ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Nonparametric Relational Model to Discover Hidden Topics

[1] K.Mallika, [2] G.VadivelMurugan
[1] Final Year PG CSE Student, Sree Sowdambika College of Engineering, Aruppukottai, Tamilnadu State, India
[2] Asst. Prof of CSE, Sree Sowdambika College of Engineering, Aruppukottai, Tamilnadu State, India

***Abstract* – Nonparametric relational topic models provide a successful way to discover the hidden topics from a document network. Most of theoretical and practical tasks, such as dimensional reduction, document clustering, and link prediction, would benefit from this revealed knowledge. 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. Specifically, each document is assigned a Gamma process, although this method provides an solution, it brings additional challenges when mathematically modeling the network structure of typical document network i.e., two spatially closer document stand to have more similar topics. we require the topics are shared the documents through gamma process. In order to resolve these challenges, we use a sub sampling strategy to assign each and every document a different Gamma process from the global Gamma process, and the sub sampling probabilities of documents are assigned with a sampling technique instead of Markov Random Field constraint that inherits the document network structure. Through the posterior inference algorithm, we can discover the hidden topics and its number simultaneously. Experimental results on the capabilities of learning the hidden topics and, more importantly, the number of topics.**

Keywords:— Text mining, network analysis, topic model, Bayesian nonparametric.

I. 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. 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 is an example of a document network in which the academic papers are linked by their citation relations; an email network is a document network in which the emails are linked by their reply relations; a webpage network 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) have already been successfully developed. Unlike the traditional topic models 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 Poisson 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.

II. PROPOSED SYSTEM

In order to overcome this drawback, we propose a Nonparametric Relational Topic (NRT) model in this paper, there are three challenges: 1) How to express the document interest on infinite number of topics? Instead 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. 2) How to make all the documents share the same set of topics? This is a common feature found in many real-world applications 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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

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 sub sampling probabilities of all the documents on topics, and make the linked documents subsample the similar topics. A sub sampling 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.

III. NONPARAMETRIC RELATIONAL MODEL

In this section, we present the proposed Nonparametric Relational (NR) 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..

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. When the number of topics cannot be reasonably prefixed with enough prior knowledge. It makes traditional topic models built by probability distributions not work. 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.

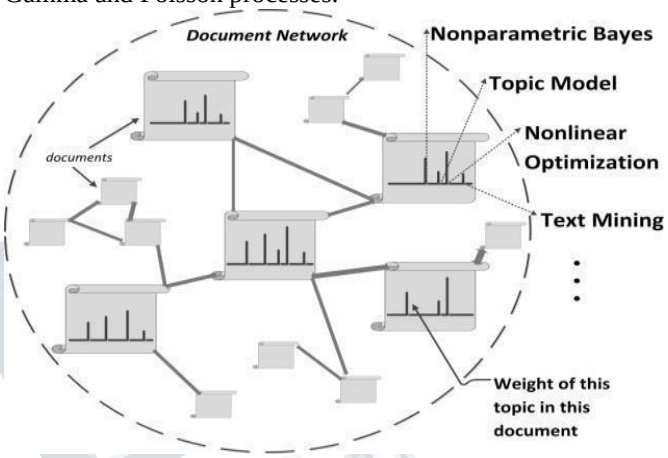


Fig:1 Illustration of Gamma process assignments for a document network. Each document is assigned a Gamma process which has infinite components (represented by the fences in the figure)

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 $R^+ \times \Theta$ where $\pi k \in R^+$ and $\theta k \in \Theta$, and it can be represented , where $\delta\theta k$ is an indicator function (i.e., $\delta\theta k(\theta^*) = 1$ if $\theta k = \theta^*$and $\delta\theta k(\theta^*) = 0$ if θk 6= θ^*); πk 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. When using Γ to express the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

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 πk is within (0,+∞) 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. 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 handle Challenge 2, we firstly generate a global Gamma process, i.e., $\Gamma_0 \sim GaP(\alpha, H)$, 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)
Suppose we have countably infinite $\{(\pi_k, \theta_k)\}_{k=1}^{\infty}$ points from a Gamma process Γ~GaP(α,H). Then, we generate a set of independent binary variables. which is proofed. The {rk} can be seen as the indicators for the reservation of the point of original/global Gamma process, so Γ^0 is called Thinned Gamma Process. We can give each *rk* a Bernoulli prior *qk*, where *qk* ∈ [0,1] is the subsampling probability of keeping topic *k*. Apparently, different realizations of {*rk*} will lead to different thinned Gamma processes. For each document, a thinned Gamma process Γ*d* is generated with Γ0 as the global process,

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

k=1 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, where *qd,k* denotes the probability of the Gamma processΓ*d* of document *d* with component *k*. Until now, Challenge 2 is handled. It makes traditional topic models built by probability distributions not work. we use a draw from a Gamma process to express the interest of a document on infinite hidden topics

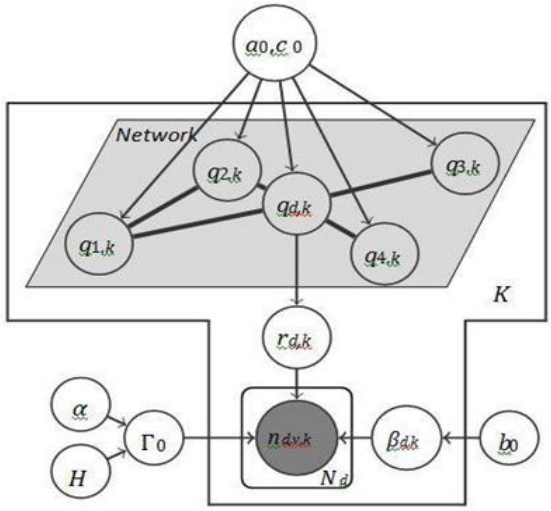


Fig. 2. Graphical representation for the Nonparametric Relational (NR) Model.

IV. MODEL INFERENCE

Naïve Bayes Algorithm
It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

4.1Gibbs 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† as the (potential) maximum number of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

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 $r_{d,k} = 1$, $p(q_{d,k}|)$

$\propto q\ a_{d,k}^{a0+1-1}(1-q_{d,k})^{c0-1}$

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) $\exists v, n_{d,v,k} > 0 \rightarrow r_{d,k} = 2$
3) $\forall v, nd, v, k = 0$
   a) if $\forall v, ud, v, k = 0$,
   b) if $\forall v, ud, v, k = 0$, $p(1)(r_{d,k} = 0)$
   c) if $\exists v, u\ d, v, k > 0$, $p(2)(r_{d,k} = 0) \propto (1 - q_{d,k})$

Accordingly, we can use a discrete distribution to sample r by, 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 where $n^{P}v$ $n_{d,v,k}$ is the number of words assigned to topic k in document d Sampling θ k. In our model, we set H where $n,v,k = {}^{P}d\ n_{d,v,k}$ 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 nd,v,1,...,nd,v,K† together due to the

$$n_{d,v} = \sum_{k=1}^{K} n_{d,v,k}$$

known according to Multinomial distribution p(nd,v,1,...,nd,v,K†|•••)∝Mult(nd,v;ξd,v,1,...,ξd,v,K†) where

$$\xi_{d,v,k} = \frac{\theta_{k,v} r_{d,k} \pi_k \beta_{d,k}}{\sum_{k}^{K^\dagger} \theta_{k,v} r_{d,k} \pi_k \beta_{d,k}}$$

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

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

Algorithm 1: Truncated Version of Gibbs Sampling for NRT

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

1: randomly set initial values for , $K, \{\theta_k\}_{k=1}^{K}$
2: *it* = 1;
3: **while** *it* ≤ *max*$_{it}$ **do**
4: **for** each topic *k* **do**
5: **for** each document *d* **do**
6: **for** each word *v* of document *d* **do**
7: Update $n_{d,v,k}$ by Eq. (13) ;
8: **end for**
9: Update $q_{d,k}$ by Eq. (5) and (6) ;
10: Update $r_{d,k}$ by Eq. (10) ;
11: Update $\beta_{d,k}$ by Eq. (11) ;
12: end for
13: Update θ_k by Eq. (12) ;
14: Update π_k by Eq. (15);
15: end for
16: it++;
17: **end while**

where n•,•,k = Pd Pv $n_{d,v,k}$ is the total number of words assign to topic k. Note that the truncation version of the model is not equal to a probability distribution-based model. 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. Note that the truncation level K† 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 quite 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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

and their parameters is time and space consuming. 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,ζ0),m. 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. e firstly generate a global Gamma process, i.e., Γ 0 ∼GaP(α,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 where Uniform(0,ζ0) is a Uniform distribution on [0,ζ0] and ζk is a fixed positive decreasing sequence limk→∞ζk = 0. With the help of slice variable ud,v,m, we can sample zd,v,m within a finite scope as follows,

$$p(z_{d,v,m} = k|\cdots) \propto \xi_{d,v,k} \cdot \frac{\Pi(u_{d,v,m} \leq \zeta_k)}{\zeta_k}$$

$$n_{d,v,k} = \sum_{m} \delta_k(z_{d,v,m})$$

$$m \in [1, n_{d,v}]$$

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 ζk is a fixed positive decreasing sequence. Sampling πk (slice sampling version) The construction of Gamma process (Γ0 ∼GaP(α,H)) [39] is,

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

where Ek andTk are two additional auxiliary variables

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

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 πk is

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

Algorithm 2: Slice Version of Gibbs Sampling for NRT

Input : Network and nd,v
Output :

1: randomly set initial values for ,
2: it = 1 ;
3: while it ≤ maxit do
4: for each topic k
do
5: for each document d do
6: for each word v of document d do
7: Sample slice variable $n_{d,v,k}$ by Eq. (16) ;
8: Update $n_{d,v,k}$ by Eq. (17) ;
9: end for
10: Update $q_{d,k}$ by Eq. (5) or (6) ;
11: Update $r_{d,k}$ by Eq. (10) ;
12: Update $\beta_{d,k}$ by Eq. (11) ;
13: end for
14: Update θ_k by Eq. (12)
15: Update π_k by Eq. (23);
16: Update κ_k by Eq. (24);
17: end for
18: it++;
19: end while

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

= if i = κk

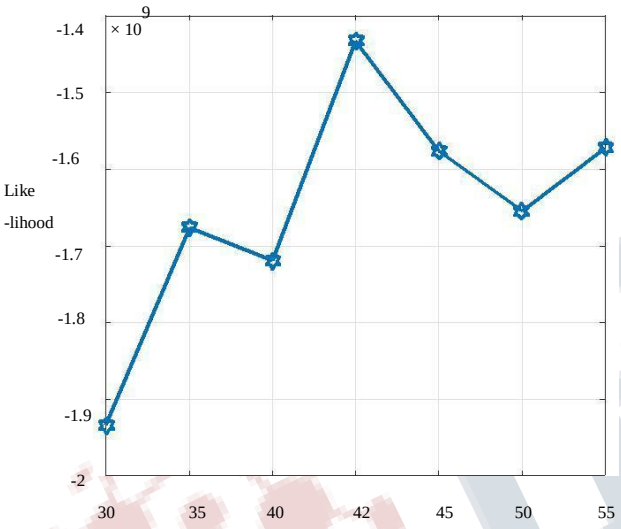
where h is an integer denotes the distance between κk with κk−1; i is the number of items in i-th. Poisson process and Ii ∼Poisson(γ); F(•|γ) and f(•|γ) are the cumulative distribution function and probability density function of Poisson distribution parameterized by γ. Note that the $u_{d,v,m}$, κ_k, 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 Algorithm2.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

V.CONCLUSION

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. 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.



Evaluation on the learned topic number from dataset through NPR 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.

ACKNOWLEDGMENT

This paper was supported by the Sree Sowdambika College of Engineering, Final Year PG CSE student Ms.K.Mallika (Reg.no:921816405006) guided by Asst.Prof of CSE Mr.G.VadivelMurugan. The authors thank to their colleagues for their help and support at different stages of the system development. Finally, we would like to thank the anonymous reviewers for their helpful comments.

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.