Infinite Author Topic Model based on Mixed Gamma-Negative Binomial Process

Junyu Xuan
University of Technology Sydney
15 Broadway
Sydney, Australia
Junyu.Xuan@student.uts.edu.au

Jie Lu
University of Technology Sydney
15 Broadway
Sydney, Australia
Jie.Lu@uts.edu.au

Guangquan Zhang
University of Technology Sydney
15 Broadway
Sydney, Australia
Guangquan.Zhang@uts.edu.au

Richard Yi Da Xu
University of Technology Sydney
15 Broadway
Sydney, Australia
Yida.Xu@uts.edu.au

Xiangfeng Luo
Shanghai University
99 Shangda Road
Shanghai, China
luoxf@shu.edu.cn

ABSTRACT

Incorporating the side information of text corpus, i.e., authors, time stamps, and emotional tags, into the traditional text mining models has gained significant interests in the area of information retrieval, statistical natural language processing, and machine learning. One branch of these works is the so-called Author Topic Model (ATM), which incorporates the authors's interests as side information into the classical topic model. However, the existing ATM needs to predefine the number of topics, which is difficult and inappropriate in many real-world settings. In this paper, we propose an Infinite Author Topic (IAT) model to resolve this issue. Instead of assigning a discrete probability on fixed number of topics, we use a stochastic process to determine the number of topics from the data itself. To be specific, we extend a gamma-negative binomial process to three levels in order to capture the author-document-keyword hierarchical structure. Furthermore, each document is assigned a mixed gamma process that accounts for the multi-author's contribution towards this document. An efficient Gibbs sampling inference algorithm with each conditional distribution being closed-form is developed for the IAT model. Experiments on several real-world datasets show the capabilities of our IAT model to learn the hidden topics, authors' interests on these topics and the number of topics simultaneously.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

topic models, nonparametric Bayesian learning, text mining, gamma processes, negative binomial process

1. INTRODUCTION

Traditional text mining algorithms only model the text corpus with two levels: document-word. Topic models are commonly regarded as the efficient tools for the text mining by learning the hidden topics [15]. Recently, interests have been paid on the side information of the text corpus, which includes the conferences of the papers [21], time stamps [24], authors [18, 20], entities [9], emotion tags [1] and other labels [28]. The incorporation of these side information into the classical topic models benefits a lot of real-world tasks. Among them, Author Topic Model (ATM) [18, 20, 17] is proposed by adding a set of variables to the original topic model aiming to indicate and inference the interests of authors together with the hidden topics.

The ability to jointly learn the hidden topics and authors' interests on these topics has a variety of application scenarios. For example, 1) an academic recommendation system can recommend authors and/or papers with similar research interests to that of the input author; 2) detecting the most and least surprising papers for an author [20]; 3) in an author-topic-based paper browser, a set of papers can be ranked according to authors and topics; 4) authors disambiguation [26].

One drawback of the existing author topic model is that the number of hidden topics needs to be fixed in advance. This number is normally chosen with domain knowledge. By fixing the number of topics, ATM can then adopt Dirichlet and Multinomial distributions with the pre-defined dimension. However, limiting each document to have exactly fixed number of topics is apparently unrealistic for many real-world applications. In this paper, we propose an infinite author topic (IAT) model to relax this assumption. Instead of using fixed-dimensional distributions, stochastic processes are used: to be specific, the gamma-negative binomial process [27] is extended to three levels for capturing the hierarchical structure: author-document-keyword. In this model,

each document is assigned with a gamma process to express the interest of this document on the hidden topics instead of a vector with a fixed dimension. This gamma process can be simply considered as an infinite discrete distribution, and is parameterized by a base measure (another gamma process) that denotes the interest of the author of this document on the hidden topics. However, a document normally has multiple authors, so we assign a document a mixed gamma process that is based on all the gamma processes of the authors of this document. Furthermore, introducing mixed gamma process will lead to intricacies in terms of model inference. Therefore, an efficient Gibbs sampling with closed-form conditional distributions is developed for the proposed model. Experiments on the two real-world datasets show the capability of our model to learn both the hidden topics and the number of topics, simultaneously.

The main contributions of this paper are,

1. propose a new nonparametric Bayesian model to relax the fixed topic number assumption of the traditional author topic models;

2. design an efficient Gibbs sampling inference algorithm for getting the solution of the proposed model.

The rest paper is structured as follows. Section 2 briefly introduces the related work. Section 3 describes some preliminary knowledge. The IAT model is proposed and presented in Section 4 with its Gibbs sampling inference algorithm. Section 5 describes the IAT model experimental results using real-world datasets. Finally, Section 6 concludes this study with a discussion on future directions.

2. RELATED WORK

In this section, we briefly review the related work of this study. The first part is about the topic models, and the second part is about nonparametric Bayesian learning.

2.1 Topic Models

Topic models [2] are Bayesian models with fixed-dimensional probability distributions. They are originally designed for the text mining task, which aim to discover the hidden topics in the text corpus to assist document clustering or classification. Due to their good extendibility and powerful representation, they have been successfully applied to many research areas, including analysis in image [12], video [10], genetics [5] and music [13]. Among these extensions, author topic models [18, 20, 17] were proposed to infer the hidden topics and author interests. The documents are supposed to be generated by its authors according to their interests over the hidden topics. This model will be explained with more details in Section 3.

ATM has attracted a lot of attentions from researchers working in the text mining area, because it provides an elegant way to incorporate the side (in this case, author) information of the documents for topic learning. This model can be extend to incorporate other side information of text corpus, such as emotional tags [1], conferences[21] and time stamps [24].

2.2 Nonparametric Bayesian Learning

Nonparametric Bayesian learning is a key approach for learning the number of mixtures in a mixture model (also called model selection problem). Without predefining the
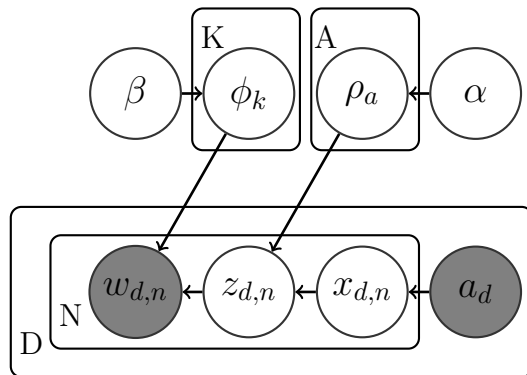

Figure 1: Classical Author-Topic Model

number of mixtures, this number is supposed to be inferred from the data, i.e., let the data speak.

The idea of nonparametric Bayesian learning is to use the stochastic processes to replace the traditional fixed-dimensional probability distributions, such as Multinomial, Poisson, and Dirichlet. In order to avoid the limitation associated with fixed dimensions, Multinomial Process (MP), Poisson Process (PP) [8] and Dirichlet Process (DP) [22] are used to replace former distributions because of their infinite properties.

The merit of these stochastic processes is that they let the data to determine the number of factors (in text mining task, topics). DP is a good alternative for the models with Dirichlet distribution as the prior. Many probabilistic models with fixed dimensions have been extended to the infinite ones by the help of stochastic processes: Gaussian Mixture Model (GMM) is extended to Infinite Gaussian Mixture Model (IGMM) [16] using DP; Hidden Markov Model is extended with infinite number of hidden states using Hierarchial Dirichlet Process [23, 7]. Through the posterior inference (i.e., Markov chain Monte Carlo (MCMC) [11]), the number of the mixtures can be inferred. Other popular processes include beta process, gamma process, poisson process, multinomial process, negative binomial process (NBP) [27, 3] have also been used in the machine learning communities recently.

To summarize, nonparametric Bayesian learning [4] has been successfully used to extend many finite models and applied to many real-world applications. However, to the best of our knowledge, there has not been any works proposed to use NBP for author topic modelling. This paper addresses this shortcoming by proposing a mixed gamma negative binomial process to extend the finite author topic model to the infinite one.

3. PRELIMINARY KNOWLEDGE

This section briefly introduces the related models which will be used in the rest of sections.

3.1 Author Topic Model

The Author Topic Model [18, 20, 17] aims to learn the hidden topics from the papers and more importantly learn the authors' interests on these topics. Based on the classical LDA [2], a set of new variables are introduced to indicate the authors' interests. The graphical representation of the model is shown in Fig. 1, and the generative procedure is

as follows,

$$\rho_a \overset{i.i.d}{\sim} Dirichlet(\alpha)$$
$$\phi_k \overset{i.i.d}{\sim} Dirichlet(\beta)$$
$$x_{d,n} \sim Unif(a_d) \tag{1}$$
$$z_{d,n} \sim Discrete(\rho_{x_{d,n}})$$
$$w_{d,n} \sim Discrete(\phi_{z_{d,n}})$$

where $\{\rho_a\}_{a=1}^{A}$ denote the authors' interests on the topics and a_d denotes the authors of a document. We can see from the Eq.(1) that the ATM is constructed by the fixed-dimensional probability distributions. One issue of this model is that the number of topics needs to be predefined, because the dimensions of the probability distributions need to be predefined. However, it is very difficult and not appropriate to predefine the topic number in many real-world scenarios.

3.2 Gamma Negative Binomial Process

3.2.1 Gamma Process

A gamma process $GaP(\alpha, H)$ [19] is a stochastic process, where H is a base (shape) measure and α is the concentration (scale) parameter. It also corresponds to a complete random measure. Let $\Gamma = \{(\pi_i, \theta_i)\}_{i=1}^{\infty}$ be a random realization of a Gamma process in the product space $R^+ \times \Theta$. Then, we have

$$\Gamma \sim GaP(\alpha, H)$$
$$= \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \tag{2}$$

where $\delta(\cdot)$ is an indicator function, π_i satisfies an improper gamma distribution $gamma(0, \alpha)$, and $\theta_i \sim H$. After the normalization of the $\{\pi\}$, we can get the famous Dirichlet process [22].

3.2.2 Negative Binomial Process

A negative binomial process $NBP(p, G_0)$ [27] is also a stochastic process parameterized by a base measure G_0 and p. Similar with the gamma process, a realization of negative binomial process $X = \{(n_i, \theta_i)\}_{i=1}^{\infty}$ is also a set of points in product space $Z^+ \times \Theta$. Then, we have

$$X \sim NBP(p, G_0)$$
$$= \sum_{i=1}^{\infty} n_i \delta_{\theta_i} \tag{3}$$

where $\{n_i\}$ are integers, so negative binomial process is normally used for the counting model [3]. Compared with Poisson process which is also suitable for the counting model, negative binomial process has a better variance-to-mean ratio (VMR) and the overdispersion level [27].

3.2.3 Gamma-Negative Binomial Process

Normally, negative binomial process is used as the likelihood part of a Bayesian model. Like a negative binomial distribution $x \sim NB(r, p)$ which has two parameters: $r > 0$ and $p \in [0, 1]$, there are two kinds of priors for a negative binomial process: one is Gamma process [27] as shown in Eq. (3); the other is the Beta process [3] as $X \sim NBP(B, r)$. In this paper, we use the Gamma process prior. A gamma-negative binomial process-based topic model is proposed in
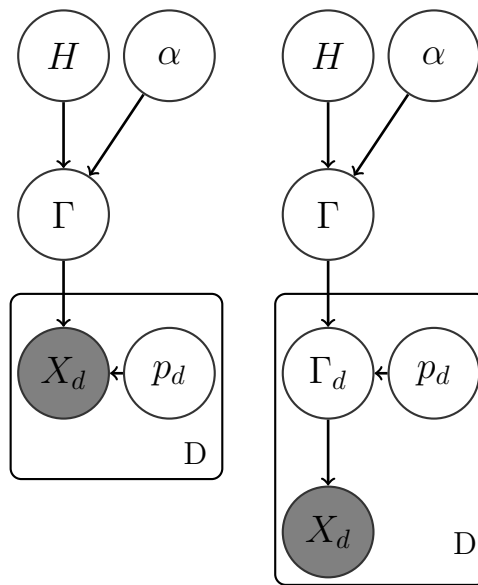

Figure 2: Gamma-Negative binomial process topic model

Table 1: Notations used in this paper

Notation	description
D	number of documents
A	number of authors
N	number of words
AD	author-document mapping matrix
DN	document-word mapping matrix
A_d	number of authors of document d

[27] as shown in Fig. 2 and it can be represented as,

$$\Gamma \sim GaP(c_0, H)$$
$$X_d \sim NBP(p_d, \Gamma) \tag{4}$$

where the base measure of the negative binomial process Γ is a random measure from a gamma process. X_d is for each document, and this hierarchial form makes the documents share a same base measure Γ. This gamma-negative binomial process can be equivalently augmented as gamma-gamma-poisson process,

$$\Gamma \sim GaP(c_0, H)$$
$$\Gamma_d \sim GaP\left(\frac{1-p_d}{p_d}, \Gamma\right) \tag{5}$$
$$X_d \sim PP(\Gamma_d)$$

where $X_d \sim PP(\Gamma_d)$ is a Poisson process with parameter Γ_d. This augmentation, which is useful for the close-form model inference algorithm design, is equal to gamma-negative binomial process model in distribution.

In this paper, we will build an infinite author topic model based on this gamma-negative binomial process model.

4. INFINITE AUTHOR TOPIC MODEL

In this section, we first propose our infinite author topic (IAT) model, and then introduce its Gibbs sampling strategy to inference the proposed model.
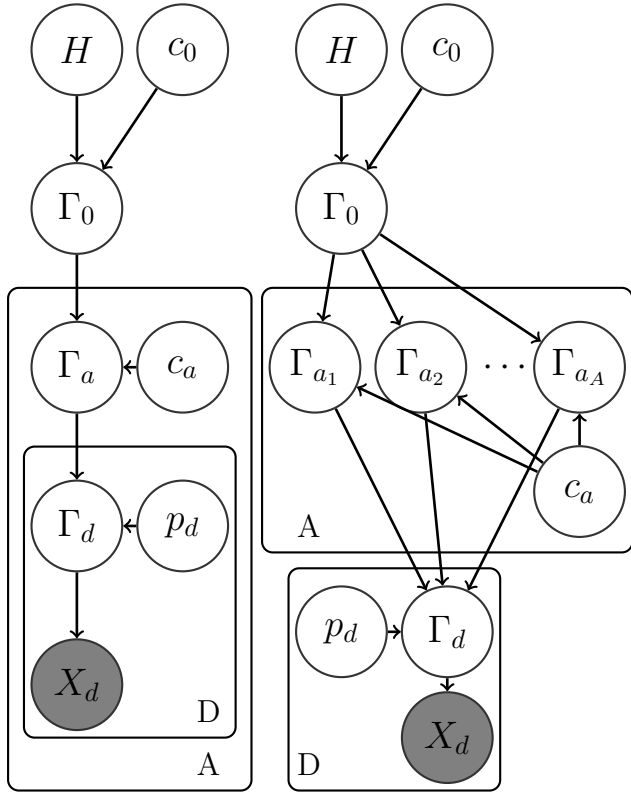
Figure 3: Gamma-Gamma-Negative Binomial Process Model (3GNB) (left one) and Infinite Author Topic Model (IAT) (right one)

4.1 Model Description

Consider the gamma-negative binomial process topic model in Eqs. (4) and (5) again: despite its successful, this model however is fundamentally the same as the basic topic models, which are used for modeling the data of two level hierarchy: document-keyword. Our aim is to extend topic model into three-level hierarchy: author-document-keyword. So we add another gamma process level to capture the additional (author) level based on the gamma-negative binomial process topic model in Eq.(5) analogues to the hierarchical form of Hieratical Dirichlet Process [23],

$$
\begin{aligned}
\Gamma_0 &\sim GaP(c_0, H) \\
\Gamma_a &\sim GaP(c_a, \Gamma_0) \\
\Gamma_d &\sim GaP((1-p_d)/p_d, \Gamma_a^d) \\
X_d &\sim PP(\Gamma_d)
\end{aligned} \tag{6}
$$

where Γ_a is the new added level for the authors. We call this model three-level gamma-negative binomial process topic model (3GNB), which is graphically shown in the left sub-figure of Fig. 3. However, there is a problem in 3GNB that it requires each document with only one author.

In the 3GNB model, each document is assigned a realization of gamma process,

$$
\Gamma_d = \sum_{k=1}^{\infty} \pi_{d,k} \delta_{\theta_k} \tag{7}
$$

where θ_k denotes the kth topic and $\pi_{d,k}$ is the weight of kth

topic. $\{\pi_{d,k}\}_{k=1}^{\infty}$ can be viewed as the interest of document d on the topics. The number of topics can potentially be infinite and therefore justifies the infinity in the summation. However, since the data is limited, the learned topics will be also limited. Similar to the document, each author is also assigned a realization of gamma process,

$$
\Gamma_a = \sum_{k=1}^{\infty} \pi_{a,k} \delta_{\theta_k} \tag{8}
$$

where $\{\pi_{a,k}\}_{k=1}^{\infty}$ is the weight of interests of author a on the topics. In the 3GNB model, the base measure for a Γ_d is from its author Γ_a. It can be seen as the 'interest inheritance'.

In order to model in the setting where a document is with multiple authors, we combine all the gamma processes of every authors of a document together by

$$
\Gamma_a^d = \Gamma_{a_1} \oplus \Gamma_{a_2} \oplus \cdots \oplus \Gamma_{a_{A_d}} \tag{9}
$$

where A_d is the number of authors of document d, $\oplus$ is the convex combination (each gamma process is with same weight in this paper) and Γ_a^d is the mixed prior for Γ_d. We can see the mixed gamma process Γ_a^d as the mixed interests of all the authors of a document. Then, the revised model is as follow

$$
\begin{aligned}
\Gamma_0 &\sim GaP(c_0, H) \\
\Gamma_a &\sim GaP(c_a, \Gamma_0) \\
\Gamma_a^d &= \Gamma_{a_1} \oplus \Gamma_{a_2} \oplus \cdots \oplus \Gamma_{a_{A_d}} \\
\Gamma_d &\sim GaP((1-p_d)/p_d, \Gamma_a^d) \\
X_d &\sim PP(\Gamma_d)
\end{aligned}
$$

and the graphical representation is shown in Fig. 3. Some frequently used notations are explained in Table 1.

4.2 Model Inference

It is difficult to perform posterior inference under infinite mixtures, a common work-around solution in nonparametric Bayesian learning is to use a truncation method. Truncation method is widely accepted, which uses a relatively big K as the (potential) maximum number of topics. Under the truncation, the model can be expressed below as a good approximation to the infinite model,

$$
\begin{aligned}
\gamma_0 &\sim Gamma(e_0, 1/f_0) \\
r_{0,k}|\gamma_0, c_0 &\sim Gamma(\gamma_0/K, 1/c_0) \\
r_{a,k}|r_0, c_a &\sim Gamma(r_{0,k}, 1/c_a) \\
p_d &\sim beta(a_0, b_0) \\
r_{a,k}^d &= r_{a_1,k} \oplus r_{a_2,k} \oplus \cdots \\
r_{d,k}|r_a, p_d &\sim Gamma(r_{a,k}^d, p_d/(1-p_d)) \\
n_{d,k} &\sim Pois(r_{d,k}) \\
N_d &= \sum_{k=1}^{K} n_{d,k} \\
\theta_{1:K} &\sim \frac{1}{\gamma_0} H \\
z_{d,n} &\sim Multi(r_{d,1}/\sum r_d, r_{d,2}/\sum r_d, r_{d,3}/\sum r_d, \cdots) \\
w_{d,n} &\sim \theta_{z_{d,n}}
\end{aligned}
$$

where $\gamma_0 = \int dH$ is the total mass of measure H, and the parameters are given the appropriate priors. Here, H is a

N-dimensional Dirichlet distribution, and each θ is a topic that is a N-dimensional vector.

The difficult part of the inference for this model is the mixed part Γ_a^d or r_a^d. Since $r_a^d = r_{a_1} \oplus r_{a_2} \oplus \cdots$ is the mixed value, it is hard to infer the posterior of r_a through its likelihood. In order to resolve this issue, we firstly introduce the Additive Property of the negative binomial distribution,

THEOREM 1. *If X_i follows a negative binomial distribution with parameters r_i and p and if the various X_i are independent, then $\sum X_i$ follows a negative binomial distribution with parameters $\sum r_i$ and p.*

In the model, we have

$$r_{d,k}|\{r_a\}, p_d \sim Gamma(r_{a,k}^d, p_d/(1-p_d))$$
$$n_{d,k} \sim Pois(r_{d,k}) \tag{10}$$

(in distribution) equal to

$$n_{d,k} \sim NB(r_{a,k}^d, p_d) \tag{11}$$

and according to THEOREM 1, it is further (in distribution) equal to

$$n_{d,k}^a \sim NB\left(\frac{r_{a,k}}{A_d}, p_d\right)$$
$$n_{d,k} = \sum_a n_{d,k}^a \tag{12}$$

where A_d is the number of authors in document d.

We have split $n_{d,k}$ the number of words assigned to topic k in document d into a number A_d of independent variables $\{n_{d,k}^a\}$. Here, $n_{d,k}^a$ denotes the number of words assigned to topic k from author a in document d. From Eq.(12), we can see that we have the likelihood part of the r_a, so we can update/inference the r_a using n_d^a. Introducing the auxiliary variables $\{n_{d,k}^a\}$ helps us resolve the difficult inference problem brought by the mixed gamma process. Note that the independence between the elements of $\{n_{d,k}^a\}$ is very important, which facilitates us update each $n_{d,k}^a$ independently.

According to the relationship between the negative binomial distribution and the gamma-poisson distribution, for each $n_{d,k}^a$, we have:

$$n_{d,k}^a \sim NB(\frac{r_{a,k}}{A_d}, p_d)$$
$$\Longrightarrow r_{d,k}^a \sim Gamma(\frac{r_{a,k}}{A_d}, p_d/(1-p_d)), \quad n_{d,k}^a \sim Pois(r_{d,k}^a) \tag{13}$$

We want to highlight that $r_{d,k}^a$ is different from $r_{a,k}^d$: $r_{a,k}^d$ is the mixed Gamma process of multiple author Gamma processes Γ_a of Gamma process Γ_d of document d and $r_{d,k}^a$ is the interest of document d on topic k inherited from author a.

Due to the non-conjugacy of gamma distribution and negative binomial distribution, it is difficult to update r_a with a gamma prior. In order to make the inference with only close-formed conditional distributions, we use the following results on the negative binomial process,

THEOREM 2. *[14, 27] If X follows a negative binomial distribution $X \sim NB(r, p)$ with parameters r and p, then X can also be generated from a compound poisson distribution as*

$$X = \sum_{t=1}^{l} u_t, t \stackrel{i.i.d}{\sim} Log(p), l \sim poiss\left(-rln(1-p)\right) \tag{14}$$

where $Log()$ is a Logarithmic distribution. Furthermore, this poisson-logarithmic bivariate count distribution, $p(X, l)$, can be expressed as

$$X \sim NB(r, p), \; l \sim CRT(X, r) \tag{15}$$

where CRT denotes Chinese restaurant Table distribution.

With THEOREM 2, the Eq. (13) is also equal to

$$n_{d,k}^a \sim NB(\frac{r_{a,k}}{A_d}, p_d)$$
$$\Longrightarrow n_{d,k}^a \sim \sum_{1}^{l_{d,k}^a} log(p_d), \quad l_{d,k}^a \sim Pois(-\frac{r_{a,k}}{A_d} \cdot ln(1-p_d))$$
$$\Longrightarrow l_{d,k}^a \sim CRT(n_{d,k}^a, \frac{r_{a,k}}{A_d}), \quad n_{d,k}^a \sim NB(\frac{r_{a,k}}{A_d}, p_d) \tag{16}$$

Finally, we can update all $n_{d,k}^a$ by,

$$(n_{d,k_1}^{a_1}, n_{d,k_2}^{a_1}, \cdots, n_{d,K}^{a_A}) \sim Mult(n_d, \frac{r_{d,k_1}^{a_1}}{r}, \frac{r_{d,k_2}^{a_1}}{r}, \cdots, \frac{r_{d,K}^{a_A}}{r})$$
$$r = \sum_a \sum_k r_{d,k}^a \tag{17}$$

and for each word n in a document d, we can assign it to a topic k and author a by

$$p(z_{d,n} = k, i_{d,n} = a) \propto \frac{r_{d,k}^a}{r}$$
$$n_{d,k} = \sum_n \delta(z_{d,n} = k)$$
$$n_{a,k} = \sum_d \sum_n \delta(z_{d,n} = k \; AND \; i_{d,n} = a) \tag{18}$$

With these changes of variables, the original model is reformulated as,

$$\gamma_0 \sim Gamma(e_0, 1/f_0)$$
$$r_{0,k}|\gamma_0, c_0 \sim Gamma(\gamma_0/K, 1/c_0)$$
$$p_d \sim beta(a_{d,0}, b_{d,0})$$
$$r_{a,k}|r_0, c_a \sim Gamma(r_{0,k}, 1/c_a)$$
$$r_{a,k}^d = r_{a_1,k} \oplus r_{a_2,k} \oplus \cdots$$
$$r_{d,k}|r_a, p_d \sim Gamma(r_{a,k}^d, p_d/(1-p_d))$$
$$r_{d,k}^a \sim Gamma(\frac{r_{a,k}}{A_d}, p_d/(1-p_d)), \quad a \in A^d$$
$$z_{d,n}^a \sim Discrete(\frac{r_{d,k}^a}{r}, \cdots) \tag{19}$$
$$n_{d,k} = \sum_n \delta(z_{d,n} = k)$$
$$n_{a,k} = \sum_d \sum_n \delta(z_{d,n} = k \; AND \; i_{d,n} = a)$$
$$n_{d,k}^a = \sum_n \delta(z_{d,n} = k \; AND \; i_{d,n} = a)$$
$$N_d = \sum_n \sum_a z_{d,n}^a$$

In the following, a Gibbs sampling algorithm is designed for the posterior inference and all the conditional distributions are listed.

Sampling z

$$p(z_{d,n} = k, i_{d,n} = a|\cdots) \propto \theta_{k,n} \cdot r_{d,k}^a \tag{20}$$

Sampling r_d^a

$$p(r_{d,k}^a|\cdots) \propto Gamma(\frac{r_{a,k}}{A_d} + n_{d,k}^a, p_d) \quad (21)$$

where $n_{d,k}^a$ is the number of words in document d with author a and topic k.

Sampling l_d^a

$$p(l_{d,k}^a|\cdots) \propto CRT\left(n_{d,k}^a, \frac{r_{a,k}}{A_d}\right) \quad (22)$$

Sampling p_d

$$r_{a,k}^d = r_{a_1,k} \oplus r_{a_2,k} \oplus \cdots$$

$$p(p_d|\cdots) \propto Beta\left(a_0 + \sum_k n_{d,k}, b_0 + \sum_k r_{a,k}^d\right) \quad (23)$$

$$p(r_{d,k}|\cdots) \propto Gamma(r_{a,k}^d + n_{d,k}, p_d)$$

Sampling r_a

$$p(r_{a,k}|\cdots)$$

$$\propto Gamma\left(r_{0,k} + \sum_{d \text{ with } a} l_{d,k}^a, \frac{1}{c_a - \sum_{d \text{ with } a} \frac{1}{A_d} \cdot \ln(1-p_d)}\right) \quad (24)$$

Sampling l_a

$$p(l_{a,k}|\cdots) \propto CRT\left(\sum_{d \text{ with } a} l_{d,k}^a, r_{0,k}\right) \quad (25)$$

Sampling $r_{0,k}$

$$p(r_{0,k}|\cdots) \propto Gamma\left(\gamma_0/K + \sum_a l_{a,k}, \frac{1}{c_0 - \sum_a \ln(1-p_a)}\right) \quad (26)$$

where

$$p_a = \frac{-\sum_{d \text{ with } a} \frac{1}{A_d} ln(1-p_d)}{c_a - \sum_{d \text{ with } a} \frac{1}{A_d} ln(1-p_d)} \quad (27)$$

Sampling l_k'

$$p(l_k'|\cdots) \propto CRT\left(\sum_a l_{a,k}, \gamma_0/K\right) \quad (28)$$

Sampling γ_0

$$p(\gamma_0|\cdots) \propto Gamma\left(e_0 + \sum_k l_k', \frac{1}{f_0 - \ln(1-p')}\right) \quad (29)$$

where

$$p' = \frac{-\sum_a ln(1-p_a)}{c_0 - \sum_a ln(1-p_a)} \quad (30)$$

Sampling θ_k

$$p(\theta_k|\cdots) \propto H(\theta_k) \prod_d \theta_{z_{d,n}=k,n} \quad (31)$$

We can see from these conditional distributions that all of them are closed-form which is very easy to updated and implemented. Note that the sampling of the CRT distribution can be found in [27]. The whole procedure is summarized in Algorithm 1.

Note that after we obtain all the samples of the posterior $p(\theta, r_a, r_d, r_0, z_{d,n}^a, p_d, \gamma_0, n_{d,k}^a|N_d, AD, DN, e_0, f_0, c_0, c_a, a_0, b_0)$

Algorithm 1: Gibbs Sampler for IAT

Input: D, A, N, AD, DN
Output: K_{real}, $\{\theta\}$, $\{r_a\}$, $\{r_d\}$
initialization;
while $iter \leq max_{iter}$ **do**
  for $d = 1; d \leq D$ **do**
    for $n = 1; n \leq N_d$ **do**
      Update $z_{d,n}$ and $i_{d,n}$ by Eq. (20);
    for $a = 1; a \leq A_d$ **do**
      Update $r_{d,k}^a$ by Eq. (21); Update $l_{d,k}^a$ by Eq. (22);
    Update $r_{d,k}$ and p_d by Eq. (23);
  for $a = 1; a \leq A$ **do**
    Update $r_{a,k}$ by Eq. (24); Update $l_{a,k}$ by Eq. (25);
  Update $r_{0,k}$ by Eq. (26); Update l_k' by Eq. (28);
  Update γ_0 by Eq. (29); Update θ by Eq. (31);
  $iter++$;
Identify K_{real};
Select the sample with largest likelihood and
$K = K_{real}$;
return $\{\theta\}$, $\{r_a\}$, $\{r_d\}$;

Table 2: Statistics of Datasets

Datasets	D	A	N
NIPS	1,740	2,037	13,649
DBLP	28,569	28,702	11,771

Table 3: Groups of Datasets DBLP

	D training	D test	A	N
group 1	1,072	319	1,115	3,783
group 2	1,071	316	1,094	3,782
group 3	1,075	305	1,071	3,788
group 4	1,076	339	1,104	3,823
group 5	1,079	310	1,111	3,841

Table 4: Groups of Datasets NIPS

	D training	D test	A	N
group 1	1,503	237	2,037	5,110
group 2	1,495	245	2,037	5,110
group 3	1,511	229	2,037	5,110

of latent variables and remove the burn-in stage, we firstly identify the topic number with largest frequency as the K_{real}, and then find the sample with largest likelihood and $K = K_{real}$ from these samples. The output of Gibbs sampler are the latent variables θ, r_a and r_d in this sample.

5. EXPERIMENTS

In this section, we evaluate the proposed infinite author topic model (IAT), and compare it with the finite author-topic model (ATM) on different datasets.

5.1 Datasets

Two public datasets used in this paper are:

- **NIPS papers**[1] This dataset contains papers from the NIPS conferences between 1987 and 1999. More description can be found in the [20];

- **DBLP papers**[2] The abstracts and authors of papers are extracted through DBLP interface from four areas: database, data mining, information retrieval and artificial intelligence. More description can be found in the [6].

Some statistics of two datasets are shown in Table 2. For each dataset, we randomly select some documents as training data and test data. The Table 4 and Table 3 show the selection results on two datasets. The number of selected training and test documents are specialized in column *D training* and column *D test* in Table 4 and 3. The requirements of selections is: the training and test documents must share some authors and some words. This requirement makes sure the learned topics and authors' interests can be used to predict the test documents.

5.2 Evaluation Metrics

In order to evaluate the performance of the proposed model, we calculate the perplexity of the test documents using the learned topics and author interests on these topics. Perplexity is widely used in language modeling to assess the predictive power of a model [20, 2]. It is a measure of how surprising the words in the test documents are from the model's perspective. It can be computed as,

$$
\begin{aligned}
Perplexity &= \exp\left(-\sum_d p(\mathbf{w_d}|a_d)\right) \\
&= \exp\left(-\sum_d \sum_k p(\mathbf{w_d}|\theta_k)p(\theta_k|a_d)\right)
\end{aligned}
\tag{32}
$$

where a_d is the authors of test document d. The smaller the value of perplexity is, the better the predictive ability of a model has. Since we use the same test documents for different models, the normalization is not considered because it does not influence the model comparisons.

Another evaluation metric is the training data likelihood,

$$
logLikelihood = \sum_d \log p(w_d|\theta, \mathbf{r_a}, \mathbf{r_d})
\tag{33}
$$

This is a measure of the probability of the training document under the learned latent variables θ, $\mathbf{r_a}$ and $\mathbf{r_d}$. It can be understood as 'how the model fits the training data'. The bigger the value of likelihood is, the better a model fits the training data.

5.3 Results Analysis

For the DBLP dataset, the results are all shown in Fig. 4. Each row of the Fig. 4 denotes a group of DBLP dataset corresponding to Table 3. The left subfigures show the comparison on the data log-likelihood. Here, we adjust different active topic numbers for the ATM, including $K = 100$, $K = 200$, $K = 300$, $K = 400$ and $K = 500$. From these subfigures, the proposed IAT model (The hyper-parameters are set as following by experiences for the rest of this section: $a_0 = 1$, $b_0 = 1$, $e_0 = 1$, $f_0 = 1$, $c_0 = 1$ and $c_a = 1$) outperforms the ATM on different preset topic numbers. It means

[1]http://www.datalab.uci.edu/author-topic/NIPs.htm
[2]http://www.cs.uiuc.edu/ hbdeng/data/kdd2011.htm

that IAT fits the training documents better than the ATM, and, more importantly, IAT does not depend the domain knowledge to predefine the active topic number, making the method widely applicable.

The middle subfigures in Fig. 4 indicate the changing of active topics during the iteration of the IAT (The number of active topics is set as the number of training documents at the initialization step of the model). These curves show that the number of active topics dramatically drops down at the burn-in stage of the sampling, and began to stabilize after about 200 iterations. Since the documents are different in content but similar in numbers amongst the groups, the learned topic number is differ slightly amongst each others. These numbers are: group 1: $K = 519$; group 2: $K = 332$; group 3: $K = 493$; group 4: $K = 465$; group 5: $K = 504$.

In order to show the effectiveness of the proposed model, we also compare the performances of two models (IAT and ATM) on the test documents prediction using perplexity in Eq. (32). Since the training and test documents share some authors, we can compute the perplexity of the test documents according to the learned topics and authors' interests on them. At each step of iterations, the perplexity of test documents is computed using the latent variables, $\{\theta\}$, $\{r_a\}$ and $\{r_d\}$, at this iteration. The results are shown in right subfigures of Fig. 4. In each subfigure, the first bar denotes the mean of perplexities of all iterations except the burn-in stage ($1 \sim 200$ iterations) of the proposed model IAT and the others denote ATM with different (predefined) topic numbers. The standard deviations are also shown in the subfigures. The proposed model gets the best performance (smallest perplexity). The standard deviation of IAT is relatively bigger than ATM. The reason is because the number of active topics will change during the iteration but it will not change in ATM, so in theory, the random-walk space of Gibbs sampler of IAT can be larger than that of ATM. Even with this relatively larger standard deviation, the mean of perplexity of IAT is smaller than ATM.

For the NIPS dataset, the results are all shown in Fig. 5. Same with the DBLP dataset, the log likelihoods of IAT and ATM with different predefined active topic numbers are shown in the left side of the Fig. 5. Unsurprisingly, the subfiguers in the middle column show the convergence of IAT (group 1: 367; group 2: 529; group 3: 354). Specially, we found that the log-likelihoods of ATM increases when topic number decreases. Therefore, we have compared with ATM with only two (the minimum number) topics as shown in the left subfigures in Fig. 5. It can be seen that the proposed IAT model also gets larger log likelihood and smaller perpetuity when compared with ATM except the case where ATM is set to have 10 topics in group 2. Even so, the ATM in group 2 with 10 topics has almost same performance with IAT on the Log-likelihood of training documents. Moreover, we can see that it takes 800 iterations to reach this stability for the ATM with 10 topics, but IAT only takes less than 50 iterations to reach the same stability.

6. CONCLUSIONS AND FURTHER STUDY

We have developed an infinite author topic model that can automatically learn completely the latent features of the author-document-keywords hierarchy, which include hidden topics, authors' interests on these topics and the number of topic from text corpora. The stochastic processes are adopted instead of the fixed-dimensional probability distri-
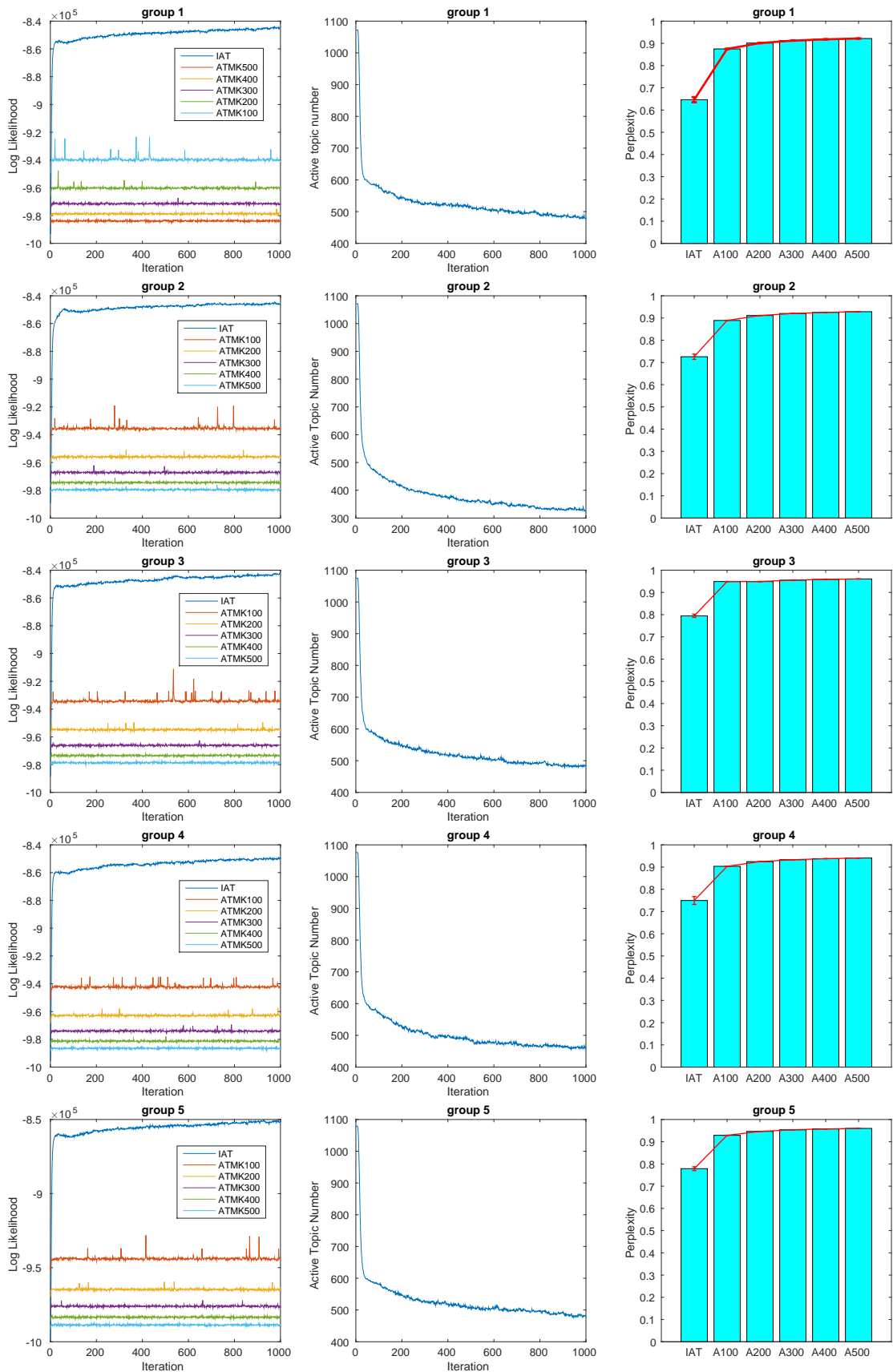
Figure 4: Results from IAT and ATM on five groups of DBLP dataset. Each row denotes a group. In each row, the left subfigure shows the Log-likelihoods comparison between IAT and ATM with different (predefined) topic numbers: $K = 100$, $K = 200$, $K = 300$, $K = 400$, and $K = 500$; The middle subfigure shows the change of active topic number of IAT during the iteration of Gibbs sampling; the right subfigure shows the perplexity comparison between IAT and ATMs.
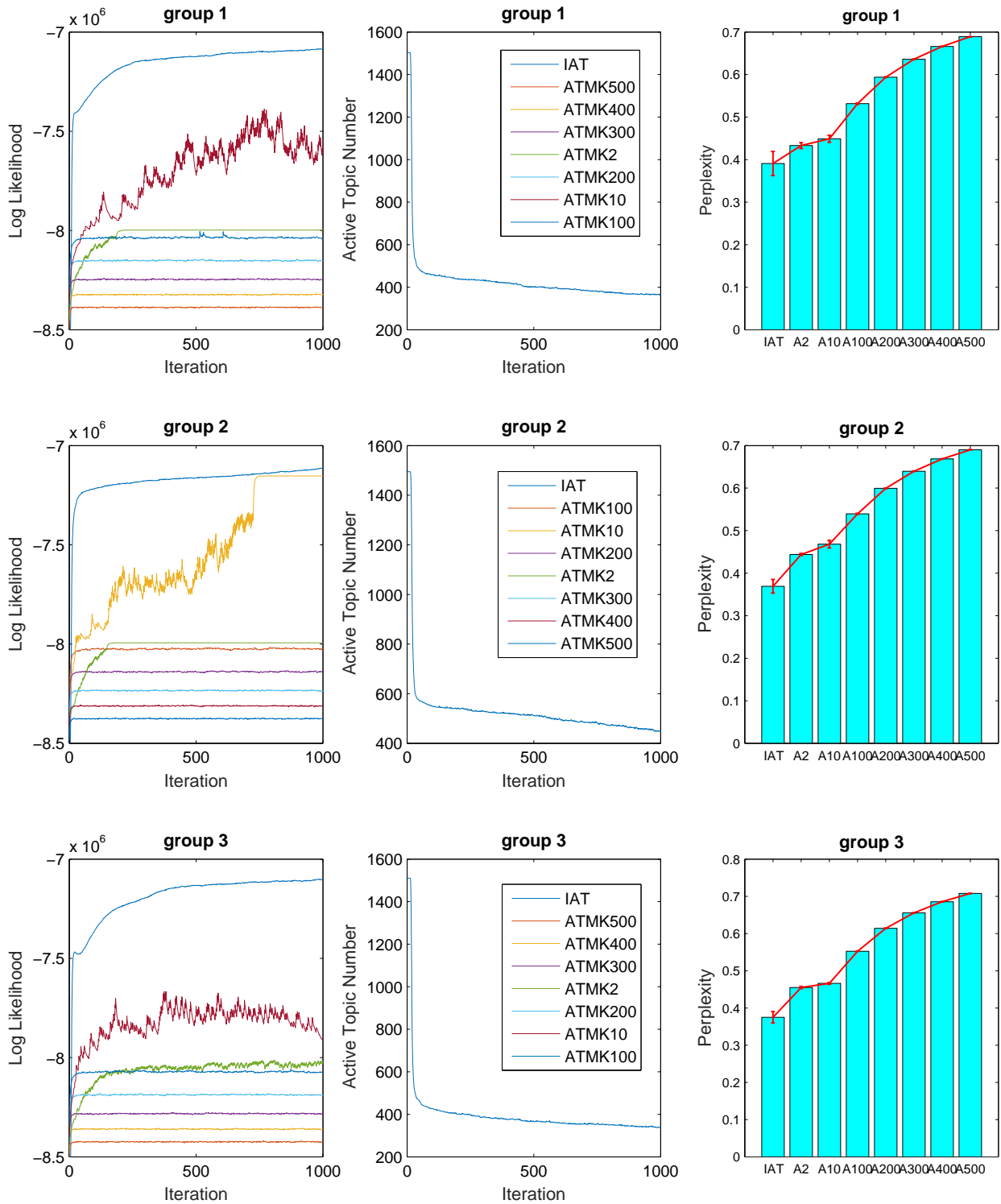
Figure 5: Results from IAT and ATM on three groups of NIPS dataset. Each row denotes a group. In each row, the left subfigure shows the Log-likelihoods comparison between IAT and ATM with different (predefined) topic numbers: $K = 2$, $K = 10$, $K = 100$, $K = 200$, $K = 300$, $K = 400$, and $K = 500$; The middle subfigure shows the change of active topic number of IAT during the iteration of Gibbs sampling; the right subfigure shows the perplexity comparison between IAT and ATMs.

butions. The model uses a mixed author gamma process as the base measure of the document gamma process to capture the author-document mapping. We have demonstrated that the designed Gibbs sampling algorithm can be used to learn such infinite author topic model based on the various real-world datasets.

Other potential applications of this work include multi-label learning [25]: The 'authors' in the proposed model can be seen as labels, and the inference of the model can be seen as the training of the multi-label classifier. The learned topics can be seen as having infinite features space. This is our further study.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1658–1670, Sept. 2012.

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

[3] T. Broderick, L. Mackey, J. Paisley, and M. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2014.

[4] W. L. Buntine and S. Mishra. Experiments with non-parametric topic models. KDD '14, pages 881–890, New York, NY, USA, 2014. ACM.

[5] X. Chen, X. Hu, T. Y. Lim, X. Shen, E. K. Park, and G. L. Rosen. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):980–991, July 2012.

[6] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. KDD '11, pages 1271–1279, New York, NY, USA, 2011. ACM.

[7] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[8] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. KDD '13, pages 266–274, New York, NY, USA, 2013. ACM.

[9] H. Kim, Y. Sun, J. Hockenmaier, and J. Han. Etm: Entity topic models for mining documents associated with entities. ICDM '12, pages 349–358, Washington, DC, USA, 2012. IEEE Computer Society.

[10] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. WWW '08, pages 1009–1018, New York, NY, USA, 2008. ACM.

[11] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of*

[12] C.-T. Nguyen, N. Kaothanthong, T. Tokuyama, and X.-H. Phan. A feature-word-topic model for image annotation and retrieval. *ACM Transactions on Web*, 7(3):12:1–12:24, Sept. 2013.

[13] A. Pinto and G. Haus. A novel xml music information retrieval method using graph invariants. *ACM Transactions on Information Systems*, 25(4), Oct. 2007.

[14] M. H. Quenouille. A relation between the logarithmic, poisson, and negative binomial series. *Biometrics*, 5(2):162–164, 1949.

[15] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[16] C. E. Rasmussen. The infinite gaussian mixture model. NIPS '12, pages 554–560, 1999.

[17] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):4:1–4:38, Jan. 2010.

[18] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[19] A. Roychowdhury and B. Kulis. Gamma processes,stick-breaking, and variational inference. *arXiv preprint arXiv:1410.1068*, 2014.

[20] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. KDD '04, pages 306–315, New York, NY, USA, 2004. ACM.

[21] J. Tang, R. Jin, and J. Zhang. A topic modeling approach and its integration into the random walk framework for academic search. ICDM '08, pages 1055–1060, Dec 2008.

[22] Y. W. Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2010.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[24] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[25] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, Aug 2014.

[26] J. Zhao, P. Wang, and K. Huang. A semi-supervised approach for author disambiguation in kdd cup 2013. KDD Cup '13, pages 10:1–10:8, New York, NY, USA, 2013. ACM.

[27] M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2013.

[28] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(1):2237–2278, Aug. 2012.

[ ] *Computational and Graphical Statistics*, 9(2):249–265, 2000.