This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

Bayesian Nonparametric Relational Topic Model through Dependent Gamma Processes

Junyu Xuan, Jie Lu, *Senior Member, IEEE*, Guangquan Zhang, Richard Yi Da Xu, and Xiangfeng Luo, *Member, IEEE*

Abstract—Traditional relational topic models provide a successful way to discover the hidden topics from a document network. Many theoretical and practical tasks, such as dimensional reduction, document clustering, and link prediction, could benefit from this revealed knowledge. However, existing relational topic models are based on an assumption that the number of hidden topics is known a priori, which is impractical in many real-world applications. Therefore, in order to relax this assumption, we propose a nonparametric relational topic model using stochastic processes instead of fixed-dimensional probability distributions in this paper. Specifically, each document is assigned a Gamma process, which represents the topic interest of this document. Although this method provides an elegant solution, it brings additional challenges when mathematically modeling the inherent network structure of typical document network, i.e., two spatially closer documents tend to have more similar topics. Furthermore, we require that the topics are shared by all the documents. In order to resolve these challenges, we use a subsampling strategy to assign each document a different Gamma process from the global Gamma process, and the subsampling probabilities of documents are assigned with a Markov Random Field constraint that inherits the document network structure. Through the designed posterior inference algorithm, we can discover the hidden topics and its number simultaneously. Experimental results on both synthetic and real-world network datasets demonstrate the capabilities of learning the hidden topics and, more importantly, the number of topics.

Index Terms—Text mining, network analysis, topic model, Bayesian nonparametric

1 Introduction

Understanding a corpus is significant for businesses, organizations and individuals for instance the academic papers of IEEE, the emails in an organization and the previously browsed webpages of a person. One commonly accepted and successful way to understand a corpus is to discover the hidden topics in the corpus [1]–[3]. The revealed hidden topics could improve the services of IEEE, such as the ability to search, browse or visualize academic papers; help an organization understand and resolve the concerns of its employees; assist internet browsers to understand the interests of a person and then provide accurate personalized services. Furthermore, there are normally links between the documents in a corpus. A paper citation network [4] is an example of a document network in which the academic papers are linked by their citation relations; an email network [5] is a document network in which the emails are linked by their reply relations; a webpage network [6], [7] is a document network in which webpages are linked by their hyperlinks. Since these links also express the nature of the documents, it is apparent that hidden topic discovery should consider these links as well.

Similar studies focusing on the hidden topics discovering from the document network using some Relational Topic Models (RTM) [8]–[10] have already been successfully developed. Unlike the traditional topic models [1], [2] that focus on mining the hidden topics from a document corpus (without links between documents), the RTM can make discovered topics inherit the document network structure. The links between documents can be considered as constrains of the hidden topics.

One drawback of existing RTMs is that they are built with fixed-dimensional probability distributions, such as Dirichlet, Multinomial, Gamma and Possion distribution, which require their dimensions be fixed before use. Hence, the number of hidden topics must be specified in advance, and is normally chosen using domain knowledge. This is difficult and unrealistic in many real-world applications, so RTMs fail to find the number of topics in a document network.

In order to overcome this drawback, we propose a Nonparametric Relational Topic (NRT) model in this paper, which removes the necessity of fixing the topic number. When aiming to build a NRT for a document network, there are three challenges: 1) How to express the document interest on infinite number of topics? In-

- *J. Xuan is with Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Australia and the School of Computer Engineering and Science, Shanghai University, China (e-mail: Junyu.Xuan@uts.edu.au).*
- *J. Lu, G. Zhang and R. Y. D. Xu are with Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Australia (e-mail: Jie.Lu@uts.edu.au; Guangquan.Zhang@uts.edu.au; Yida.Xu@uts.edu.au).*
- *X. Luo is with the School of Computer Engineering and Science, Shanghai University, China. (e-mail: luoxf@shu.edu.cn).*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

stead of probability distributions, stochastic processes are adopted by the proposed model to express the interest of a document on the 'infinite' number of topics. Stochastic process can be simply considered as 'infinite' dimensional distributions[1]. 2) How to make all the documents share the same set of topics? This is a common feature found in many real-world applications, and many literatures [9], [10] have exploited this property in their work. In order to achieve the above requirement, we use a global Gamma process to represent a set of base components each document has its own Gamma process thinned from the global one. The thinned Gamma processes help documents share the same set of topics. This is important because users are not interested in analyzing documents in a database without sharing any common topics. 3) How to make two linked documents have similar topics? We handle this challenge by controlling the subsampling probabilities of all the documents on topics, and make the linked documents subsample the similar topics. A subsampling Markov Random Field is proposed as the model constraint. Finally, two sampling algorithms are designed to learn the proposed model under different conditions. Experiments with document networks show some efficiency in learning what the hidden topics are and superior performance the model's ability to learn the number of hidden topics. It is worth noting that, although we use document networks as examples throughout this paper, our work can be applied to other networks with node features.

The main contributions of this paper are to:

1) propose a new Bayesian nonparametric model which can relax the topic number assumption used in the traditional relational topic models;
2) design two sampling inference algorithms for the proposed model: a truncated version and an slice version to facilitate the inference for the proposed model.

The rest paper is structured as follows. Section 2 summarizes the related work. The proposed NRT model is presented in Section 3 and we have illustrated the detailed derivations of its sampling inference in Section 4. Section 5 presents experimental results both on the synthetic and real-world data. 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 paper. The first part summarizes the literature on relational topic models. The second part summarizes the literatures on Bayesian nonparametric learning.

1. We only consider the pure-jump processes in this paper. Some continuous processes cannot be simply considered as the 'infinite' dimensional distributions.

2.1 Topic models with network

Our work in this paper aims to model the data with the network structure as a constraint. Since social network and citation network are two explicit and commonly-used networks in the data mining and machine learning areas, some extensions of the traditional topic models try to adapt to these networks. The co-occurrence relations between words are considered by a Graph Topic Model [11]. For the social network, an Author-Recipient-Topic model [12] was proposed to analyze the categories of roles in social networks based on the relationships of people in the network. A similar task was investigated in [13] where social network structure was inferred from informal chatroom conversations utilizing the topic model [14]. As an important issue of social network analysis, communities [15] were extracted using a Social Topic Model [16]. The Mixed Membership Stochastic Blockmodel is another way to learn the mixed membership vector (i.e., topic distribution) for each node from a network structure [17], but it did not consider the content/features of each node. For the citation network, Relational Topic Model (RTM) was proposed to infer the topics [9] and discriminative topics [10] from citation networks by introducing a link variable between two linked documents. Unlike RTM, a block was adopted to model the link between two document [18]. Considering the physical meaning of citation relations, a variable was introduced to indicate if the content of citing paper was inherited from cited paper or not [19]. In order to keep the document structure, Markov Random Field (MRF) was combined with topic model [20]. The communities in citation network were also investigated [21]. In summary, existing relational topic models are all inherited from traditional topic models, so the number of topics needs to be fixed. It is unrealistic, in many real-world situations, to fix this number in advance. Our work tries to resolve this issue through the nonparametric learning techniques reviewed in the following subsection.

Note that there is another similar research field that is very similar but different with relational topic model. A relational data is composed of two parts: a network structure (e.g., document network) and node features (e.g., document-word mapping). Relational topic model (RTM) is a kind of model to discover topics from node features constrained by the (node) network structure; contrarily, some works try to detect node community constrained by the node features. For example, GAMer is a combination of subspace learning and dense graph mining [22]; a simple probabilistic generative model is built for the friend circles in a social network [23]; two sources of information are linked through the node community membership [22] and using seed groups as lower bounds of communities [24]. It is interesting that although they are working on the same data, their aim and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

output are totally different, so we want to consider them as two different research fields. Furthermore, we will compare the proposed model with one model (CESNA model [22]) from this field in the experiment section.

2.2 Bayesian model selection and Bayesian nonparametric learning

Aforementioned Bayesian models need to select an appropriate number of topics, i.e., model selection problem. There are mainly two kinds of Bayesian model selection approaches: separate estimation and comparative estimation. For separate estimation, two models are compared through their posterior distribution given data, such as: Bayes Factor (BF) [25], An Information theoretic Criterion (AIC) [26], Bayesian Information Criterion (BIC) [27], and Deviance Information Criterion (DIC) [28], and so on. For comparative estimation, the distance between two posterior distributions from two models is evaluated through KL divergence [29] or entropy [30]. Although these methods achieve success, they require multiple runs of the learning algorithm with different topic numbers which limits the practicality of these approaches.

Bayesian nonparametric learning [31] is another principle way to learn the number of mixtures in a mixture model. Without predefining the number of mixtures, this number is supposed to be inferred from the data, i.e., let the data speak. The traditional elements of probabilistic models are fixed-dimensional distributions, such as Gaussian distribution, Dirichlet distribution [1], Logistic Normal distribution [32], and so on. All these distributions need to predefine their dimensions. In order to avoid this, Gaussian process and Dirichlet process [33] are used to replace former fixed-dimensional distributions because of their infinite properties. Since the data is limited, the learned/used atoms will also be limited even with these 'infinite' stochastic processes. Infinite mixture models are the extension of Finite Mixture Models through the 'infinite' stochastic processes where there are a finite number of hidden components (topics) used to generate data. One classic infinite mixture model is the Infinite Gaussian mixture model [34]. An example use for a Dirichlet process is the hierarchical topic model composed by Latent Dirichlet Allocation (LDA) [1] with a nested Chinese restaurant process [35]. By using a nested Chinese restaurant process as the prior, not only is the number of them not fixed, the topics in this model are also hierarchically organized. In order to learn the stochastic processes-based models with an infinite property, the inference methods should be properly designed. There are two popular and successful methods to do this: Markov Chain Monte Carlo (MCMC) [36] and variational inference [37]. To summarize, nonparametric learning has been successfully used for extending many models and applied in many real-world applications. However, there is still no work on the nonparametric extension of relational topic models. This paper uses a set of Gamma processes to extend the finite relational topic model to the infinite one.

TABLE 1
Important Notations in this paper

Symbol	meaning in this paper
D	the number of documents
V	vocabulary size
K	the number of topics
d	document index
v	word index
k	topic index
N_d	number of words in document d
H	the base probability measure of a Gamma process
α	concentration parameter of a Gamma process
Γ	a random draw/realization of a Gamma process
π_k	weight of the topic k
$\pi_{d,k}$	weight of the topic k in document d
Θ	parameter (keyword distribution) space
θ_k	parameter (keyword distribution) of topic k
r_k	binary (indicator) variable of topic k
$r_{d,k}$	binary (indicator) variable of topic k in document d
$q_{d,k}$	the subsampling probability of document d keeping topic k
C	a clique of document network
$\psi(C)$	energy function on clique C
$n_{d,v}$	number of word v in document d
$n_{d,v,k}$	number of word v assigned to topic k in document d
$z_{d,v,m}$	topic index assigned to m-th word v in document d
$u_{d,v,m}$	auxiliary slice variable assigned to m-th word v in document d

3 Nonparametric Relational Topic Model

In this section, we present the proposed Nonparametric Relational Topic (NRT) model for the document network in detail. When aiming to build a NRT, we are going to face three challenges: i) How to express the document interest on infinite number of topics? ii) How to make all the documents share the same set of topics? iii) How to make two linked documents share similar topics? In the following, we will introduce our idea to handle the above three challenges one by one. Some frequently used notations are summarized in Table 1.

Challenge 1 *How to express the document interest on infinite number of topics?*

When the topic number is prefixed, it is simply to draw a random variable from a fixed-dimensional probability distribution (such as Dirichlet distribution and Logit-normal distribution) as the topic interest of a document in the traditional topic models. However, the probability distributions have to be abandoned for the model building when the number of topics cannot be reasonably prefixed with enough prior knowledge. It makes traditional topic models built by probability distributions not work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

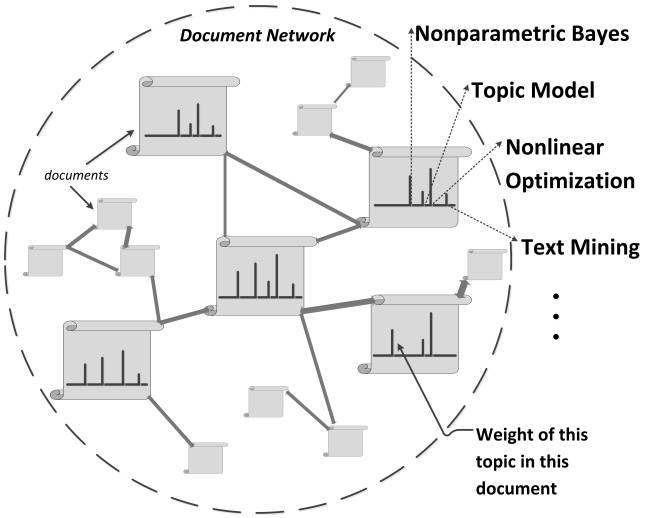


Fig. 1. Illustration of Gamma process assignments for a document network. Each document is assigned a Gamma process which has infinite components (represented by the fences in the figure). Each fence denotes a hidden topic, and some examples are given in the figure. The length of the fences denote the weights of different topics in a document.

Considering the observation, i.e., word counts in documents, we use Poisson process to model the observation, and then we use a draw from a Gamma process to express the interest of a document on infinite hidden topics due to the conjugacy between Gamma and Poisson processes. A Gamma process $GaP(\alpha, H)$ is a stochastic process, where H is a base (shape) measure parameter on topic space Θ and α is the concentration (scale) parameter.

Let $\Gamma = \{(\pi_k, \theta_k)\}_{k=1}^{\infty}$ be a random draw of a Gamma process in the product space $\mathbb{R}^+ \times \Theta$ where $\pi_k \in \mathbb{R}^+$ and $\theta_k \in \Theta$, and it can be represented as $\Gamma = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad \theta_k \sim H$, where δ_{θ_k} is an indicator function (i.e., $\delta_{\theta_k}(\theta^*) = 1$ if $\theta_k = \theta^*$ and $\delta_{\theta_k}(\theta^*) = 0$ if $\theta_k \neq \theta^*$); π_k satisfies an improper Gamma distribution $Gamma(0, \alpha)$ and that is why it is called *Gamma process*. Γ can also be seen as a complete random measure. More information about Gamma process can be found in [38], [39]. When using Γ to express the document interest, the $\{\theta_k)\}_{k=1}^{\infty}$ denotes the infinite number of topics and $\{\pi_k)\}_{k=1}^{\infty}$ denotes the weights of infinite number of topics in a document. Note that π_k is within $(0, +\infty)$ not $[0, 1]$, but $\{\pi_k)\}_{k=1}^{\infty}$ can also be seen as the weights of topics in a document. As illustrated in Fig. 1, our idea is to assign each document a Gamma process. In this figure, each document is with a '*fence*' in which each bar has two properties: *position* that denotes the topic and *length* that denotes the weight of the corresponding topic in this document. Note that each document could set its fence positions at will. Due to the infinity of Γ, we can handle Challenge 1 for now.

Challenge 2 *How to make all the documents share the same set of topics?*

Since we consider the situation with infinite number of topics, it hopes that there are some topics that are shared by documents even with infinite number of candidate topics. Let us consider an extreme situation: each document in a document network is with and only with its own topics that are different from others. Apparently, this situation is not what we want because the motivation of the document modeling or topic models is to discover the shared knowledge (i.e. topics) of a document corpus.

Considering the continuity of the parameter space Θ (the *base line* of the *fence* in Fig. 1, equivalently), the probability that two documents are with same topics is 0. In order to handel Challenge 2, we firstly generate a global Gamma process, i.e., $\Gamma_0 \sim GaP(\alpha, H)$, which is equal to $\Gamma_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad \theta_k \sim H$, where $\{\pi_k, \theta_k\}_{k=1}^{\infty}$ is the global set of topics. Our idea is to consider $\{\pi_k, \theta_k\}_{k=1}^{\infty}$ as a global topic pool, and each document just selects its own topics from this pool. In this way, the probability of sharing topics between different documents will not be 0. We use a thinned Gamma process to realize this idea. Its definition is as follow,

Definition 1 (Thinned Gamma Process [40])**.** *Suppose we have countably infinite points $\{(\pi_k, \theta_k)\}_{k=1}^{\infty}$ from a Gamma process $\Gamma \sim GaP(\alpha, H)$. Then, we generate a set of independent binary variables $\{r_k\}_{k=1}^{\infty}$ ($r_k \in \{0, 1\}$). The new process,*

$$\Gamma' = \sum_{k=1}^{\infty} \pi_k r_k \delta_{\theta_k} \tag{1}$$

is still a Gamma process, which is proofed by [40]. The $\{r_k\}$ can be seen as the indicators for the reservation of the point of original/global Gamma process, so Γ' is called Thinned Gamma Process.

We can give each r_k a Bernoulli prior $p(r_k = 1) = q_k$, where $q_k \in [0, 1]$ is the subsampling probability of keeping topic k. Apparently, different realizations of $\{r_k\}$ will lead to different thinned Gamma processes. For each document, a thinned Gamma process Γ_d is generated with Γ_0 as the global process,

$$\Gamma_d = \sum_{k=1}^{\infty} \pi_k r_{d,k} \delta_{\theta_k} \tag{2}$$

where $\{r_{d,k}\}_{k=1}^{\infty}$ is a set of indicators of document d on the corresponding components. These $\{r_{d,k}\}_{k=1}^{\infty}$ are independent identical distributed random variables with Bernoulli distributions,

$$r_{d,k} \sim Bernoulli(q_{d,k}) \tag{3}$$

where $q_{d,k}$ denotes the probability of the Gamma process Γ_d of document d with component k. Until now, Challenge 2 is handled.

rectangle

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

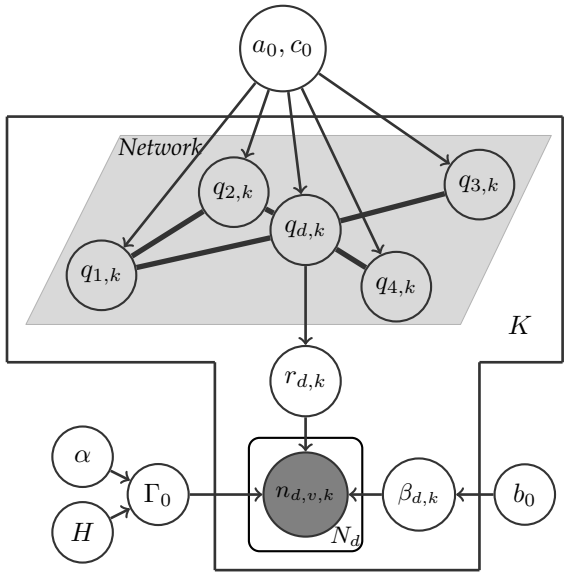


Fig. 2. Graphical representation for the Nonparametric Relational Topic (NRT) Model.

Challenge 3 *How to make two linked documents have similar topics?*

Two linked documents in a document network normally have similar topics. For example, the linked two academic papers through a citation are normally with some common researches, and two linked webpages through a hyperlink normally report similar news. Therefore, we need to control the sharing strategy of documents on the infinite topics in order to make two linked documents have similar topics.

Since we have assigned each document a thinned Gamma process, our idea is to make thinned Gamma processes dependent with each other according to the document network structure. Each thinned Gamma process is subsampled from the global Gamma process Γ_0 according to indicators $\{r_k\}$. Therefore, the dependence between the different realizations of $\{r_k\}$ will also lead to dependence of the thinned Gamma processes.

In order to obtain the dependent $\{r_k\}$ between documents, we define a Subsampling Markov Random Field (MRF) to constrain the q_k^d of all documents,

Definition 2. *[Subsampling Markov Random Field] The subsampling probabilities of all the documents on a component/topic in the global Gamma process have the following constraint,*

$$p(\boldsymbol{q_k}) = \prod_{C \in \wp(Network)} \psi(C) = \frac{1}{Z(\boldsymbol{q_k})} \exp\left(- \sum_{<d_i,d_j> \in C} \|q_{d_i,k} - q_{d_j,k}\|^2 \right) \quad (4)$$

where $\boldsymbol{q_k} = \{q_{d,k}\}_{d=1}^D$; $Network$ is the document network; $\wp(Network)$ is the clique set of $Network$; C is one clique; $\psi(C)$ is the energy function of MRF; $< d_i, d_j > \in C$ denotes there is link between d_i and d_j and this link is within clique C; and $Z(\boldsymbol{q_k})$ is the normalization part and also called partition function.

Note that the energy function $\exp\left(-\sum_{<d_i,d_j>\in C} \|q_{d_i,k} - q_{d_j,k}\|^2\right)$ in Definition 2 is designed to constrain the distance between sub-sampling probabilities of different documents on topic k. The more closely two documents are posited in the network, the more close their sub-sampling probabilities on topic k. Through this subsampling MRF constraint, the marginal distribution of each subsampling probability will depend on the values of its neighbors. Therefore, the sub-sampling probabilities of linked documents will be similar, which ensures the Challenge 3 is handled.

To sum up, the proposed Nonparametric Relational Topic (NRT) Model is graphically illustrated in Fig. 2 and its generative procedure is,

$$\begin{aligned}
&\Gamma_0 \sim GaP(\alpha, H) \\
or \quad &\Gamma_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \\
&p(q_{d,k}) \propto Beta(q_{d,k}; a_0, c_0) \\
&\qquad \cdot \exp\left(- \sum_{\substack{<d, d_i> \in C \\ C \in \wp(Network)}} \|q_{d,k} - q_{d_i,k}\|^2 \right) \\
&r_{d,k} \sim Bernoulli(q_{d,k}) \\
&\Gamma_d = \sum_{k=1}^{\infty} r_{d,k} \pi_k \delta_{\theta_k}
\end{aligned}$$

With the $\{\Gamma_d\}_{d=1}^D$ for all the documents in hand, the generative procedure of the documents is as follow,

$$\begin{aligned}
\beta_{d,k} &\sim Gamma(b_0, 1) \\
n_{d,v,k} | n_{d,v} &\sim Poisson(\theta_{k,v} r_{d,k} \pi_k \beta_{d,k}) \\
v &\in [1, V] \\
n_{d,v} = \sum_{k=1}^{\infty} n_{d,v,k} &\sim Poisson(\sum_{k=1}^{\infty} \theta_{k,v} r_{d,k} \pi_k \beta_{d,k}) \\
\theta_k &\sim H
\end{aligned}$$

where $n_{d,v}$ is the number of word v in document d (same word may appear several times in a document), $n_{d,v,k}$ is the number of word v in document d assigned to topic k, and $\beta_{d,k}$ is a parameter. Considering the relationship between the Poisson distribution and the Multinomial distribution, the likelihood part is equal to,

$$\begin{aligned}
z_{d,v,m} &\sim Discrete\left(\frac{\theta_{k,v} r_{d,k} \pi_k \beta_{d,k}}{\sum_k \theta_{k,v} r_{d,k} \pi_k \beta_{d,k}} \right) \\
m &\in [1, n_{d,v}] \\
n_{d,v,k} &= \sum_m \delta_k(z_{d,v,m})
\end{aligned}$$

This form is more convenient for the slice sampling design for the model which will be explained in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

following Section. a_0, b_0, c_0, α are model parameters. H is a base measure for the global Gamma process, and it is set as a Dirichlet distribution parameterized by η. Note that the $\boldsymbol{q}$ are not only with Beta distribution prior but also with a MRF constraint at the same time.

4 Model Inference

The inference of the proposed (NRT) model is to compute the posterior distribution of latent variables given data (i.e., document network),

$$p(K, \pi, q, r, \theta, \beta | \{n_{d,v}\}_{d\in[1,D], v\in[1,V]}, Network)$$

It is apparently that this posterior distribution is a high-dimensional and multi-variable distribution which analytical form is extremely hard to obtain. Therefore, we first use Gibbs sampling method to get samples of this posterior distribution with a truncation (define a relatively large topic number), which is a commonly-adopted strategy in the Bayesian nonparametric learning area in Section 4.1. Furthermore, we also develop an exact sampling method without the truncation requirement based on slice sampling technique [41] in Section 4.2.

4.1 Gibbs Sampling

It is difficult to perform posterior inference under infinite mixtures, and a common work-around solution in Bayesian nonparametric learning is to use a truncation method. This method is widely accepted, which uses a relatively big $K^\dagger$ as the (potential) maximum number of topics. As required by the Gibbs sampling framework, we list all the conditional distributions for the latent variables of the model in the following.

Sampling $q_{d,k}$. Since there are additional constraints for the sub-sampling probabilities, they do not have a closed-formed posterior distribution.

If $r_{d,k} = 1$,

$$p(q_{d,k}|\cdots) \propto q_{d,k}^{a_0+1-1}(1-q_{d,k})^{c_0-1} \cdot \exp\left(-\sum_{\substack{<d,d_i>\in C \\ C\in\wp(Network)}} \|q_{d,k} - q_{d_i,k}\|^2\right) \quad (5)$$

If $r_{d,k} = 0$,

$$p(q_{d,k}|\cdots) \propto q_{d,k}^{a_0-1}(1-q_{d,k})^{c_0+1-1} \cdot \exp\left(-\sum_{\substack{<d,d_i>\in C \\ C\in\wp(Network)}} \|q_{d,k} - q_{d_i,k}\|^2\right) \quad (6)$$

Given this conditional distribution of $q_{d,k}$, we can use the efficient *A* sampling* [42] that is developed recently, because the conditional distribution can be decomposed into two parts: $q_{d,k}^{a_0-1}(1-q_{d,k})^{c_0+1-1}$ and exponential part. The first part is easily sampled using a beta distribution (proposal distribution), and the second part is a bounded function.

Sampling $r_{d,k}$.

1) $\forall j, r_{d,j} = 0 \rightarrow r_{d,k} = 1$
2) $\exists v, n_{d,v,k} > 0 \rightarrow r_{d,k} = 1$
3) $\forall v, n_{d,v,k} = 0$
   a) if $\forall v, u_{d,v,k} = 0$,

$$p(r_{d,k} = 1) \propto q_{d,k} \prod_n Poi(0; \theta_{k,v}\pi_k\beta_{d,k}) \quad (7)$$

   b) if $\forall v, u_{d,v,k} = 0$,

$$p^{(1)}(r_{d,k} = 0) \propto (1-q_{d,k}) \prod_n Poi(0; \theta_{k,v}\pi_k\beta_{d,k}) \quad (8)$$

   c) if $\exists v, u_{d,v,k} > 0$,

$$p^{(2)}(r_{d,k} = 0) \propto (1-q_{d,k}) \cdot \left(1 - \prod_v Poi\left(0; \theta_{k,v}\pi_k\beta_{d,k}\right)\right) \quad (9)$$

Accordingly, we can use a discrete distribution to sample r by,

$$p(r_{d,k} = 1|\cdots) \propto \frac{p(r_{d,k}=1)}{p(r_{d,k}=1) + p^{(1)}(r_{d,k}=0) + p^{(2)}(r_{d,k}=0)} \quad (10)$$

Sampling $\beta_{d,k}$. $\beta_{d,k}$ is a model parameter with a Gamma prior and due to the conjugate between the Gamma and Poisson distribution, we have

$$p(\beta_{d,k}|\cdots) \propto Gamma(n_{d,\cdot,k} + b_0, \frac{1}{r_{d,k}\pi_k + 1}) \quad (11)$$

where $n_{d,\cdot,k} = \sum_v n_{d,v,k}$ is the number of words assigned to topic k in document d.

Sampling θ_k. In our model, we set H as a probability (Dirichlet) distribution parameterized by η, so we have the following posterior

$$p(\theta_k|\cdots) \propto Dir(\eta + n_{\cdot,1,k}, \ldots, \eta + n_{\cdot,V,k}) \quad (12)$$

where $n_{\cdot,v,k} = \sum_d n_{d,v,k}$ is the number of word v assigned to topic k in all the documents.

Sampling $n_{d,v,k}$. (truncated version) Here, we need to sample the $n_{d,v,1}, ..., n_{d,v,K^\dagger}$ together due to the known $n_{d,v} = \sum_{k=1}^{K^\dagger} n_{d,v,k}$ according to Multinomial distribution

$$p(n_{d,v,1}, \ldots, n_{d,v,K^\dagger}|\cdots) \propto Mult(n_{d,v}; \xi_{d,v,1}, \ldots, \xi_{d,v,K^\dagger}) \quad (13)$$

where

$$\xi_{d,v,k} = \frac{\theta_{k,v} r_{d,k}\pi_k\beta_{d,k}}{\sum_k^{K^\dagger} \theta_{k,v} r_{d,k}\pi_k\beta_{d,k}}. \quad (14)$$

Sampling π_k. (truncated version) Although is from a Gamma process, it can be seen with a Gamma distribution prior given a truncation level $K^\dagger$, so we can sample it through the following posterior,

$$p(\pi_k|\cdots) \propto Gamma(1/K^\dagger + n_{\cdot,\cdot,k}, \frac{1}{\beta_{\cdot,k} + 1}) \quad (15)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

Algorithm 1: Truncated Version of Gibbs Sampling for NRT

Input: $Network$ and $\boldsymbol{n_{d,v}}$
Output: K, $\{\theta_k\}_{k=1}^K$, $\{\pi_k^d\}_{k=1}^K$

randomly set initial values for K, $\{\theta_k\}_{k=1}^K$, $\{\pi_k\}_{k=1}^K$
it = 1;
while $it \leq max_{it}$ **do**
  for each topic k **do**
    for each document d **do**
      for each word v of document d **do**
        Update $n_{d,v,k}$ by Eq. (13) ;
      end for
      Update $q_{d,k}$ by Eq. (5) and (6) ;
      Update $r_{d,k}$ by Eq. (10) ;
      Update $\beta_{d,k}$ by Eq. (11) ;
    end for
    Update θ_k by Eq. (12) ;
    Update π_k by Eq. (15);
  end for
  it++;
end while

where $n_{\cdot,\cdot,k} = \sum_d \sum_v n_{d,v,k}$ is the total number of words assign to topic k and $\beta_{\cdot,k} = \sum_d \beta_{d,k}$. Note that the truncation version of the model is not equal to a probability distribution-based model [43]. Under this truncation, there will be only limited number of topics used by documents and large number of remaining topics will be unused. This truncation can be seen as an approximation of the NRT.

The whole sampling algorithm is summarized in Algorithm 1. It is interesting that the sub-sampling probabilities of different documents are independent of each other given other variables. So the update of sub-sampling probabilities of different documents can be implemented in a parallel fashion.

Note that the truncation level $K^\dagger$ should not be simply considered as a model parameter like the topic number in traditional topic model. The topic number in traditional topic model should be carefully selected within its scope; contrarily, the setting of truncation level is quit easy, because it could be simply set as large as possible provided the computational resources could support. Therefore, truncation level could be seen as an improvement comparing with the topic number in traditional topic model.

4.2 Slice Sampling

Although the truncated method are commonly accepted in the literature, maintaining a large number of components and their parameters is time and space consuming. An elegant idea (named slice sampling [41]) to resolve this problem is to introducing additional variables to adaptively truncate/select the infinite components. The very essence of slice sampling is to design a distribution for a new variable to make the original distribution easy to sample.

Sampling $n_{d,v,k}$ (slice sampling version) In order to do slice sampling, we introduce the auxilary/slice variable as,

$$u_{d,v,m} = Uniform(0, \zeta_0), m \in [1, n_{d,v}] \tag{16}$$

where $Uniform(0, \zeta_0)$ is a Uniform distribution on $[0, \zeta_0]$ and ζ_k is a fixed positive decreasing sequence $lim_{k\to\infty}\zeta_k = 0$. With the help of slice variable $u_{d,v,m}$, we can sample $z_{d,v,m}$ within a finite scope as follows,

$$\begin{aligned} p(z_{d,v,m} = k|\cdots) &\propto \xi_{d,v,k} \cdot \frac{\Pi(u_{d,v,m} \leq \zeta_k)}{\zeta_k} \\ n_{d,v,k} &= \sum_m \delta_k(z_{d,v,m}) \\ m &\in [1, n_{d,v}] \end{aligned} \tag{17}$$

where $\Pi(u_{d,v,m} \leq \zeta_k) = 1$ when $u_{d,v,m} \leq \zeta_k$ is satisfied; $\Pi(u_{d,v,m} \leq \zeta_k) = 0$ when $u_{d,v,m} \leq \zeta_k$ is not satisfied. Note that the possible values of $z_{d,v,m}$ are limited by $\Pi(u_{d,v,m} \leq \zeta_k)$ because ζ_k is a fixed positive decreasing sequence.

Sampling π_k (slice sampling version) The construction of Gamma process ($\Gamma_0 \sim GaP(\alpha, H)$) [39] is,

$$\Gamma_0 = \sum_{k=1}^{\infty} E_k e^{-T_k} \delta_{\theta_k} \tag{18}$$

where E_k and T_k are two additional auxiliary variables

$$E_k \sim Exp(\frac{1}{\alpha}),\ T_k \sim Gamma(\kappa_k, \frac{1}{\alpha}),\ \theta_k \sim H \tag{19}$$

where $Exp(\frac{1}{\alpha})$ denotes an Exponential distribution parameterized by $\frac{1}{\alpha}$. According to [39], [44], all the components/points/topics could be considered as draws from a number (I that could be infinitely large) of Poisson processes, so each topic is assigned a Poisson process index κ_k and the following property holds,

$$\sum_{k=1}^{\infty} \delta_{\kappa_k}(i) \sim Poisson(\gamma), \quad \gamma = \int_{\Theta} H \tag{20}$$

which means that the number of topics from each Poisson process satisfies a Poisson distribution parameterized by γ that is the total mass of base measure H of Gamma Process. Note that γ is equal to 1 if the H is set as a probability measure. Finally, According to the construction in Eq. (18), the prior of π_k is,

$$\pi_k = E_k e^{-T_k} \sim Exp(\frac{1}{\alpha}) \cdot Gam(\kappa_k, \frac{1}{\alpha}) \tag{21}$$

and the posterior is,

$$\begin{aligned} \pi_k &= (E_k, T_k) \\ &\sim Poi(data|E_k e^{-T_k}) Exp(E_k|\frac{1}{\alpha}) Gam(T_k|\kappa_k, \frac{1}{\alpha}) \end{aligned} \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

Algorithm 2: Slice Version of Gibbs Sampling for NRT

Input: $Network$ and $\boldsymbol{n_{d,v}}$
Output: K, $\{\theta_k\}_{k=1}^K$, $\{\pi_k^d\}_{k=1}^K$

randomly set initial values for K, $\{\theta_k\}_{k=1}^K$, $\{\pi_k\}_{k=1}^K$
it = 1;
while $it \leq max_{it}$ **do**
  for each topic k **do**
    for each document d **do**
      for each word v of document d **do**
        Sample slice variable $n_{d,v,k}$ by Eq. (16) ;
        Update $n_{d,v,k}$ by Eq. (17) ;
      end for
      Update $q_{d,k}$ by Eq. (5) or (6) ;
      Update $r_{d,k}$ by Eq. (10) ;
      Update $\beta_{d,k}$ by Eq. (11) ;
    end for
    Update θ_k by Eq. (12) ;
    Update π_k by Eq. (23);
    Update κ_k by Eq. (24);
  end for
  it++;
end while

We can sample this posterior by two Gamma distributions,

$$\begin{aligned} E_k|T_k &\sim Gam(E_k|n_{\cdot,\cdot,k}+1, \frac{1}{\alpha^{-1}+\beta_{\cdot,k}e^{-T_k}}) \\ T_k|E_k &\sim Poi(n_{\cdot,\cdot,k}|\beta_{\cdot,k}e^{-T_k}E_k)\cdot Gam(T_k|\kappa_k, \frac{1}{\alpha}) \end{aligned} \tag{23}$$

where $n_{\cdot,\cdot,k} = \sum_d \sum_v n_{d,v,k}$ and $\beta_{\cdot,k} = \sum_d \beta_{d,k}$.

The conditional distribution for the indicator κ_k is,

$$p(\kappa_k = i|\cdots) \propto p(T_k|\kappa_k = i)\cdot p(\kappa_k = i|\{\kappa_l\}_{l=1}^{k-1}) \tag{24}$$

The second part on the right hand side of Eq. (24) is,

$$p(\kappa_k = i|\cdots) = \begin{cases} 0, & \text{if } i < \kappa_{k-1} \\ \dfrac{1-F(I_{i-1}|\gamma)}{1-F(I_{i-1}-1|\gamma)}, & \text{if } i = \kappa_{k-1} \\ \dfrac{(F(I_{i-1}|\gamma)-F(I_{i-1}-1|\gamma))}{1-F(I_{i-1}-1|\gamma)} \cdot (1-f(0|\gamma))\, f(0|\gamma)^{h-1}, & \text{if } i = \kappa_{k-1}+h \end{cases} \tag{25}$$

where h is an integer denotes the distance between κ_k with κ_{k-1}; I_i is the number of items in i-th Poisson process and $I_i \sim Poisson(\gamma)$; $F(\cdot|\gamma)$ and $f(\cdot|\gamma)$ are the cumulative distribution function and probability density function of Poisson distribution parameterized by γ.

Note that the $u_{d,v,m}$, κ_k, E_k and T_k are introduced additional variables. They are not in the original model, and their appearances are only for the sampling without the help of the truncation level. The whole slice sampling algorithm is summarized in Algorithm 2.

5 Experiments

In this section, we evaluate the effectiveness of the proposed NRT model in learning the hidden topics from document networks. First, we introduce two evaluation metrics for the quantification of the effectiveness and comparisons in Section 5.1. Then, a series of experiments on the synthetic datasets to testify the model's different aspects in Section 5.2. Finally, we show the usefulness of the proposed model through comparing other models on two real-world datasets in Section 5.3.

5.1 Evaluation Metrics

Since NRT builds on two parts of knowledge (i.e., the network structure and document content) from a document network data, we make predictions for one of them based on the other. Two evaluation metrics used in state-of-the-art relational topic models have been adopted here for the quantitative comparison [9], [10]: *LinkRank*, *WordRank*, and *AUC*. *LinkRank* is defined as the average rank of positive links of test documents with training documents (The lower *LinkRank* is better); *WordRank* is defined as the average rank of words of test documents (The lower *WordRank* is better); *AUC* is the area under ROC that is the curve to show the positive link prediction of test documents (The higher *AUC* is better). The exact definitions could be found in [9], [10]. Note that the false links and words are considered in these metrics.

5.2 Experiments on synthetic data

We generated synthetic data to explore the NRT's ability to learn the hidden topics and infer the number of hidden topics from the document network, and to show the impact of SMRF and model parameter.

5.2.1 Synthetic data generation

At first, we choose a set of ground truth numbers symbolised by K, D and V that refer to the number of topics, documents and (different) words, respectively. Then, K global topics are generated through a V-dimensional Dirichlet distribution parameterized by $\{\alpha_1, \ldots, \alpha_V\}$ where $\alpha_i = 1\ \forall i$. Next, we generate the document interests on these topics through a K-dimensional Dirichlet distribution parameterized by $\beta_1, \ldots, \beta_K\}\ \forall \beta_i = 1$. With topics and the document interests on these topics in hand, we can generate each document d as follows: 1) N_d is uniformly chosen to be a number between $\frac{N}{2}$ and N where N is set as the maximum number of words in a document; 2) Repeat the following operations N_d times: a topic index is drawn from the document's topic interest and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

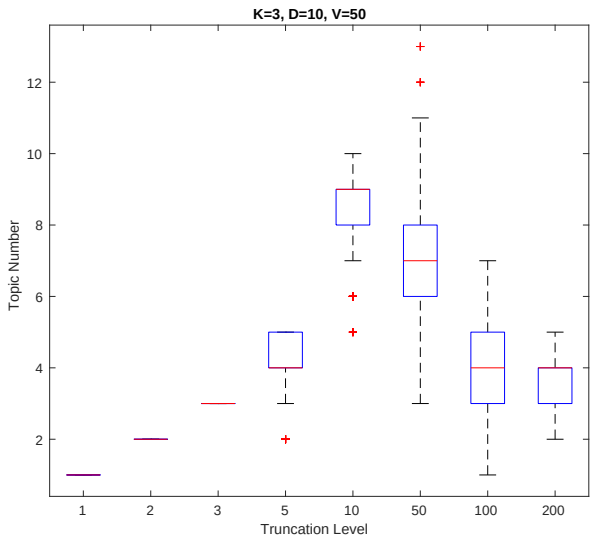


Fig. 3. The boxplot of the learned topic numbers by truncated inference method given different truncation levels.

then draw a word from the selected topic. Finally, we can obtain a $D \times V$ matrix with rows as documents and columns as words, and each entry of this matrix $n_{d,v}$ denotes the frequency a particular word v in a particular document d. The next step is to generate the relations between documents. For each pair of documents, we compute the inner product between their topic interests. In order to sparsify these relationships, we only retain the ones where their inner products are greater than 0.2.

5.2.2 Influence of truncation level

There are two inference methods proposed in this paper: one is truncation version and the other is slice version. For the truncation version in Algorithm 1, a truncation level needs to be given in advance. In order to show the influence of this parameter, we have fed different truncation levels (i.e., $K^{\dagger} \in \{1, 2, 3, 5, 10, 50, 100, 200\}$) and a generated dataset using the procedure in Section 5.2.1 with the setting (i.e., $K = 3$, $D = 10$ and $V = 50$) into Algorithm 1. For each run, it takes 10,000 iterations with 2,000 burn-in samples and 10 interval samples. The results are plotted in Fig. 3 which shows not only the topic number means from eight truncation levels but also the some basic statistics of 800 samples at each truncation level. It can be seen that the topic number dose not exceed the truncation levels when they are smaller than the real one (i.e., 3 for this dataset). When the truncation level is larger than 3, there will be a fluctuation of the learned topic numbers but the learned topic number will still not exceed the truncation level, so the fluctuation is small when the truncation level is not very large (such as 5 in the Fig. 3). As the increasing of the truncation level, the approximation of the truncation version distribution is more accurate, so the learned topic number will be closer to the real one and the variance is smaller.

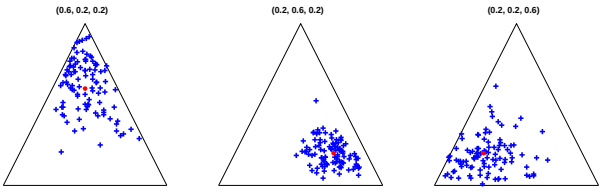


Fig. 4. The illustration of topics learning results. Three red/circle nodes denote three benchmark topics that are also given at the top of each subfigure; the blue/cross nodes denote learned topics from NRT.

5.2.3 Topics learning

One ability of NRT model is to discover the hidden topics from a document network. This subsection aims to show this ability. At first, we generate a synthetic dataset using the revised procedure in Section 5.2.1 with a setting (i.e., $K = 3$, $D = 10$, $V = 3$) and the topics are predefined as benchmarks rather than randomly sampled ones. The three topics are $(0.6, 0.2, 0.2)$, $(0.2, 0.6, 0.2)$, and $(0.2, 0.2, 0.6)$, which correspond to three points in the 3-dimensional simplex. After running NRT model (using truncation-based inference in Algorithm 1), we keep 100 samples with 3 topics. In each sample, there are three learned topics which are linked to the benchmark topics according to the similarity measure, and we choose the *best linking status* as the final one for each sample. The *best linking status* means the the total similarity between each pair of topics reaches maximum. For example, there are three learned topics in a sample: $(0.25, 0.5, 0.25)$, $(0.5, 0.25, 0.25)$, and $(0.25, 0.25, 0.5)$. We should link the first learned topic to $(0.2, 0.6, 0.2)$, the second learned topic to $(0.6, 0.2, 0.2)$, and the third learned topic to $(0.2, 0.2, 0.6)$. In Fig. 4, three red/circle nodes denote three benchmark topics and the blue/cross ones are from samples. We can see from this figure that the samples from NRT centers on the benchmark topics with a certain variance, which shows the effectiveness of NRT on the topics learning.

5.2.4 Topic number learning

Another ability of NRT model is to discover the hidden topics without the requirement of the predefined topic number. In order to show this ability, we use the synthetic data generation procedure in Section 5.2.1 with different settings: $K = 3$, $D = 10$, $V = 30$; $K = 12$, $D = 80$, $V = 100$; $K = 20$, $D = 500$, $V = 2000$; $K = 50$, $D = 3000$, $V = 2000$. For each setting, we run the NRT model (using truncation-based inference in Algorithm 1 for $K = 3, 12, 20$; using slice-based inference in Algorithm 2 for $K = 50$) with 10,000

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

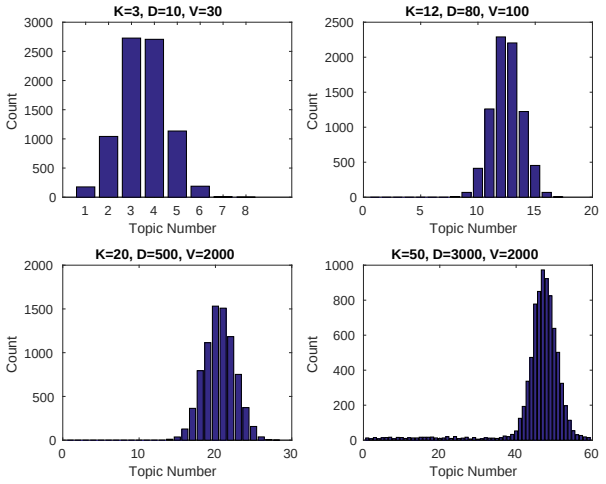


Fig. 5. Learned topic number distribution from NRT with synthetic datasets under different settings. Normally, the expectation of this distribution will be regarded as the learn topic number of a document network.

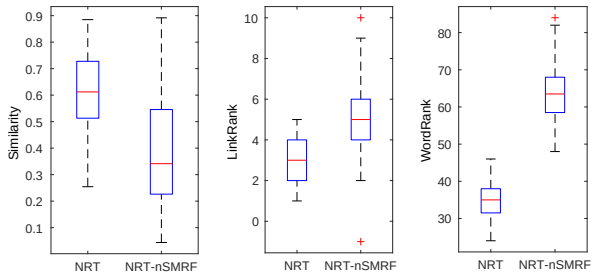


Fig. 6. Effectiveness of SMRF in NRT. The first subfigure is for the comparison between average similarity of topic interests of all test linked document pairs from both NRT and NRT without SMRF; The second and third subfigures are for comparison on *LinkRank* and *WordRank*.

iterations with first 2,000 samples as burn-in stage. In Fig. 5, we plot the topic numbers in the remaining 8,000 samples from NRT model on four synthetic datasets. From this figure, we can draw the conclusion that NRT has the ability to learn out the topic number from a document network to some extent.

5.2.5 Effectiveness of SMRF

We use SMRF that is proposed in Definition 2 to add the network structure into the model. In order to evaluate the performance of this SMRF, we compare *NRT with SMRF* and *NRT without SMRF* using generated dataset by Section 5.2.1 with setting: $K = 10$, $D = 30$, $V = 200$. Among all the documents, 23 documents are considered as the training documents with 44 links, and 7 documents are reserved as the test documents with 10 links. Here, we use the truncation-based inference in Algorithm 1. After the mixing of

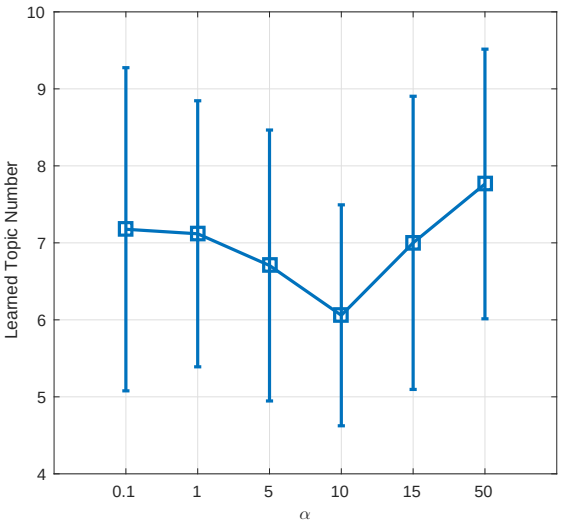


Fig. 7. Sensitivity of model parameter α on learned topic number from NRT. The errorbars in the figure show the standard deviations.

sampling (10,000 iterations with 2,000 burn-in samples), we take 100 samples with 80 as the interval. At first, we evaluate the average similarity between topic interests of all test linked document pairs. The assumption is that the more similar topic interests of two linked documents, the learned topics are more reasonable because the test links are generated using through the topic interest similarity. The result is plotted in the first subfigure of Fig. 6, which shows that NRT with SMRF constraint could recover the test links better. Next, we compare them using the metrics proposed in Section 5.1. The results are shown in the second and third subfigures of Fig. 6. We can see from these figures that SMRF helps NRT obtain better performance on the *LinkRank* and *WordRank*.

5.2.6 Sensitivity of model parameter α

α is the parameter of global Gamma process in NRT. Since there is a SMRF constraint in the model, it is difficult to theoretically deduce the distribution of the learned topic number. Therefore, we do this experiment to investigate the influence from α to the final learned topic number. We compare NRT with different values $\alpha = \{0.1, 1, 5, 10, 15, 50\}$ using generated dataset by Section 5.2.1 with setting: $K = 10$, $D = 30$, $V = 200$ using the truncation-based inference in Algorithm 1. Note that there are two sources that would affect the learned topic number: the model itself (model parameter α) and the data. In order to remove the effect from the data and focus on the investigation of the influence from the model parameter, the observation is ignored during the inference procedure. The results are shown in Fig. 7 from which we can draw the conclusion that α has little impact on the learned topic number. The reason is that α is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

concentration parameter of global Gamma process, so it has little impact on the topic number but the diversity of the topic interests of each document. The larger α is, the more diverse the weights of topics in each document.

5.3 Experiments on real-world data

5.3.1 Datasets and Setup

The real-world document network datasets[2] used in this study are:

- **Cora Dataset** It consists of 2,708 scientific publications with their citation relations. The citation network consists of 5,429 links. The dictionary consists of 1,433 unique words.
- **Citeseer Dataset** The CiteSeer dataset consists of 3,312 scientific publications. The citation network consists of 4,732 links. The dictionary consists of 3,703 unique words.
- **WebKB Dataset** The WebKB dataset consists of 877 webpages. The hyperlink network consists of 1,608 links. The dictionary consists of 1,703 unique words.

For each dataset, we use 5-fold cross validation to evaluate the performance. For each fold of dataset, the procedure is as follow: 1) train the model using the training data; 2) and compute two evaluation metrics on test data based on the trained model. The better model is expected to achieve better performance on the test data prediction. The average prediction results of 5-fold will be reported and plotted in the following section. In this section, we use the slice-based inference algorithm in all the following experiments. The comparative models are as follows:

- **RTM** Relational Topic Model (RTM) [9]. We used the implementation of RTM from *A Fast And Scalable Topic-Modeling Toolbox*[3] for comparison.
- **dRTM** Discriminative Relational Topic Model (dRTM) [10] is an extension of RTM with topic discriminative constrains. Note that dRTM still needs to prefix the topic number.
- **CESNA** Communities from Edge Structure and Node Attributes (CESNA) [22] is a community detection model. Its implementation online (*SNAP*[4]) is used for the following comparison. The number of communities also need to be fixed in advance.

5.3.2 Results and Discussions

The comparative results of four models on three datasets are given in Fig. 8, 9 and 10. Since all the models need the topic number as an input except NRT, the x-axis in each figure denotes the topic number. Since NRT does not need the predefined topic number as an input, it does not impacted by it so its result is plotted as a line in figures with topic number as x-axis. In each figure, there are three subfigures: the first subfigure shows the results on the link prediction through *LinkRank*; the second subfigure shows the results on the document prediction through *WordRank*; the third subfigure shows positive link prediction through *AUC*. Note that the slice version of NRT in Algorithm 2 is used as the implementation of NRT. The reason is that slice version is more efficient than truncated version because the slice version does not need to keep the (relatively) large number of hidden topics in memory (the initial guess for the number of topics is normally set as larger than the number of documents).

We compared our method with three comparative models in terms of link and document prediction. In terms of link prediction, our algorithm outperformed others in most categories, where we noticed some less accurate results under some RTM settings. In terms of term prediction, NRT's performance was consistently better than othere with a single exception from dRTM with topic number 10 on *Citeseer*. We can see that there is a fluctuation for other models during the change of topic number; contrarily, NRT is not impacted by the topic number setting, because it has the ability to learn it from the data. Take *cora* dataset as an example. The candidates of possible topic number are at least within $[1, 2708]$. However, for the proposed NRT model, the active topic number is automatically learned from the data (for *cora* dataset it is around 42). Without any prior domain knowledge, this topic number can achieve relatively good results on the link and document prediction considering its large range $[1, 2708]$. In terms of overall result, we argue that in the absence of an accurate domain knowledge of K value, the NRT algorithm has allowed us achieving better and more robust performance compared with the current state-of-the-art methods.

In order to show the reasonability of the learn topic number, we further evaluate NRT with different fixed topic number through Bayesian model comparison using *Cora* dataset as an example. At first, NRT is degenerated from a Bayesian nonparametric model to a fix-dimensional probabilistic model through changing the Gamma and Poisson processes to Gamma and Poisson distributions with fixed dimension K (Note that the first parameter of distribution of global π_k should be changed from $1/K^\dagger + n_{\cdot,\cdot,k}$ to $1 + n_{\cdot,\cdot,k}$). Then, we compare posterior probability of the model given the data (Since we believe all the models have the equal weights in the prior, the data likelihoods of the models could be compared instead). The model with large posterior probability is more reasonable to the given data. It is worth noticed that RTM cannot be used here for he Bayesian model comparison, because RTM and NRT are built by different blocks (i.e., probabilistic distributions or stochastic processes), different

2. http://linqs.cs.umd.edu/projects/projects/lbc/
3. http://www.ics.uci.edu/ãsuncion/software/fast.htm#rtm
4. http://snap.stanford.edu/snap

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

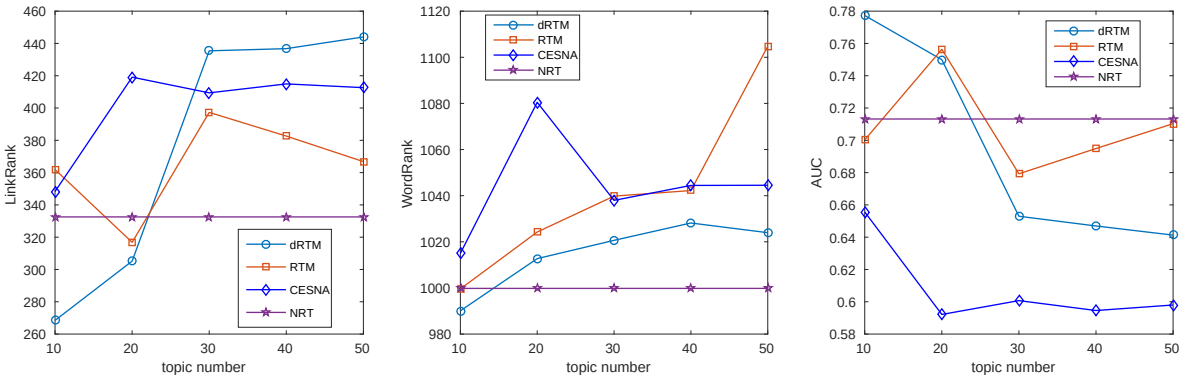


Fig. 8. Prediction results with evaluation metrics (i.e., *LinkRank*, *WordRank*, and *AUC*) on *Citeseer* dataset using 5-fold cross validation. Note that NRT does not need topic number as an input.

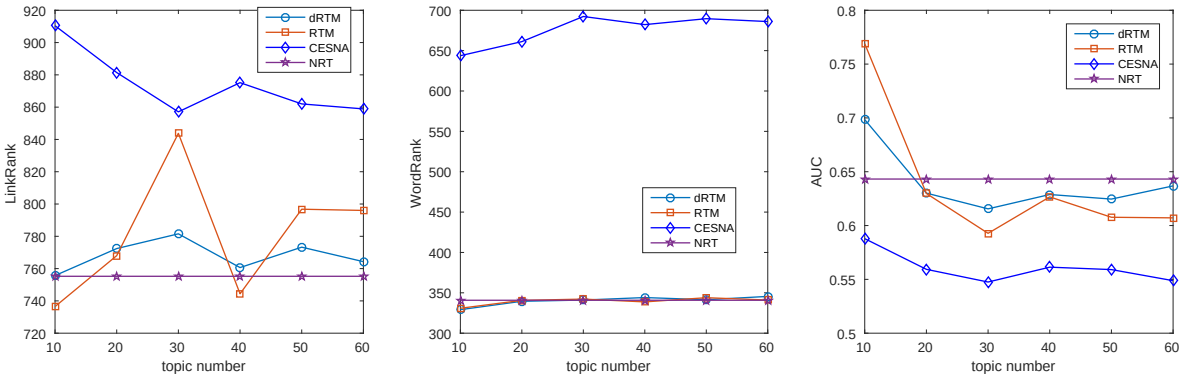


Fig. 9. Prediction results with evaluation metrics (i.e., *LinkRank*, *WordRank*, and *AUC*) on *Cora* dataset using 5-fold cross validation.

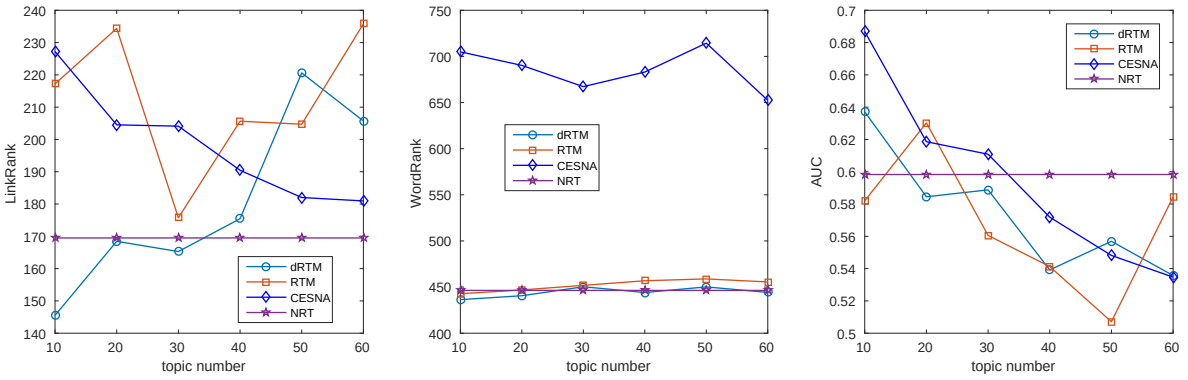


Fig. 10. Prediction results with evaluation metrics (i.e., *LinkRank*, *WordRank*, and *AUC*) on *WebKB* dataset using 5-fold cross validation.

variables and different modeling ideas. The optimized topic numbers from RTM and the probabilistic model degenerated from NRT may be different. Finally, the results are shown in Fig. 11, and we can draw the conclusion that the learned topic number from NRT is a reasonable one. Note that the learned topic number from NRT is not necessarily the global optimized one but a locally optimized one. However, the above experiments on synthetic and real-world datasets have shown the efficiency of this locally optimized one on the tasks.

6 Conclusions and Future Study

Despite of the success of existing relational topic models in discovering hidden topics from document networks, they are based on the unrealistic assumption, for many real-world applications, that the number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

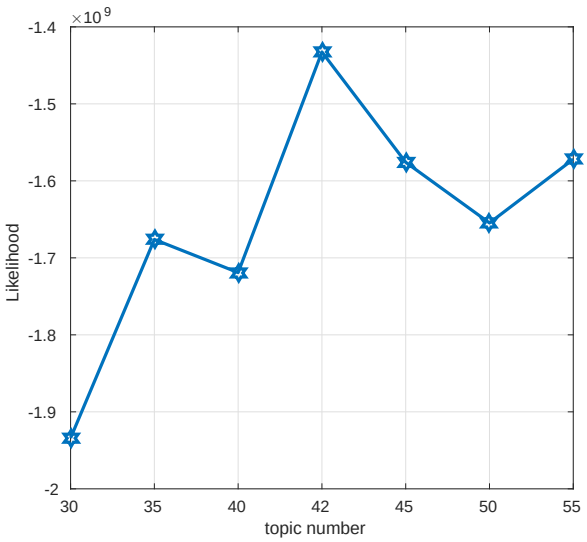


Fig. 11. Evaluation on the learned topic number from NRT on *Cora* dataset through Bayesian model comparison.

of topics can be easily predefined. In order to relax this assumption, we have presented a nonparametric relational topic model. In our proposed model, the stochastic processes are adopted to replace the fixed-dimensional probability distributions used by existing relational topic models which lead to the necessity of pre-defining the number of topics. At the same time, introducing stochastic processes leads to the difficulty with model construction and inference, and we have therefore presented a thinned Gamma process-based model and also presented truncated Gibbs and slice sampling algorithms for the proposed model. Experiments on both the synthetic dataset and the real-world dataset have demonstrated our method's ability to inference the hidden topics and their number.

In the future, we are interested in making the sampling algorithm scalable to large networks by using new network constrain methods instead of MRFs. Current MRF-based methods do not make the inference efficient enough. We believe that the network constraint methods can avoid this issue. Another interesting study would be the integration of additional information mined from the documents [45], i.e., ontology [46] from webpages.

ACKNOWLEDGMENTS

Research work reported in this paper was partly supported by the Australian Research Council (ARC) under discovery grant DP140101366 and the China Scholarship Council. This work was jointly supported by the National Science Foundation of China under grant no.61471232 and Shanghai Committee of Science and Technology International Cooperation Foundation under grant no.16550720400.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[3] J. Xuan, J. Lu, G. Zhang, R. Yi Da Xu, and X. Luo, "Infinite author topic model based on mixed gamma-negative binomial process," in *2015 IEEE International Conference on Data Mining*, Nov 2015, pp. 489–498.

[4] Z. Guo, Z. Zhang, S. Zhu, Y. Chi, and Y. Gong, "A two-level topic model towards knowledge discovery from citation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 780–794, April 2014.

[5] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine learning: ECML 2004*. Springer, 2004, pp. 217–226.

[6] H. W. Park, "Hyperlink network analysis: A new method for the study of social structure on the web," *Connections*, vol. 25, no. 1, pp. 49–61, 2003.

[7] C. Wang, J. Lu, and G. Zhang, "A constrained clustering approach to duplicate detection among relational data," in *Proceedings of 11th Pacific-Asia Conference in Knowledge Discovery and Data Mining*, ser. PAKDD '07, Nanjing, China, 2007, pp. 308–319.

[8] J. Chang and D. M. Blei, "Relational topic models for document networks," in *AISTATS*, 2009, pp. 81–88.

[9] J. Chang, D. M. Blei *et al.*, "Hierarchical relational models for document networks," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 124–150, 2010.

[10] N. Chen, J. Zhu, F. Xia, and B. Zhang, "Discriminative relational topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2014.

[11] J. Xuan, J. Lu, G. Zhang, and X. Luo, "Topic model for graph mining," *IEEE Transactions on Cybernetics*, vol. 45, no. 12, pp. 2792–2803, Dec 2015.

[12] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email." *Journal of Artificial Intelligence Research*, vol. 30, pp. 249–272, 2007.

[13] Y. Cha and J. Cho, "Social-network analysis using topic models," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 565–574.

[14] V. Tuulos and H. Tirri, "Combining topic models and social networks for chat data mining," in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, Sept 2004, pp. 206–213.

[15] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 101–110.

[16] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson, "Social topic models for community extraction," Tech. Rep., 2008.

[17] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, pp. 1981–2014, Jun. 2008.

[18] Y. Zhu, X. Yan, L. Getoor, and C. Moore, "Scalable text and link analysis with mixed-topic link models," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 473–481.

[19] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: How can citations help?" in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 957–966.

[20] Y. Sun, J. Han, J. Gao, and Y. Yu, "itopicmodel: Information network-integrated topic modeling," in *The Ninth IEEE International Conference on Data Mining*, Dec 2009, pp. 493–502.

[21] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: Joint models of topic and author community," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 665–672.

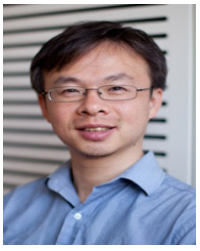
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2636182, IEEE Transactions on Knowledge and Data Engineering

[22] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 1151–1156.
[23] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.
[24] M. Revelle, C. Domeniconi, M. Sweeney, and A. Johri, *Finding Community Topics and Membership in Graphs*. Cham: Springer International Publishing, 2015, pp. 625–640.
[25] A. E. R. Robert E. Kass, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995. [Online]. Available: http://www.jstor.org/stable/2291091
[26] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.
[27] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 03 1978. [Online]. Available: http://dx.doi.org/10.1214/aos/1176344136
[28] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 583–639, 2002. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00353
[29] K. Mengersen and C. P. Robert, *Testing for mixtures: a Bayesian entropic approach*. Oxford Sci. Publ., Oxford University Press, 1996.
[30] R. C. H. C. Sujit K. Sahu, "A fast distance-based approach for determining the number of components in mixtures," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, vol. 31, no. 1, pp. 3–22, 2003. [Online]. Available: http://www.jstor.org/stable/3315900
[31] S. J. Gershman and D. M. Blei, "A tutorial on bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.
[32] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, pp. 17–35, 2007.
[33] S. Ghosal, *The Dirichlet process, related priors and posterior asymptotics*. Chapter, 2010, vol. 2.
[34] C. E. Rasmussen, "The infinite gaussian mixture model," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K. Müller, Eds. MIT Press, 2000, pp. 554–560.
[35] T. Griffiths, M. Jordan, J. Tenenbaum, and D. M. Blei, "Hierarchical topic models and the nested chinese restaurant process," *Advances in Neural Information Processing Systems*, vol. 16, pp. 106–114, 2004.
[36] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.
[37] L. Carin, D. M. Blei, and J. W. Paisley, "Variational inference for stick-breaking beta process priors," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 889–896.
[38] V. Rao and Y. W. Teh, "Spatial normalized gamma processes," in *Advances in neural information processing systems*, 2009, pp. 1554–1562.
[39] A. Roychowdhury and B. Kulis, "Gamma processes,stick-breaking, and variational inference," *arXiv preprint arXiv:1410.1068*, 2014.
[40] N. J. Foti, J. D. Futoma, D. N. Rockmore, and S. Williamson, "A unifying representation for a class of dependent random measures," in *AISTATS*, 2013, pp. 20–28.
[41] R. M. Neal, "Slice sampling," *Annals of Statistics*, pp. 705–741, 2003.
[42] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3086–3094.
[43] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.
[44] Y. Wang and L. Carin, "Levy measure decompositions for the beta and gamma processes," in *Proceedings of the 29th International Conference on Machine Learning*, ser. ICML '12. New York, NY, USA: ACM, 2012.
[45] C. Wang, J. Lu, and G. Zhang, "Mining key information of web pages: A method and its application," *Expert Systems with Applications*, vol. 33, no. 2, pp. 425–433, 2007.
[46] ——, "Integration of ontology data through learning instance matching," in *2006 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2006*, 2006, pp. 536–539.

Junyu Xuan is a postdoctoral research fellow with the Faculty of Engineering and Information Technology at University of Technology Sydney. His main research interests include Machine Learning, Text Mining, Web Mining and Complex Network. He has published about 20 papers, including TOIS, TSMC, TCYB, ICDM, IJCNN, and so on.

Jie Lu is a full professor and Associate Dean Research with the Faculty of Engineering and Information Technology at the University of Technology Sydney. Her research interests lie in the area of learning-based decision support systems. She has published 10 research books and 400 papers, won 8 Australian Research Council discovery grants and 20 other grants. She serves as Editor-In-Chief for KBS and IJCIS, and delivered 14 keynotes in international conferences.

Guangquan Zhang is an associate professor with the Faculty of Engineering and Information Technology at the University of Technology Sydney. His main research interests lie in the area of uncertain information processing. He has published 4 monographs and over 300 papers in refereed journals, conference proceedings and book chapters. He has won 7 Australian Research Council discovery grants and guest edited many special issues for international journals.



Richard Yi Da Xu is a Senior Lecturer in Faculty of Engineering and Information Technology at the University of Technology Sydney. His current research interests include machine learning, computer vision, and statistical data mining. He has published about 50 papers, including TIP, TKDE, TNNLS, PR, TKDD, AAAI, ICIP, and so on.

Xiangfeng Luo is a professor in the School of Computers, Shanghai University, China. His main research interests include Web Wisdom, Cognitive Informatics, and Text Understanding. He has published over 140 papers in refereed journals, conference proceedings and book chapters, including THMS, TSMC, TBD, TLT, and so on. He has won 4 grants from National Science Foundation of China and 5 other grants.