP) dataset[1]and the NSF Research Awards Abstracts (NSF) dataset[2], as considered in [6]. The DBLP dataset is a collection of bibliographic information on major computer science journals and proceedings. We use a DBLP subset[3] containing the titles and abstracts of $N = 28569$ documents, with $W = 11771$ words in the vocabulary, $A = 28702$ authors and $V = 20$ conferences. Each conference is labeled with one of the four research areas: data mining, information retrieval, database, and artificial intelligence; each document is given the same label as the conference it appears in, and each author is labeled with the research area where he publishes the most number of documents [23].

The NSF dataset is made up of $N = 129000$ abstracts describing NSF awards for research from 1990 to 2003. We consider a subset, containing $N = 16405$ documents and $A = 9989$ investigators. These documents belong to the largest 10 research programs, such as applied mathematics, economics and geophysics. There are in total 20,717 links between the documents and investigators and $W = 18674$ unique words. Note that there are no venue information for this dataset.

We use both the accuracy (AC) and normalized mutual information (NMI) described in [6] for performance evaluations. The AC is defined as $AC = \frac{\sum_{i=1}^{n} \delta(a_i, map(l_i))}{n}$, where n denotes the total number of objects to be labeled, $\delta(x, y)$ equals one if $x = y$ and equals zero if $x \neq y$. The mapping function $map(l_i)$ maps each class label l_i to the corresponding label from the dataset. The mutual information $\text{MI}(C, C')$ between the labeled cluster C and the learned cluster C' is defined as

$$\text{MI}(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)} \tag{30}$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from the corpus belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ denotes the joint probability that an arbitrarily selected document belongs to the clusters c_i as well as c'_j at the same time. The NMI is defined as $\text{MI}(C, C')/\text{MI}(C, C)$.

We compare the proposed cFTM with nine algorithms: nonnegative matrix factorization (NMF) [16], probabilistic latent semantic analysis (PLSA) [14], laplacian probabilistic semantic indexing (Lap-PLSI) [3], latent Dirichlet allocation (LDA) [1], author-topic model (ATM) [22], ranking-based clustering (NetClus) [23], topic modeling with biased propagation including the biased random walk framework (TMBP-RW) and biased regularization framework (TMBP-Regu) [6], and focused topic model (FTM) [28]. Note that in the special case that $\lambda_{n1} = 1$, $\lambda_{n2} = 0$ and $\lambda_{n3} = 0$, the cFTM reduces to the FTM in which the contextual information is not utilized.

The parameter settings for the algorithms that are compared to are described in [6]. In cFTM, we set the parameters as $\alpha = 1/3$, $g = h = 10^{-6}$, $\eta = 0.05$, $c_1 = 1$, $c_0 = 1$, $M = 20$ and $K_{\max} = 50$. The gamma shape parameters $\gamma_1, \gamma_2, \gamma_3$ are given the prior of Gamma$(5, 0.1)$ and they are sampled with the Hybrid Monte Carlo algorithm [28]. The

[1] http://www.informatik.uni-trier.de/∼ley/db/

[2] http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html

[3] http://www.cs.uiuc.edu/∼hbdeng/data/kdd2011.htm

model proved insensitive to these parameter settings, and many other related settings give similar results. For the FTM and cFTM, we consider 6000 MCMC iterations, with the first 3000 samples discarded as burn-in and the remaining ones collected.

5.1 Classification Performance Comparison

For the DPLP dataset, the cFTM infers 12 frequently used topics, 10 author clusters (discarding clusters with less than 5 authors), and 4 venue clusters, according to the MCMC sample with the maximum likelihood. With the topic proportion vector for each document learned by the cFTM, we use the K-means to cluster these vectors into 4 and 10 classes for the DBLP and NSF datasets, respectively, equal to the number of labels in these two datasets (this is done only to quantify performance relative to "truth," allowing quantitative comparisons to other models; this is not a necessary step in our actual model). Based on these classes, we calculate the AC and NMI with the provided class labels.

As shown in Table 1, the proposed cFTM achieves the best overall performance on the DBLP data, notably outperforming the state-of-the-art TMBP algorithm [6] in both the document and author classifications, and having comparable results in venue classification. A characteristic distinguishing the cFTM from the TMBP algorithm [6] is that the cFTM automatically learns the importance weights of the author and venue information on word assignment, in a document dependent manner, while the TMBP algorithm fixes these weights to be the same for all the documents and uses cross-validation to tune the values. Furthermore, in the cFTM, the number of topics is automatically inferred with the HBP prior and the author and venues are automatically clustered with the DP, while in the TMBP algorithm [6], the number of topics is tied to the number of class labels and the author and venue clusterings are not considered. In addition, the cFTM is able to decouple the across document popularity and within document prevalence of the usages of topics under the FTM framework, while the TMBP algorithm, building on the PLSA framework, does not have this property. These differences may explain the performance gain achieved by the cFTM.

We also report in Table 2 the document classification performance of different methods on the NSF data, where the cFTM infers 35 topics and 26 investigator clusters. Among these methods, the ATM, TMBP and cFTM have comparable performance, outperforming all the other methods. Since in the NSF data, there is no venue information, the improvement of cFTM method over other methods is less significant compared to the results in the DBLP dataset. Another reason is that the heterogeneous information network of NSF is much sparser than that of DBLP: there are only 1.26 links per document for the NSF, while there are 3.61 links per document for the DBLP (here a "link" is defined as the connection between a investigator and a document) [6].

5.2 Inferred Author/Venue Relationships

We examine in detail the topic modeling and author and venue clustering results of the cFTM, concentrating on the DBLP data. Shown in Table 4 are typical topics characterized by their most-probable words, where Topics 4, 10, 6 and 11 correspond well to the research areas of database, data mining, information retrieval and artificial intelligence, respectively. For example, the top three key words of Topic 4 (database) are "data", "query" and "database" and the top three key words of Topic 11 (artificial intelligence) are "learning", "methods" and "knowledge". It is interesting to notice that the top key words of Topic 7, such as "mining", "learning" and "retrieval", appear frequently in both the data mining and information retrieval research areas. We also notice several topics frequently used across the documents are devoted to general terminology commonly found in all the four research areas, e.g., the top key words of Topic 5 are "method", "developed" and "applied". Compared to the top words from the topics extracted from TMBP [6], the cFTM method has the advantage that it absorbs the common words across different areas into several globally popular topics, while other topics focus on specific aspects of the corpus, which are only used by a subset of the documents.

Shown in Figure 2 are the author clustering results on the DBLP dataset. Based on the MCMC sample with the maximum likelihood, the cFTM infers 10 author clusters

Table 2: Classification performance comparison of different algorithms on the NSF dataset. Except for the results of the FTM and cFTM, all the other results were reported in [6].

Metric	AC (Paper)	NMI (Paper)
NMF	45.97	40.02 [6]
PLSA	63.00	64.48 [6]
LapPLSI	63.65	64.58 [6]
LDA	65.06	63.36 [6]
ATM	**65.69**	69.58 [6]
NetClus	63.51	66.11 [6]
TMBP-RW	64.84	68.74 [6]
TMBP-Regu	65.15	69.83 [6]
FTM	63.75	64.60
cFTM	65.49± 0.57	**70.07 ±0.26**

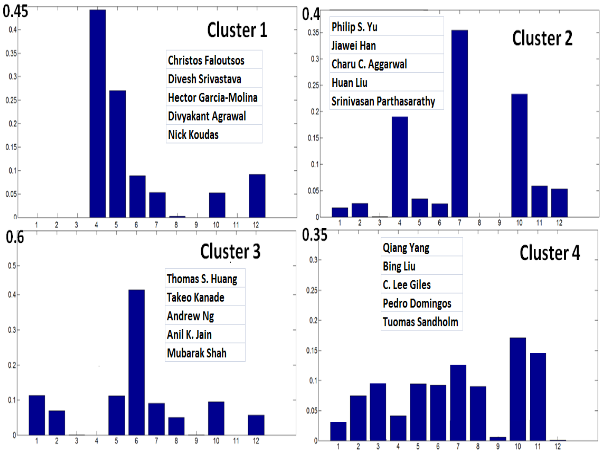


Figure 2: The topic usage probabilities of author clusters inferred by the cFTM on the DBLP dataset. Ten author clusters are inferred. We also show in each cluster the names of the authors who have the most number of publications. The horizontal axis represents topic indices (see Table 3), and the vertical axis reflects the probability of each topic appearing in a author cluster.

Table 1: Classification performance comparison of different algorithms on the DBLP dataset. Except for the results of the FTM and cFTM, all the other results were reported in [6].

Object	Paper	Paper	Author	Author	Venue	Venue	Average	Average
Metric (%)	AC	NMI	AC	NMI	AC	NMI	AC	NMI
NMF	44.55	22.92	-	-	-	-	44.55	22.92
PLSA	59.45	32.75	65.0	37.97	80.0	74.74	68.15	48.49
LapPLSI	61.35	33.93	-	-	-	-	60.70	33.37
LDA	47.00	20.48	-	-	-	-	47.00	20.48
ATM	77.00	52.21	74.13	40.67	-	-	75.57	46.44
NetClus	65.00	40.96	70.82	47.43	79.75	76.69	71.86	55.03
TMBP-RW	73.10	53.13	82.59	67.76	81.75	**77.53**	79.15	66.14
TMBP-Regu	79.15	59.16	89.81	74.25	**82.75**	76.56	83.90	69.99
FTM	69.37	43.51	-	-	-	-	69.37	43.51
cFTM	**82.73** **±0.65**	**62.91** **±0.51**	**92.51** **±0.71**	**76.20** **±0.39**	81.97 ±0.36	76.05 ±0.43	**85.73** **±0.57**	**71.72** **±0.45**

in total, with the first four largest clusters and their representative authors shown in Figure 2. The largest cluster (Cluster 1) contains about 4,700 authors and the smallest cluster contains about 30 authors. As shown in Figure 2, Cluster 1 consists of researchers in database, such as Christos Faloutsos, Divesh Srivestava and Hector Garcia-Molina. The authors in Cluster 3 are experts focusing on information retrieval, such as Thomas Huang, Takeo Kanade and Andrew Ng. Cluster 4 consists of authors who have publications in both data mining and artificial intelligence, such as Qiang Yang, B. Liu and C.Lee Giles. Authors in Cluster 4 typically have frequently published papers not only in artificial intelligence venues such as CIKM and WWW but also in data mining venues such as KDD and SDM.

The author topic usage probability vector for each cluster is also intuitive. Taking Cluster 1 for example, the topic usage probability vector has a large weight on Topic 4, which is characterized by words in database as shown in Table 4. The topic usage probability vector of Cluster 3 has a large weight in Topic 6, which is characterized by words in information retrieval. It is also not surprising that the usage probabilities of Topic 5 across clusters have a low variance, since Topic 5 contains common words frequently used in all documents regardless their research areas.

The venue clustering results on the DBLP data with the cFTM is investigated in Figure 4. We find that the venue clustering results are also quite intuitive. For example, the data mining conferences KDD, PKDD and PAKDD always stay in the same cluster in all the 3000 collected MCMC samples. Similarly, the database conferences ICDE, EDBT and PODS share the same cluster in all the collection samples. Note that in 83% of the collection samples AAAI and CVPR are in the same cluster, likely because AAAI consists of papers from both information retrieval and artificial intelligence.

6. CONCLUSIONS

Employing nonparametric Bayesian priors, including the Dirichlet process and hierarchical beta process, we propose a new contextual focused topic model (cFTM). The cFTM utilizes both the text and contextual information to model a document corpus. It infers a set of semantically meaningful topics to summarize the corpus, as well as the relational information between the authors and venues. It automatically infers the number of topics, the number of author and venue clusters, and the probabilistic importance of the author and venue information on word assignment in a document dependent manner. Efficient MCMC inference is presented. Example results on the DBLP and NSF datasets are used to demonstrate the consistent and promising performance of the proposed cFTM, with quantitative comparison to other state-of-the-art methods and intuitive qualitative analysis. The computational expense of the proposed model is comparable to that of related topic models [28]. In non-optimized Matlab, running on a 2.26GHz CPU computer, each MCMC sample required approximately 2.3 seconds to compute, when considering the DBLP data.

Acknowledgements

The research reported here was supported by ARO, NGA, ONR and DARPA (MSEE program).

7. REFERENCES

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

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks*, volume 30, 1998.

[3] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, 2008.

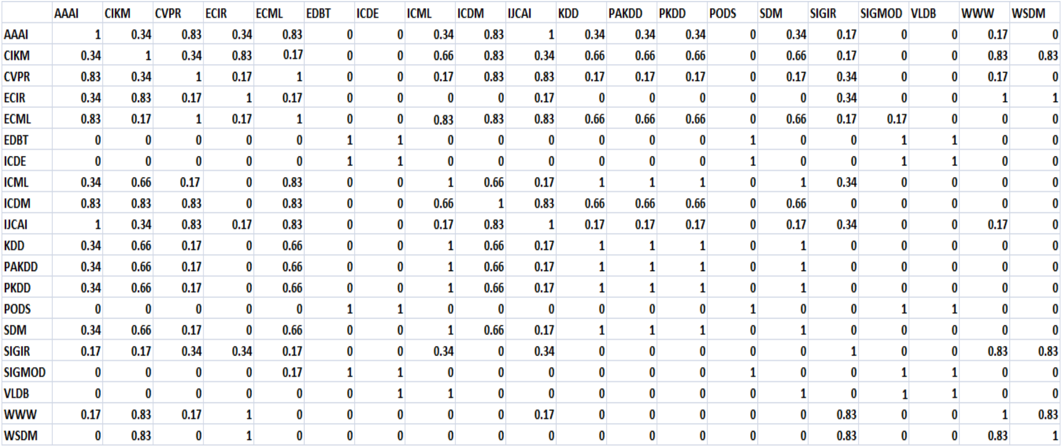
[4] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, 2009.

[5] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, 2000.

[6] H. Deng, J. Han, B. Zhao, Y. Yu, and C. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*, 2011.

[7] H. Deng, M. R. Lyu, and I. King. Effective latent space graph-based re-ranking model with global consistency. In *WSDM*, 2009.

[8] H. Deng, M. R. Lyu, and I. King. A generalized



	AAAI	CIKM	CVPR	ECIR	ECML	EDBT	ICDE	ICML	ICDM	IJCAI	KDD	PAKDD	PKDD	PODS	SDM	SIGIR	SIGMOD	VLDB	WWW	WSDM
AAAI	1	0.34	0.83	0.34	0.83	0	0	0.34	0.83	1	0.34	0.34	0.34	0	0.34	0.17	0	0	0.17	0
CIKM	0.34	1	0.34	0.83	0.17	0	0	0.66	0.83	0.34	0.66	0.66	0.66	0	0.66	0.17	0	0	0.83	0.83
CVPR	0.83	0.34	1	0.17	1	0	0	0.17	0.83	0.83	0.17	0.17	0.17	0	0.17	0.34	0	0	0.17	0
ECIR	0.34	0.83	0.17	1	0.17	0	0	0	0	0.17	0	0	0	0	0	0.34	0	0	1	1
ECML	0.83	0.17	1	0.17	1	0	0	0.83	0.83	0.83	0.66	0.66	0.66	0	0.66	0.17	0.17	0	0	0
EDBT	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0
ICDE	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0
ICML	0.34	0.66	0.17	0	0.83	0	0	1	0.66	0.17	1	1	1	0	1	0.34	0	0	0	0
ICDM	0.83	0.83	0.83	0	0.83	0	0	0.66	1	0.83	0.66	0.66	0.66	0	0.66	0	0	0	0	0
IJCAI	1	0.34	0.83	0.17	0.83	0	0	0.17	0.83	1	0.17	0.17	0.17	0	0.17	0.34	0	0	0.17	0
KDD	0.34	0.66	0.17	0	0.66	0	0	1	0.66	0.17	1	1	1	0	1	0	0	0	0	0
PAKDD	0.34	0.66	0.17	0	0.66	0	0	1	0.66	0.17	1	1	1	0	1	0	0	0	0	0
PKDD	0.34	0.66	0.17	0	0.66	0	0	1	0.66	0.17	1	1	1	0	1	0	0	0	0	0
PODS	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0
SDM	0.34	0.66	0.17	0	0.66	0	0	1	0.66	0.17	1	1	1	0	1	0	0	0	0	0
SIGIR	0.17	0.17	0.34	0.34	0.17	0	0	0.34	0	0.34	0	0	0	0	0	1	0	0	0.83	0.83
SIGMOD	0	0	0	0	0.17	1	1	0	0	0	0	0	0	1	0	0	1	1	0	0
VLDB	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	0	0
WWW	0.17	0.83	0.17	1	0	0	0	0	0	0.17	0	0	0	0	0	0.83	0	0	1	0.83
WSDM	0	0.83	0	1	0	0	0	0	0	0	0	0	0	0	0	0.83	0	0	0.83	1

Figure 3: The venue clustering results of the proposed cFTM algorithm on the DBLP dataset. The (i, j)th element of the matrix represents the fraction of samples among the 3000 collected MCMC samples in which the ith and jth venues are in the same cluster.

Table 3: Most probable words of eight representative topics inferred by the cFTM model on the DBLP dataset. The top key words shown in each topic are ranked based on their probabilities to be used.

Topic 4	Prob	**Topic 10**	Prob	**Topic 6**	Prob	**Topic 11**	Prob
data	0.12374	data	0.14957	information	0.159	learning	0.22921
query	0.10022	mining	0.12614	retrieval	0.143	methods	0.19101
database	0.09728	information	0.10091	web	0.09	knowledge	0.06013
system	0.05506	clustering	0.05262	search	0.05	systems	0.05637
algorithm	0.05452	classification	0.04902	text	0.046	reasoning	0.05011
distributed	0.04651	algorithm	0.04469	model	0.045	model	0.04886
queries	0.0417	based	0.04037	user	0.039	algorithm	0.02882
algorithms	0.0302	analysis	0.02884	document	0.024	logic	0.01817
performance	0.02753	experimental	0.02343	semantic	0.015	representation	0.01091
Topic 3	Prob	**Topic 5**	Prob	**Topic 7**	Prob	**Topic 8**	Prob
show	0.21758	method	0.14874	mining	0.17164	query	0.17989
results	0.17505	methods	0.1235	learning	0.14517	structure	0.15442
system	0.14233	developed	0.0929	algorithm	0.13634	database	0.08111
systems	0.13524	applied	0.06015	clustering	0.09355	language	0.04784
propose	0.02891	based	0.05961	discovery	0.03972	computer	0.03692
set	0.018	results	0.04511	retrieval	0.03575	specific	0.0338
large	0.01528	experimental	0.04296	detection	0.03266	information	0.02341
analysis	0.01091	research	0.02095	image	0.01722	space	0.01457
efficient	0.01076	analysis	0.0129	search	0.0106	order	0.01353

Table 4: The author clustering results on the DBLP dataset. Within each cluster, the five authors with the most number of publications are displayed.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
C. Faloutsos	P. S. Yu	T. S. Huang	Q. Yang	H. P. Kriegel
D. Srivastava	J. Han	T. Kanade	B. Liu	E. J. Keogh
H. G. Molina	C. C. Aggarwal	A. Ng	C. L.Giles	S. B. Zdonik
D. Agrawal	H. Liu	A. K. Jain	P. Domingos	T. Li
N. Koudas	S. Parthasarathy	M. Shah	T. Sandholm	C. Zaniolo
Cluster 6	**Cluster 7**	**Cluster 8**	**Cluster 9**	**Cluster 10**
S. Chaudhuri	G. Weikum	H. Mannila	W. B.Croft	C. T.Yu
R. Ramakrishnan	J. F. Naughton	K. Wang	Z. Chen	W. Y.Ma
W. Wang	E. A.Rundensteiner	V. Kumar	C. Zhai	R. Rastogi
J. Pei	H. Pirahesh	R. Jin	J. Allan	J. Gehrke
H. Wang	D. Gunopulos	R. J. Mooney	R. Kumar	J. X. Yu

co-hits algorithm and its application to bipartite graphs. In *KDD*, 2009.

[9] T. Ferguson. Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1973.

[10] S. Ghosal. Dirichlet process, related priors and posterior asymptotics. *Bayesian Nonparametrics*, pages 35–79, 2010.

[11] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[12] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

[13] M. Hoffman, D. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, 2010.

[14] T. Hofmann. Probabilistic semantic indexing. *SIGIR*, pages 50–57, June 1999.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Journal of the ACM (JACM)*, 1999.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2000.

[17] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002.

[18] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008.

[19] R. Nallapati, A. Ahmed, E. P. Xing, and W. Cohen. Joint latent topic models for text and citations. In *KDD*, 2008.

[20] L. Ren, L. Du, L. Carin, and D. Dunson. Logistic stick-breaking process. In *The Journal of Machine Learning Research*, volume 12, 2011.

[21] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1994.

[22] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, 2004.

[23] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, 2009.

[24] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, 2008.

[25] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[26] 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.

[27] R. J. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, 2007.

[28] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

[29] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR*, 2005.

[30] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS*, 2003.

[31] D. Zhou, B. Scholkopf, and T. Hofmann. Semisupervised learning on directed graphs. In *NIPS*, 2004.

[32] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

[33] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin. Dependent hierarchical beta process for image interpolation and denoising. In *AISTATS*, 2011.

[34] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.