MEI: Mutual Enhanced Infinite Community-Topic Model for Analyzing Text-augmented Social Networks

DONGSHENG DUAN1, YUHUA LI2,*, RUIXUAN LI2, ZHENGDING LU2, AIMING WEN1

Intelligent and Distributed Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, P. R. China
Email: 1{duandongsheng, wenaiming}@smaill.hust.edu.cn
2{idcliuyhua, rxli, zdlu}@hust.edu.cn
*Corresponding author: idcliuyhua@hust.edu.cn

Community and topic can help summarize the text-augmented social networks. Existing works mixed up community and topic by regarding them as the same. However, there is inherent difference between community and topic such that treating them as the same is not so flexible. We propose a mutual enhanced infinite community-topic model (MEI) to detect communities and topics simultaneously in text-augmented social networks. Community and topic are correlated via community-topic distribution. The mutual enhancement effect between community and topic is validated by introducing two novel measures perplexity with community (perplexityc) and MRK with topic (MRKt). To determine the numbers of communities and topics automatically, Dirichlet Process Mixture model (DPM) and Hierarchical Dirichlet Process mixture model (HDP) are used to model community and topic respectively. We further introduce parameter to model the weight of community and topic responsible for community inference. Experiments on the co-author network built from a subset of DBLP data show MEI outperforms the baseline models in terms of generalization performance. Parameter study shows that MEI is averagely improved by 3.7% and 15.5% in perplexityc and MRKt respectively by setting low weight for topic. We also experimentally validate that MEI can determine the appropriate numbers of communities and topics.

Keywords: mutual enhanced infinite community-topic model, community, topic, dirichlet process, gibbs sampling

Received 00 Month 2011; revised 00 Month 2011

1. INTRODUCTION

With the prevailing of online social systems, such as online social networking site, content sharing sites and academic search systems, various kinds of social networks become ubiquitous and the number of users in such systems has been growing rapidly. In the most concise way, a social network is usually modeled as a graph in which nodes represent users and links represent relationships between the users. As we can see, users in online social systems could not only create relationships with others but also generate texts to share their own ideas, views and experiences with neighboring users. In such kind of applications, we are given not only links but also user-generated texts, such as papers published by authors in a co-author network. We call such a social network text-augmented social network. To understand the underlying communities and their formation rules hidden in text-augmented social network, community and topic detection are widely studied in the past decade. To make the following presentation more concise, we refer to text-augmented social network as social network to simply the presentation except in the context where we would like to emphasize our problem setting.

Overall speaking, community detection aims at partitioning a social network into a number of communities such that there are much denser links between users from the same community than those from different ones. On the other hand, the goal of topic modeling is to discover semantic coherent
clusters of words, known as topics, in a collection of texts. As we can see, community detection focuses on the links while topic modeling focuses on the texts. Therefore, in the following we refer to community as link emphasized community and topic as text emphasized topic. With the boom of both links and texts in online social systems, detecting communities and their topics simultaneously becomes significantly important for better understanding the ongoing increasing social networks.

For example, Figure 1 shows an artificial co-author network, where authors are linked together via co-authorships (as the solid lines show) and each author is associated with papers she/he published (as the dashed lines show). In this paper, we would like to detect research communities (as the two colored circles show) and their research topics (as the colored bars below each community show) which can be also understood as the community profile [1] in such a co-author network.

The most straightforward way for detecting communities and their topics are as follows. First, detect communities by using well established community detection algorithms. Then, perform topic modeling on the user-generated texts to obtain the topic proportions of each user. Finally, compute the topic proportion of each community as the (weighted) average of the topic proportions of users belonging to that community.

However, there are several drawbacks about the above straightforward way. First, community detection and topic modeling are not considered in a unified way such that it can not detect communities and topics simultaneously. Second, texts are not considered in community detection while links are not considered in topic modeling such that each task could not take advantage of the full information in a social network. Third, it is not convincing to treat the (weighted) average of the topic proportions of users in each community as the topic proportion of that community and how to set the corresponding weights is also a challenging issue.

In fact, there are models that are more sophisticated than the straightforward way. Some models use texts to improve community detection [2, 3, 4] while some others incorporate links into topic models [5, 6, 7, 8]. However, these models can not detect communities and topics simultaneously, and still have the second drawback as the straightforward way. There are also some models that integrate community and topic into a unified model, such as LinkLDA, Pairwise LinkLDA and Link-PLSA-LDA [9]. However, these models treat community and topic as the same latent variable while in real social networks a community can be interested in more than one topics and one topic can also be interested in by more than one communities. Therefore, modeling community and topic by using a shared latent variable is not appropriate to model real social networks.

Motivated by the above analysis, we propose a novel model for simultaneously detecting communities and topics in text-augmented social networks. In this model, we explicitly distinguish community and topic from each other by modeling them via different latent variables. On the other hand, it is observed that there are correlations between communities and topics. Users from the same community tend to be interested in similar topics. Thus, we correlate community and topic together via community-topic distribution in our model. Moreover, most previous works for community detection or topic modeling require the numbers of latent classes, i.e. community or topic, to be specified in advance. However, the appropriate numbers of latent classes are difficult to estimate in prior. To alleviate this problem, we leverage non-parametric Bayesian approach, Dirichlet Process Mixture model (DPM) and Hierarchical Dirichlet Process (HDP) mixture model, to automatically determine the numbers of both communities and topics. Since community and topic can be enhanced by each other via community-topic distribution during parameter learning process and the numbers of both communities and topics are allowed to grow infinitely when it is necessary, our model is called mutual enhanced infinite community-topic model (MEI for short).

The main contributions of this paper are summarized as follows.

- An infinite probabilistic generative model MEI is proposed to detect communities and their topics simultaneously in a text-augmented social network. The model explicitly distinguishes community and topic from each other by modeling them as different latent variables, and correlates them together via community-topic distribution. Moreover, non-parametric Bayesian approaches DPM and HDP are employed to infer the appropriate numbers of communities and topics automatically. Compared with previous models, our model captures both the difference and correlation between community and topic.
- We modify two measures, i.e. perplexity and
MRK, for evaluating the word and link prediction performance of MEI model, resulting in perplexity, and \( MRK_t \) respectively, and experimentally demonstrate that MEI has better generalization performance in terms of the modified measures than the original ones, which validates the mutual enhancement effect of MEI model between community and topic.

- We study the effect of tuning parameter on the performance of MEI and observe that MEI can be improved by 3.7\% and 15.5\% in perplexity, and \( MRK_t \) respectively by setting low weight for topic part. The performance of MEI is also evaluated and compared with several baseline models. Experimental results show MEI outperforms the baseline models in generalization performance in terms of both perplexity, and \( MRK_t \).

The rest of the paper is organized as follows. We review the related works in section 2 and introduce the mutual enhanced infinite community-topic model in detail in section 3. Section 4 gives out the Gibbs Sampling based parameter learning algorithm. The experimental settings and results are reported in section 5 and we conclude this paper in section 6.

2. RELATED WORKS

In this section, we review works that are most related to ours. For the purpose of clarity, we categorize the related works into four types, namely link emphasized community detection, text combined community detection, text emphasized topic modeling, and link combined topic modeling.

Link emphasized community detection has been studied for several decades (see [10] for an overview). Overall speaking, community detection algorithms can be classified into two classes, i.e., measurement based algorithms and probabilistic generative models. For measurement based algorithms, community detection is to partition a social network into communities such that corresponding evaluation measure is (near) optimal. Well-known such measures include normalized cut [11] and modularity [12]. However, this kind of algorithms could only describe the community structure of the observed networks but could not predict the link structure of unobserved ones. On the other hand, probabilistic generative models [13, 14, 15, 16] build a generative process for links by introducing a hidden variable to indicate the community label of each user. Besides the community labels, they also produce the pairwise community link probability which represents the probability there is a link between users from any two communities. Thus, they can generalize their results to unobserved networks. However, the generalization performance is very limited since they focus only on links in a social network.

The goal of text combined community detection [2, 3, 4] is to improve community detection results by incorporating the user-generated texts. [2] proposes a discriminate content (DC) model for text analysis and incorporates this model into popularity-based conditional link (PCL) model for community detection. In that work, authors present some other possible combinations of models for text and link analysis, such as PCL+PLSA and PHITS+DC. Although [3] aims at community outlier detection, it also produces the normal communities in a social network. In that paper, a community is defined as a collection of well-connected users with similar text information and outliers are users whose texts are significantly different from their neighboring communities. [4] addresses the scalability issue in text combined community detection. However, this type of algorithms just detect communities in a social network while do not detect the corresponding topics for the communities.

Text emphasized topic modeling is a well-known text mining tool, which becomes popular in the past decade due to its solid theoretical foundation and promising performance. PLSA [13] and LDA [17] are two basic topic models. The idea of topic models is to model the co-occurrence relationship between words in the texts by using a hidden variable, i.e., topic, which indicates the high level concept in the texts. A text is modeled as a distribution over topics and a topic is in turn modeled as a distribution over words. However, the generalization performance of basic topic modeling methods is limited by the fact that they only consider texts while ignore links in a social network.

There are also some models that integrate community and topic into a unified model. We call these models as link combined topic modeling. [7] improves PLSA by constraining that the topic proportions of neighboring users should be as similar as possible. Link-PLSA [18] models links by using the same latent variable responsible for generating texts and combines the likelihood of links into that of texts as an additional term. [19] proposes Link-LDA, which can actually be thought as an extension that replaces PLSA with LDA in Link-PLSA model. As we can see, both Link-PLSA and Link-LDA simply treat links and texts as the same. [20] further incorporates heterogenous information network into topic models. On the other hand, [9] proposes Pairwise LinkLDA and Link-PLSA-LDA, which model links by mixed membership stochastic block (MMSB) model and Link-PLSA respectively. [21] proposes relational topic model (RTM) that models links as a binary random variable that is conditioned on the topics in the connected documents. Although they model links and texts differently, both community and topic are modeled by using the same latent variable.

But in real social networks, community and topic are different from each other, thus treating them as the same is not appropriate to model real social networks. Experimental section shows these models have worse generalization performance than ours.

Finally, we would like to mention Group-Topic (GT)
model [22] which looks like MEI with the first glance. However, the problem definition of GT is different from ours. GT specifies the link between two users by identifying whether they behave the same or not during an event, such as voting on a resolution. In our situation, there are large volume of publications (corresponds to events in GT model), and for each publication only a few authors (co-authors in the publication) behave the same while most of the authors behave differently. The severe skewed effect in the scientific publications makes GT not so appropriate for our problem setting. In contrast, we directly treat co-authorships as the links, and associate users with texts she/he published while GT considers the case where texts are associated with links. Besides the problem definition, there are the following inherent differences between GT and MEI. Firstly, GT adopts a mixture of unigrams to discover topics while MEI uses LDA. Blei et al. [17] experimentally show that LDA has better generalization performance than mixture of unigrams on scientific publications. Secondly, GT produces the grouping of entities for each topic while MEI produces the topic distribution of each community. Lastly, GT requires the user to specify the numbers of both groups and topics while MEI can detect them automatically. At almost the same time with our work, [23] proposes a Latent Geographical Topic Analysis (LGTA) model that combines geographical clustering and topic modeling into one framework and allows geographical configuration and estimation of topics to benefit each other, while our model combines link emphasized community detection with topic modeling and makes community and topic mutually enhanced.

3. MUTUAL ENHANCED INFINITE COMMUNITY-TOPIC MODEL

In this section, we first give out the formal definition of the problem considered in this paper. Then, we present finite version of our model. Next, the infinite version of the model is proposed by using Hierarchical/Dirichlet Process (H/DPM). Finally, we explain how the proposed model can allow the numbers of communities and topics to grow infinitely by using the well-known Chinese Restaurant Process metaphor.

3.1. Problem Definition

The input of our problem is a text-augmented social network where each user is associated with texts she/he generated, which is formally defined as follows.

Definition 1. (Text-augmented Social Network) The text-augmented social network considered in this paper is defined as text-augmented graph $\mathcal{G} = (V, E, W)$, where $V$ is the set of users in the text-augmented social network, $E = \{(i, j) | i, j \in V\}$ is the set of social relationships between users, and $W = \{w_i\}_{i \in V}$ represents the user-generated texts with $w_i$ being the texts generated by user $i$.

The inputs of our problem include the links $E$ and texts $W$ in the text-augmented social network. For the purpose of model illustration, a variable $r$ is instead used to describe the links $E$. $r_{ij}$ is defined as 1 if $(i, j) \in E$, otherwise $r_{ij} = 0$. For texts, each element $w_i$ in $W$ is actually a sequence of words $\{w_{i1}, \ldots, w_{iN_i}\}$ where $N_i$ is the number of words in $w_i$. All the unique words in $W$ form the vocabulary.

Given the text-augmented social network defined above, we would like to detect communities and topics hidden in such a social network. For each user in $V$, we would like to infer which community it belongs to. We use $k$ to denote community and $k_i$ to indicate the community label of user $i$. On the other hand, for each word in $W$, we would like to infer which topic it belongs to. We use $z$ to denote topic and $z_{in}$ to indicate the topic labels of word $w_{in}$ which is the $n$-th word in $w_i$. When the community assignment of each user and topic assignment of each word are known, we could work out the topic distribution of each community denoted by $\phi$, the word distribution of each topic denoted by $\psi$ and the pairwise community interaction distribution denoted by $\eta$.

3.2. Mutual Enhanced Community-Topic Model

The graphical representation of the finite version of the proposed model, i.e. mutual enhanced community-topic model (ME), is shown in Figure 2. This model is actually a combination of Stochastic Block Model (SBM) [24] and Latent Dirichlet Allocation (LDA) [17]. Specifically, SBM is a generative model for the link emphasized community structure of the social network, which uses pairwise community specific Bernoulli distributions to model the presence and absence of links between pairs of users. LDA is a generative model for the user-generated texts, which models the generating process of texts by using two Multinomial distributions, i.e user-specific topic distribution and topic-specific word distribution. The detailed generative processes of ME for links and texts are very similar to SBM and LDA respectively. The novel parts of ME compared to the previous models are community-topic distributions $\{\phi_{g}\}_{g=1}^{K}$, which correlate community and topic together.

In ME model, the community variable $k$ is used to model the link structure of the social network while the topic variable $z$ is used to model the user-generated texts. Although community and topic are used to model different aspects of the social network, they can be integrated together to refine each other. On one hand, communities are coherent parts of the social network such that there are much denser links within each part than those between different ones. According to the homophily phenomenon [25] in the
social network, users from the same community tend to be interested in similar topics. Therefore, concatenating the texts generated by users in the same community together can benefit the topic detection results. On the other hand, users interested in similar topics tend to build connections with each other to form a community, thus topics can also be leveraged to improve the community detection results. The mutual enhanced process is controlled by community-topic model.

In Figure 2, the numbers of communities and topics are fixed to be $K$ and $T$ respectively. However, it is usually a difficult task to specify the numbers of communities and topics in advance. Fortunately, Dirichlet Process Mixture model (DPM) and Hierarchical Dirichlet Process mixture model (HDP) allow the numbers of latent classes to grow infinitely in a probabilistic model, and it is widely used to infer the appropriate number of latent classes in mixture models automatically. We employ H/DPM into ME, which results in mutual enhanced infinite community-topic model denoted as MEI for short. Next, we describe MEI in detail.

3.3. Mutual Enhanced Infinite Community-Topic Model

Based on ME, MEI utilizes H/DPM to determine the numbers of both communities and topics automatically. More precisely, DPM and HDP are used to model community part and topic part of ME respectively. In an informal but convenient way, we state that in community part the topical vectors $z$ of users are observable while in topic part the community assignment $k$ of users are known. For the purpose of clarity, Figure 3 illustrates the community part and topic part of ME in a graphical view.

Following [26], DPM model for the community part is defined as

$$
\begin{align*}
\pi_k | \theta_k & \sim f(\pi_k; \theta_k) \\
\theta_k | G & \sim G \\
\theta_j | G' & \sim G'
\end{align*}
$$

where $G$ and $G'$ are Dirichlet Processes with base measures $H$ and $H'$ respectively and they share the same concentration parameter $\alpha$, $f(z_i; \theta_i)$ is a multinomial distribution with parameters $\theta_i = \phi_{k_i} = [\phi_{k_i1}, \ldots, \phi_{k_iT}]$ and $f'(r_{ij}; \theta_{ij})$ is a Bernoulli distribution with parameters $\theta_{ij} = \eta_{k_i,k_j}$ where $k_i$ denotes the community assignment of user $i$ and $T$ is the number of topics. For the purpose of derivation simplicity, the base measures $H$ and $H'$ are chosen to follow symmetric Dirichlet prior with parameter $\beta$ and symmetric Beta prior with parameter $\rho$ respectively.

The graphical representation of infinite version of community part is shown in Figure 4(a). Compared with the finite version of community part, we can see that the number of communities, that is $K$, disappears in the infinite version, since the number of communities can be allowed to grow infinitely whenever it is necessary. In the community part we can see that the community assignments of users are affected by both the links and topical vectors. To distinguish the affects of the two kinds of information on the communities, we introduce parameters $\lambda$ and $1 - \lambda$ to model the weights of topic part and community part respectively for the inference of communities.

Following [27], HDP mixture model for topic part is defined as

$$
\begin{align*}
w_{gn} | \theta_{gn} & \sim f(w_{gn}; \theta_{gn}) \\
\theta_{gn} | G_g & \sim G_g \\
G_g | G_0 & \sim DP(\alpha_0, G_0) \\
G_0 & \sim DP(\gamma, H_0)
\end{align*}
$$

where $w_{gn}$ is the $n$-th word in community $g$, $G_g$ is a Dirichlet Process with concentration $\alpha_0$ and base measure $G_0$ which is drawn from an overall
Dirichlet Process with concentration $\gamma$ and base measure $H$, $f(w_{gn}; \theta_{gn})$ is a multinomial distribution with parameters $\theta_{gn} = \{\psi_{zn_1}, \ldots, \psi_{zn_V}\}$, where $z_{gn}$ is the topic assigned to word $w_{gn}$ and $V$ is the number of unique words in the vocabulary. The prior of base measure $H_0$ is defined as a symmetric Dirichlet distribution with parameter $\mu$. Figure 4(b) shows the graphical view of the infinite topic part.

For the HDP in our model, each community corresponds to a table in our model, each community corresponds to a table in a restaurant and each user corresponds to a customer. When a new customer is coming, she/he chooses a non-empty table $g$ to sit at with probability $\frac{C_g}{M-1+\alpha}$, which is proportional to the number $C_g$ of customers already sitting at that table, and chooses a new table with probability $\frac{\alpha}{M-1+\alpha}$. For the HDP in our model, each community corresponds to a restaurant and there are infinite number of tables in each restaurant. Each word corresponds to a restaurant and there are infinite number of tables in each restaurant. Each word corresponds to a customer and each topic corresponds to a dish on the global menu. In each restaurant, a customer chooses a non-empty table $t$ to sit at with probability $\frac{C_t}{\sum_{l=1}^{D_t}C_{t,l}+\mu}$, which is proportional to the number $C_t$ of customers already sitting at that table, and chooses a new table with probability $\frac{\mu}{\sum_{l=1}^{D_t}C_{t,l}+\mu}$. For each table, the waiter serves an existing dish $l$ with probability $\frac{C_{t,l}}{D_t}$, which is proportional to the number $D_t$ of tables already served that dish, and serves a new dish with probability $\frac{\mu}{\sum_{l=1}^{D_t}C_{t,l}+\mu}$.

As the above description, there is probability for assigning a new community to an user and a new topic to a word, thus DPM and HDP indeed allow the numbers of communities and topics to grow infinitely whenever the observational data tell us it is necessary.

4. MODEL LEARNING VIA GIBBS SAMPLING

In this section, Gibbs sampling based approach is presented for the learning of MEI model. Then the model parameters are estimated after the sampling process, and in turn the time complexity of the learning algorithm is analyzed and compared with several baseline models theoretically and finally the settings for some hyper-parameters are discussed.

4.1. Sampling Equations

Inspired by the Gibbs sampling equation for DPM [26] and HDP [27], we list the sampling equations for our model. The detailed derivation of these equations can be found in the Appendix. Notice that in the following, table refers to that in HDP not DPM, since in DPM a table is just a community.

Sampling equation for the community assignment of each user $i$.

$$p(k_i = g|k_{-i}, z_{-i}, r, \alpha, \beta, \rho) \propto \frac{C_{g}^{-1}}{M^{-1+\alpha}} \left[ \prod_{l=1}^{D_t} \frac{\Gamma(C_{t,l}+\alpha)}{\Gamma(C_{t,l}+\alpha+\beta)} \Gamma(C_{t,l}+\beta) \prod_{l=1}^{\gamma} \Gamma(C_g+\beta) \right]^{\lambda} \times \left[ \prod_{h=1}^{K} \frac{\Gamma(C_{h,g}+\rho)}{\Gamma(C_{h,g}+\rho+\beta)} \Gamma(C_{h,g}+\beta) \right]^{\gamma(1-\lambda)} C_{g}^{-1} > 0$$

$\text{Otherwise}$

$$\frac{C_{g}^{-1}}{M^{-1+\alpha}} \left[ \prod_{l=1}^{D_t} \frac{\Gamma(C_{t,l}+\alpha)}{\Gamma(C_{t,l}+\alpha+\beta)} \Gamma(C_{t,l}+\beta) \prod_{l=1}^{\gamma} \Gamma(C_g+\beta) \right]^{\lambda} \times \left[ \prod_{h=1}^{K} \frac{\Gamma(C_{h,g}+\rho)}{\Gamma(C_{h,g}+\rho+\beta)} \Gamma(C_{h,g}+\beta) \right]^{\gamma(1-\lambda)}$$

(3)

Sampling equation for the table assignment of each word $w_{gn}$.

$$p(t_{gn} = l|t_{-gn}, w_{gn} = v, w_{-gn}, z, \alpha, \gamma, \mu) \propto \frac{C_{g}^{-1}}{\alpha+\mu} \frac{\alpha+\mu}{\sum_{l=1}^{D_{t}}C_{t,l}+\mu} \frac{\sum_{l=1}^{D_{t}}C_{t,l}+\mu}{\sum_{l=1}^{\gamma}C_{t,l}+\mu} \frac{C_{g}^{-1}}{\gamma+\mu} \frac{\gamma+\mu}{\Gamma(\mu)} C_{g}^{-1} > 0$$

$\text{Otherwise}$

$$\frac{C_{g}^{-1}}{\alpha+\mu} \frac{\alpha+\mu}{\sum_{l=1}^{D_{t}}C_{t,l}+\mu} \frac{\sum_{l=1}^{D_{t}}C_{t,l}+\mu}{\sum_{l=1}^{\gamma}C_{t,l}+\mu} \frac{C_{g}^{-1}}{\gamma+\mu} \frac{\gamma+\mu}{\Gamma(\mu)}$$

(4)

Sampling equation for the topic assignment for a new table $t^{\text{new}}$ when the word $w_{gn}$ is sampled to that table.

$$p(z_{gn}^{t^{\text{new}}} = l|z_{-gn}^{t^{\text{new}}}, w_{gn} = v, w_{-gn}, \gamma, \mu) \propto \frac{D_{t^{\text{new}}}}{\gamma+\mu} \frac{D_{t^{\text{new}}}}{\sum_{l=1}^{D_{t}}C_{t,l}+\mu} \frac{D_{t^{\text{new}}}}{\Gamma(\mu)} \frac{D_{t^{\text{new}}}}{\Gamma(\mu)}$$

$\text{Otherwise}$

$$\frac{D_{t^{\text{new}}}}{\gamma+\mu} \frac{D_{t^{\text{new}}}}{\sum_{l=1}^{D_{t}}C_{t,l}+\mu} \frac{D_{t^{\text{new}}}}{\Gamma(\mu)} \frac{D_{t^{\text{new}}}}{\Gamma(\mu)}$$

(5)

Sampling equation for the topic assignment of each table $t$ in each community $g$.

$$p(z_{gt} = l|z_{-gt}, w_{gt}, w_{-gt}, \gamma, \mu) \propto \frac{D_{t^{g}}}{{\gamma}+\mu} \frac{D_{t^{g}}}{{\sum_{l=1}^{D_{t}^{g}}}C_{t,l}+\mu} \frac{D_{t^{g}}}{{\Gamma(\mu)}} \frac{D_{t^{g}}}{{\Gamma(\mu)}}$$

$\text{Otherwise}$

$$\frac{D_{t^{g}}}{{\gamma}+\mu} \frac{D_{t^{g}}}{{\sum_{l=1}^{D_{t}^{g}}}C_{t,l}+\mu} \frac{D_{t^{g}}}{{\Gamma(\mu)}} \frac{D_{t^{g}}}{{\Gamma(\mu)}}$$

(6)
In all the above sampling equations, $\Gamma(\cdot)$ represents Gamma function. Parameter $\lambda$ bounded by the range of $(0, 1)$ in Eqn. 3 indicates the weight that we put on the topic part of our model when sampling the community for each user while $1 - \lambda$ is the weight on community part. In this setting, the larger the value of $\lambda$, the more we emphasize the affect of topical vectors on the community membership of users, and the less we emphasize the links between users. In our previous study [28], we omitted this parameter so that the less we emphasize the links between users. In our vectors on the community membership of users, and the value of $\lambda$ on community part. In this setting, the larger the $\lambda$ for each user while 1 of (0, 1) in Eqn. 3 indicates the weight that we put on community for each user.

4.2. Parameter Estimation Algorithm

The average time complexity of each iteration (line 3 to line 5) in Algorithm 1 is $O(M \times K \times (T + K) + N \times Ta \times T + K \times Ta \times T \times V)$, where $M$ is the number of users, $K$ represents the average number of communities in each iteration, $T$ represents the average number of topics, $Ta$ is the average number of tables for each community, $N$ is the total number of words in the texts generated by all the users, and $V$ is the number of words in the vocabulary. The time complexity is analyzed as follows.

In each iteration, the community for each of $M$ users is first sampled. We need to compute Eqn. 3 for $M \times K$ times. The time complexity of computing Eqn. 3 is $O(T + K)$ since we need perform two iterations, one for $T$ times and the other for $K$ times, which results in $O(M \times K \times (T + K))$ complexity for line 3 in Algorithm 1. Similarly, we can obtain that the complexity of line 4 is $O(N \times Ta \times T)$ and that of line 5 is $O(K \times Ta \times T \times V)$.

Put the three parts together, we obtain the average time complexity of each iteration.

For the purpose of comparison, we also present some baseline models and their time complexities. The baselines include:

- **SBM**: SBM only models the community structure in a social network but ignores the user-generated texts.
- **LDA**: LDA only models the topics shared by a set of user-generated texts but ignores the links between users.
- **LinkLDA**: LinkLDA models both texts and links. However, it models topic and community by the same latent variable which is a distribution over both words and links.
- **Pairwise LinkLDA**: Like LinkLDA, Pairwise LinkLDA also models the topic and community by the same latent variable. Other than LinkLDA, it applies Mixture Membership Stochastic Block (MMSB) model for link emphasized community detection.

To compare all the models fairly, we also perform non-parametric Bayesian inference for the baseline models by using Gibbs sampling. It is worthy to point out that there are also some other works for analyzing text-augmented social networks, but the above methods are directly related to MEI. We also present the time complexity of each sampling iteration for the baseline models as we do for MEI. The time complexities of the models are listed in Tab. 2.

For the comparison purpose, we refine the time complexity to be concise one via some intuitions. The intuitions include, the number of users $M$ is significantly less than that of words $N$ in the texts, the average number of communities $K$ is less than that of topics $T$ because of the first intuition, the average number of tables $Ta$ is comparable to that of topics $T$ and the number of links $E$ is much less than the number of possible user pairs $N^2$, which are usually the facts in real social networks. With the intuitions, we rewrite the time complexity of MEI as $O(N \times Ta \times T + K \times Ta \times T \times$
By using the same idea, we also present concise form of obtained by selecting the larger one from the two terms. With the concise forms, we could conveniently see the relationship between the number of links between user and other models in Tab. 2. Based on the above intuitions, we can conclude that $O(N \times T)$ dominates $O(M \times K \times (T + K))$. Since we are not clear about the relationship between $N$ and $K \times V$, $O(K \times T \times V)$ is thought to be comparable with $O(N \times T)$, thus the concise form for MEI is obtained by selecting the larger one from the two terms. By using the same idea, we also present concise form of time complexity for other models in Tab. 2.

With the concise forms, we could conveniently see that the time complexity of the models satisfies the partial relationship as $SBM \prec MEI \preceq LDA \preceq LinkLDA \prec Pairwise LinkLDA$. In subsection 5.5, we will experimentally validate the theoretical results.

### 4.4. Hyper-parameter Setting

In the MEI model, there are some hyper-parameters, including the concentration parameters of Dirichlet Process, $\alpha$, $\alpha_0$, $\gamma$, and Dirichlet prior parameters $\beta$, $\mu$, and Beta prior parameter $\rho$. For the Dirichlet prior parameters $\beta$ and $\mu$ and $\rho$, we set all of them to be 0.01 empirically.

For the concentration parameters, instead of setting them manually we sample them iteratively by using the methods proposed in [27] and [29]. Those methods assume that the concentration parameters have Gamma priors and sample them with the help of one or two auxiliary variables. Specifically, $\alpha$, $\alpha_0$ and $\gamma$.

### TABLE 1. The meaning of symbols involved in sampling equations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i, j$</td>
<td>User indexes</td>
</tr>
<tr>
<td>$k_i$</td>
<td>The community assignment of user $i$</td>
</tr>
<tr>
<td>$z_t$</td>
<td>The topic assignment of words generated by user $i$</td>
</tr>
<tr>
<td>$k_{-i}$</td>
<td>The community assignments of users except user $i$</td>
</tr>
<tr>
<td>$z_{-i}$</td>
<td>The topic assignments of all the words except those generated by user $i$</td>
</tr>
<tr>
<td>$r$</td>
<td>The social relationships between users</td>
</tr>
<tr>
<td>$g, h$</td>
<td>Community indexes</td>
</tr>
<tr>
<td>$l$</td>
<td>Topic index</td>
</tr>
<tr>
<td>$M, K, T, V$</td>
<td>The total number of users, non-empty communities, non-empty topics and unique words respectively</td>
</tr>
<tr>
<td>$C_g$</td>
<td>The number of users already assigned to community $g$ except user $i$</td>
</tr>
<tr>
<td>$C_{lt}$</td>
<td>The number of words in community $g$ assigned to topic $t$ except words generated by user $i$</td>
</tr>
<tr>
<td>$C_{l}$</td>
<td>The number of words generated by user $i$ that are assigned to topic $t$</td>
</tr>
<tr>
<td>$C_{gh}$</td>
<td>The number of links between users from community $g$ and $h$</td>
</tr>
<tr>
<td>$C_{gh}$</td>
<td>The number of absent links between users from community $g$ and $h$</td>
</tr>
<tr>
<td>$C_{ih}$</td>
<td>The number of absent links between user $i$ and users from community $h$</td>
</tr>
<tr>
<td>$t_{gw}$</td>
<td>Table index, $t_{gw}$ represents an empty table</td>
</tr>
<tr>
<td>$t_{gw}$</td>
<td>The table assignment of the $n$-th word in community $g$</td>
</tr>
<tr>
<td>$t_{-gw}$</td>
<td>The table assignment of all the words except the $n$-th word in community $g$</td>
</tr>
<tr>
<td>$w_{gw}$</td>
<td>The $n$-th word in community $g$</td>
</tr>
<tr>
<td>$v$</td>
<td>Word index in the vocabulary</td>
</tr>
<tr>
<td>$w_{-v}$</td>
<td>Words except the $n$-th word in community $g$</td>
</tr>
<tr>
<td>$z$</td>
<td>Topic assignment of all the words</td>
</tr>
<tr>
<td>$C_{lgt}$</td>
<td>The number of words assigned to the $t$-th table in community $g$ except the $n$-th word in community $g$</td>
</tr>
<tr>
<td>$M_{gt}$</td>
<td>The total number of words generated by users in community $g$</td>
</tr>
<tr>
<td>$D_t$</td>
<td>The number of tables assigned to topic $t$</td>
</tr>
<tr>
<td>$D$</td>
<td>The total number of non-empty tables</td>
</tr>
<tr>
<td>$z_{gt}$</td>
<td>The topic assignment of the $t$-th table in community $g$</td>
</tr>
<tr>
<td>$C_{-gt}$</td>
<td>The times that word $v$ is assigned to topic $t$ except the instances on table $t$ in community $g$</td>
</tr>
<tr>
<td>$C_{nt}$</td>
<td>The times that word $v$ is assigned to table $t$</td>
</tr>
<tr>
<td>$\alpha, \alpha_0, \gamma$</td>
<td>Concentration parameters for the Dirichlet process and Hierarchical Dirichlet process</td>
</tr>
<tr>
<td>$\beta, \mu, \rho$</td>
<td>Parameters for the conjugate prior distributions</td>
</tr>
</tbody>
</table>

### TABLE 2. Time complexity of MEI and the baseline models

<table>
<thead>
<tr>
<th>Models</th>
<th>Time Complexity</th>
<th>Concise form</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEI</td>
<td>$O(M \times K \times (T + K) + N \times T \times M \times T \times V)$</td>
<td>$O(\max(N, K \times V) \times T \times T)$</td>
</tr>
<tr>
<td>SBM</td>
<td>$O(M \times K \times K)$</td>
<td>$O(M \times K \times K)$</td>
</tr>
<tr>
<td>LDA</td>
<td>$O(N \times T \times M \times T \times V)$</td>
<td>$O(M \times V \times T \times T)$</td>
</tr>
<tr>
<td>LinkLDA</td>
<td>$O((N + E) \times T \times M \times T \times (V + M))$</td>
<td>$O(\max(M^2, M \times V) \times T \times T)$</td>
</tr>
<tr>
<td>Pairwise LinkLDA</td>
<td>$O((N + N^2) \times T \times M \times T \times (V + T))$</td>
<td>$O(N^2 \times T \times T)$</td>
</tr>
</tbody>
</table>
are supposed to have \( \text{Gamma}(1, 0.1), \text{Gamma}(1, 1), \text{Gamma}(1.0, 0.1) \) as priors respectively in our model, and we set the iteration number for sampling these hyper-parameters to be 20.

5. EXPERIMENTS

5.1. Evaluation Measures

To evaluate the performance of MEI model, two widely used measures, perplexity and link prediction performance in terms of MRK, are used. Since MEI simultaneously models texts and links, perplexity is used to measure how well MEI models the texts, and MRK is used to measure how well MEI models the links. To evaluate the effect of mutual enhancement between community and topic, we modify the original evaluation measures and propose perplexityc by incorporating the community information into the computation of original perplexity and MRKt by incorporating topic information into the computation of link presence probability.

5.1.1. Perplexity with Community

Perplexity [30] is a widely used measure to evaluate the generalization performance of a probabilistic model. Lower perplexity value indicates better generalization performance. For MEI, the perplexity for a set of held-out user-generated texts \( \{w_{test}\} \) with \( M^{test} \) users and \( N_i^{test} \) words for each user \( i(1 \leq i \leq M^{test}) \) is computed as

\[
\text{perplexity}(w^{test}) = \exp\left\{-\frac{\sum_{i=1}^{M^{test}} \sum_{t=1}^{N_i^{test}} \ln p(w_{i,t}^{test})}{\sum_{i=1}^{M^{test}} N_i^{test}}\right\}
\]

The key to computing perplexity is to compute the prediction probability \( p(w_{i,t}) \) of a single word \( w_{i,t} \) in held-out texts. The original method computes \( p(w_{i,t}) \) according to the topic distribution of user \( i \) and the word distribution of topics learned from the training data. Formally, it is computed as follows.

\[
p(w_{i,t}) = \sum_{t=1}^{T} \zeta_{it} \psi_{t,w_{i,t}}
\]

where \( \zeta \) is actually the topic distribution of users and \( \zeta_{it} \) represents the proportion on topic \( l \) of user \( i \), which is estimated by the following formula after the Gibbs sampling process.

\[
\zeta_{it} = \frac{C_{it} + \beta}{\sum_{l=1}^{T} C_{il} + T \beta}
\]

If user \( i \) does not exist in the training data, we compute the prediction probability for words generated by such a user by supposing that the user corresponds to all the users in the training data with equal probability, that is \( \frac{1}{M} \). However, MEI not only produces the topic distribution of each user, but also more importantly produces the topic distribution of the community that the user belongs to. Therefore, with MEI we can alternatively compute the prediction probability \( p(w_{i,t}) \) by incorporating community information as the following equation, while we could not compute it in this way with other existing topic models since they either do not model community and topic together or model them as the same latent class.

\[
p_c(w_{i,t}) = \sum_{g=1}^{K} p^{test}(k_i = g) \sum_{t=1}^{T} \Phi_{gl} \psi_{t,w_{i,t}}
\]

In the above equation, if user \( i \) in the test data is also in the training data and it has been assigned to community \( g \) in the training process, then \( p^{test}(k_i = g) = 1 \) and \( p^{test}(k_i = h) = 0 \) for \( h \neq g \), otherwise \( p^{test}(k_i = g) \) is estimated as \( \pi_g \) which is the proportion of users in community \( g \). Note that, we use \( p_c(w_{i,t}) \) instead of \( p(w_{i,t}) \) to distinguish the prediction probability with community information from that without community information. Similarly, we refer to perplexity computed based on \( p_c(w_{i,t}) \) as \( \text{perplexity}_c \).

Note that, if we regard each user as a single community then \( \text{perplexity}_c \) is the same as original perplexity. In this sense, \( \text{perplexity}_c \) is a generalization of original perplexity.

5.1.2. MRK with Topic

To evaluate the ability of MEI model summarizing the link structure, the trained model is used to predict the links between held-out users. The probability of the presence of a test link \( r_{ij}^{test} \) between two users \( i \) and \( j \) is computed as follows.

\[
p(r_{ij}^{test} = 1) = \frac{\sum_{g=1}^{K} \sum_{h=1}^{K} p^{test}(g|w_i)p^{test}(h|w_j) \eta_{gh}}{\sum_{g=1}^{K} \sum_{h=1}^{K} p^{test}(g|w_i)p^{test}(h|w_j) \eta_{gh}}
\]

where \( \eta_{gh} \) is the community pairwise interaction probability learned from the training data and \( p^{test}(g|w_i) \) represents the probability of user \( i \) belonging to community \( g \) conditioned on text \( w_i \). Therefore, the key to computing \( p(r_{ij}^{test} = 1) \) is to estimate \( p^{test}(g|w_i) \).

Existing models, such as LinkLDA and Pairwise LinkLDA, regard community and topic as the same so that community \( g \) also means topic \( g \). According to Bayesian theorem, \( p^{test}(g|w_i) \) is proportional to \( p^{test}(w_i|g)p^{test}(g) \), where \( p^{test}(g) \) is estimated as \( \pi_g \) and \( p^{test}(w_i|g) \) is estimated by using the word distribution of communities (or topics) learned from the training data. In this case, \( p^{test}(g|w_i) \) is computed as follows.

\[
p^{test}(g|w_i) \propto \pi_g \prod_{n=1}^{N_i^{test}} \psi_{gw_{i,n}}
\]

Since SBM does not model the texts at all, \( \psi_{gw_{i,n}} \) is not defined, thus we simply set all \( \psi_{gw_{i,n}} \) as 1. With this setting, \( p^{test}(g|w_i) \) turns out to be \( p^{test}(g) = \pi_g \) which
is the proportion of users in community \( g \) learned from the training data.

However, MEI not only produces the community structure of the social network, but also produces the topic distribution of each community, with which we could compute the probability of each user belonging to an existing community. In another word, with MEI we can alternatively compute \( p_{t}^{\text{test}}(g|w_i) \) by incorporating the topic information of communities as follows. We use \( p_{t}^{\text{test}}(g|w_i) \) aiming at distinguishing \( p_{t}^{\text{test}}(g|w_i) \) with topic information from that without topic information.

\[
p_{t}^{\text{test}}(g|w_i) \propto p_{t}^{\text{test}}(g) \prod_{n=1}^{N_{\text{test}}} \sum_{l=1}^{T} \phi_{gl} \psi_{lw_i}
\]

Like [9], the performance of link prediction is evaluated by a rank value as follows. For each user \( i \), the predicted probabilities that \( i \) links to all the other users can be computed by using Eqn. 8, then the links are ranked according to these probabilities. It is expected that the probabilities of present links in the test data should be higher than absent ones. The lower the largest rank of the existing links for each user, the better the link prediction performance. We use the mean of the rank values for all the users, \( \text{MRK} \) for short, to evaluate the link prediction performance. Similarly, we use \( \text{MRK}_t \) to refer to \( \text{MRK} \) computed based on \( p_{t}^{\text{test}}(g|w_i) \).

Note that, when we set \( \phi_{gg} \) to be 1 and set \( \phi_{gl} \) to be 0 for all \( l \neq g \), \( p_{t}^{\text{test}}(g|w_i) \) becomes the same as \( p_{t}^{\text{test}}(g|w_i) \), and accordingly, \( \text{MRK}_t \) reduces to \( \text{MRK} \). Therefore, \( \text{MRK}_t \) is actually a generalized \( \text{MRK} \).

5.2. Dataset

In the following experiments, we use the papers published in SIGMOD, KDD, WWW and SIGIR from year 2006 to 2010 to test the performance of the proposed model. In the dataset, authors correspond to the users and co-authorships correspond to the links in the social network. The text associated with each author is the concatenation of the titles of all the papers his/her published. As a preprocessing, we remove the authors who have less than 3 papers and also delete the stop words and words that occur less than 10 times. Finally, there are totally 874 authors, 2157 co-authorships and 21503 words left.

For the test purpose, we divide the dataset into 5 parts with each part corresponding to one year. We conduct experiments by choosing arbitrary 4 parts as the training data and the left part as the test data. These experiments are denoted as E2006, E2007, E2008, E2009 and E2010 respectively. For example, E2010 represents the experiment using the data from year 2006 to 2009 to train the models and the data of year 2010 to test the models.

5.3. Generalization Performance Study

In this subsection, the generalization performance of MEI is studied and compared with baseline models in terms of \( \text{perplexity}_c \) and \( \text{MRK}_t \). For each experiment, taking E2006 as an example, we train the models using the data from the year 2007 to 2010 and compute the \( \text{perplexity}_c \) and \( \text{MRK}_t \) on the data of year 2006 using the learned model parameters. In the experiments of this subsection, the Gibbs sampling iteration numbers are all set to 1000 with burn-in period of 800 and sample gap of 10 and the initial number of communities, tables and topics are all set to 150. Empirical results in subsection 5.4 show that initializing the number of latent classes as 150 makes Gibbs sampling process converge after several hundred iterations, such that we believe that 800 iterations are sufficient for Gibbs sampling chains to archive the steady state distribution in our data set.

5.3.1. Mutual Enhancement Effect Study

As we state before, MEI has the effect of mutual enhancement between community and topic. Recall that in Section 5.1, we modify the computation of original \( \text{perplexity}_c \) and \( \text{MRK}_t \) by incorporating community information and topic information, resulting in \( \text{perplexity}_c \) and \( \text{MRK}_t \) respectively. Here, we experimentally show the performance of MEI in terms of both the original measures and our modified ones. Figure 5 and 6 show the results for \( \text{perplexity}_c \) and \( \text{MRK}_t \) respectively. We see that MEI has better word and link prediction performance in terms of our modified measures than the original ones. From the results, we conclude that the word prediction performance of MEI increases when the community information is incorporated whereas the link prediction performance of MEI increases when the topic information is incorporated, which validates the mutual enhancement effect between community and topic of MEI model.

Based on the observations above, in the following

![FIGURE 5. Performance of MEI model in terms of original perplexity and our modified one perplexityc. The lower the better.](image-url)
FIGURE 6. Performance of MEI model in terms of original MRK and our modified one MRK\(_t\). The lower the better.

FIGURE 7. Perplexity\(_c\), of MEI model with various values of \(\lambda\). The lower the perplexity\(_c\), the better the model’s word prediction performance.

FIGURE 8. MRK\(_t\) of MEI model with various values of \(\lambda\). The higher the MRK\(_t\), the better the model’s link prediction performance.

experiments perplexity and MRK of MEI are computed as perplexity\(_c\) and MRK\(_t\) respectively. However, the baseline models described in Section 4.3 either do not model community and topic together, such as LDA and SBM, or model them as the same latent class, such as LinkLDA and Pairwise LinkLDA. Therefore, for the baseline models we compute original perplexity and MRK, which are special cases of perplexity\(_c\) and MRK\(_t\) respectively.

5.3.2. Parameter Study
As we can see from section 4, the sampling equation of community assignment for each user involves a tuning parameter \(\lambda\), which is used to model the importance of community part and topic part. Before comparing the generalization performance of our model with that of existing baselines, we study the effect of \(\lambda\) on the performance of MEI model. In this experiment, we allow \(\lambda\) to vary in the range \((0,1)\). Figure 7 and 8 show the perplexity\(_c\) and MRK\(_t\) value for MEI model with some typical values of \(\lambda\) respectively. For example, for each value of \(\lambda\), say 0.3, the value on y-axis corresponding to E2006 on x-axis in Figure 7 represents the computed perplexity\(_c\) for experimental setting E2006 when \(\lambda\) is set to be 0.3.

From the experimental results, we can see that the lower the value of \(\lambda\), the better the generalization performance of MEI model. Actually, lower \(\lambda\) suggests that the community memberships of users are more dependent on the links between users than the topical vectors associated with users. That is also why we distinguish the link emphasized community with the text emphasized topic, since even increasing the weight of topic part for the sampling of communities will hurt the performance of the model.

Notice that setting \(\lambda = 0.5\) means community part and topic part play the same significant role on the sampling for communities of users. Taking this setting as the baseline, our experiments show that low value for \(\lambda\), e.g. 0.1, improves perplexity\(_c\) and MRK\(_t\) of MEI model by 3.7% and 15.5% respectively by averaging over the results of all the experimental settings.

Although there are improvements if setting the parameter \(\lambda\) properly, the experimental results also show that the performance of MEI model is not so sensitive to the parameter \(\lambda\), because the improvement is not so significant, especially compared with the performance of existing baselines, which are shown in the experiments of the next two subsections. Therefore, \(\lambda\) and \(1 - \lambda\) are removed in Eqn. 3 in the following experiments, in which we put the same weight on both community part and topic part.

In the following we compare the performance of MEI with existing baselines. Due to the stochastic nature of Gibbs Sampling based learning algorithm, each experiment is performed for five times and the average value and standard deviation of perplexity\(_c\) and MRK\(_t\) are compared among different models.
5.3.3. Perplexity Comparison Result
The comparison results of perplexity of different models with various experimental settings are illustrated in Figure 9. SBM does not model the user-generated texts, therefore the perplexity of SBM does not make any sense thus is omitted in the figure.

As Figure 9 shows, MEI has the lowest perplexity (i.e., best word prediction performance) among all the models. The underlying reason is that MEI predicts words written by authors not only according to their own past publications but also according to their community members’ publications. In another words, MEI accounts for the influence of communities (environment) over the behavior of members. In contrast, LDA predicts words only in terms of an author’s own publications while ignoring the communities’ influence to the members thus produces higher (worse) perplexity. LinkLDA and Pairwise LinkLDA perform even worse in terms of perplexity, since they mix up community and topic by using the same latent variable, i.e. topic, making the detected topics decentralized by the links and ordered node pairs respectively.

5.3.4. MRK\textsubscript{t} Comparison Result
The comparison results of MRK\textsubscript{t} of different models with various experimental settings are illustrated in Figure 10. LDA does not model the links of the social network, therefore MRK\textsubscript{t} of LDA is not shown.

As Figure 10 shows, MEI significantly outperforms all the baselines in terms of MRK\textsubscript{t}, which indicates its superior link prediction performance. SBM performs the worst for link prediction as it only takes advantage of the link information. For an unknown user, SBM does not know which community the user is likely or unlikely to belong to, and simply assigns the user to each community with equal probability. LinkLDA and Pairwise LinkLDA perform more or less the same as SBM. The underlying reason is as follows.

Both LinkLDA and Pairwise LinkLDA regard the community and topic as the same latent variable. In another word, one topic corresponds to one community in the two models. However, in real social networks a community may cover a broad range of topics and a topic may be discussed in more than one communities. Modeling community and topic by the same latent variable makes the community and topic couple very tightly. The two models predict a link between two users with a high probability if and only if their topics are similar enough. This condition for link prediction is very strong. In real case two authors from the same community but with different research interests may co-author papers in the future.

On the contrary, MEI first predicts which community the two test authors might belong to according to his/her published papers, then predicts the link between the two authors via the community-community link proportions. MEI may predict a co-authorship between two authors studying different topics with a high probability if authors working on the two topics often co-author in the training data. MEI obtains much better link prediction performance through discriminating community and topic explicitly and correlating them together through community-topic distributions. As a brief conclusion, MEI has the best generalization performance in terms of perplexity and MRK\textsubscript{t} among all the compared models.

5.4. Determine the Numbers of Communities and Topics
Since DPM and HDP are leveraged in the model, MEI can automatically determine the appropriate number of communities and topics. In this subsection, we show the process of MEI converging to the appropriate number of latent classes. In this experiment, all the data from year 2006 to 2010 is used. The number of iterations is set to 10000.

Initially, we do not know anything about the numbers
of communities and topics in the dataset, thus the numbers of both two latent classes are set to 1 as the initial value. Figure 11 shows how the numbers of communities and topics changes as a function of the iteration time. For the purpose of comparison, the number of communities detected by SBM and the number of topics detected by LDA are also illustrated. In the figure, K represents the number of communities and T denotes the number of topics, thus K/MEI means the number of communities detected by MEI and other notations can be explained in the same manner.

The results show that MEI and SBM converge to more or less the same number (about 20) of communities under this initialization. But the number of topics produced by MEI and that by LDA differ significantly. The number of topics detected by LDA is much larger than MEI under this initialization. The reason is that the topics produced by LDA are shared among users while those produced by MEI are shared among communities and there are much fewer communities than users in social networks.

From the results above, the numbers of communities and topics detected by the three models are all not larger than 120. Therefore, similar experiments are conducted but with the number of both communities and topics initialized to be 150, which is sufficiently large for the selected dataset. Under this initialization, the variation trend of the numbers of communities and topics versus iteration time is recorded in Figure 12. Again, MEI and SBM converge to more or less the same number (about 50) of communities under this initialization whereas the number of topics detected by MEI and LDA are different from each other. The number of topics produced by LDA is also much larger than MEI under this initialization, the similar result as previous initialization.

From the results of the above two extreme initializations, it can be seen that MEI can automatically detect the appropriate numbers of communities and topics. Although the numbers of communities and topics detected by the models are not consistent under different initializations, the convergence directions are the same. We observe that in Figure 11 the numbers of latent classes still increase after 10000 iterations but in a very low rate, while in Figure 12 the numbers of latent classes keep almost unchanged even after hundreds of iterations. Theoretically, we believe that both initializations converge to the same number of communities and topics when performing infinite iterations. According to the observations, we empirically set the initial numbers of both communities and topics as 150 in Section 5.3 to speed up the convergence rate of Gibbs sampling.

5.5. Scalability Study

In the subsection 4.3, we give out the time complexity analysis for MEI and several baseline models. In this subsection, we perform experiments to study the scalability of Gibbs sampling approaches for the models.

Similar with the experimental setting in the previous subsection, we also set two extreme initializations for the numbers of communities and topics in this subsection. Scalability study is performed on the two extreme settings. For a model, the average running time of one iteration is recorded. In this experiment, we plot the running time as a function of the number of edges in the text-augmented co-author network. Figure 13 and Figure 14 show the results for the two extreme initializations respectively.

Comparing the results in the two figures, we see that the running time in Figure 13 is significantly less than that in Figure 14, which is due to the different initialization. Nevertheless, we can find similar patterns from the two figures. SBM has the least running time and Pairwise LinkLDA has the greatest one. In Figure 13, LDA, LinkLDA need more or less the same running time while MEI need less time than LDA and LinkLDA. In Figure 14, LDA, LinkLDA and MEI all need more or less the same running time. These experimental results well validate the theoretic results of time complexity analyzed in subsection 4.3. On the
FIGURE 13. Average running time of one iteration with respect to the number of links in co-author network when the number of both communities and topics is initialized to be 1.

FIGURE 14. Average running time of one iteration with respect to the number of links in co-author network when the number of both communities and topics is initialized to be 150.

other hand, the theoretic results can be used to explain the comparison results revealed by our experiments.

5.6. Case Study

In this subsection, some communities and topics detected by MEI are manually checked. In this experiment, the number of both communities and topics is initiated to be 150 and the parameters for Gibbs sampling are the same as those in subsection 5.3. Under this parameter setting, MEI discovers 50 communities and 94 topics. Tab. 3 shows top 9 communities selected from these 50 communities. The title for each community is the research group or the research interest of the first author through checking his/her homepage. The top 5 authors and the number of their published papers for each community are listed just below each community and in turn top 5 topics and their corresponding probabilities. The titles are our interpretation of the topics.

As examples, let’s see some detected communities. The first community is entitled with “Context, Learning, and User Experience for Search” (CLUES) which aims at the web search related problems. As the results show, the community is also interested in graph mining and efficient query. Vanja Josifovski in the 5th community is the leader of the Performance Advertising Group at Yahoo! Research and MEI identifies one of its main topics as sponsored advertising. The main topic in the 8th community is graph mining. The authors in this community, e.g. Jiawei Han, really study graph mining related work, such as frequent graph mining. The main author in the 13th community, e.g. Jie Tang, is known to study social network mining especially academic search through the investigation to his homepage. Through manually checking, the remaining communities and their topic proportions detected by MEI also capture the background truth.

The results also show that one community discusses a wide range of topics. For example, community 8 is interested in graph mining, web search and video, although with the emphasis on graph mining. On the other hand, one topic can be studied by several communities, such as web search which is interested by almost all the selected communities. However, web search can be regarded as the background topic in the selected dataset. Besides web search, graph mining is also interested by several different communities.

Nevertheless, the background truth of communities and topics in the DBLP data can be complicated to be quantified. Therefore, we manually check the affiliations and research interests of authors from their homepages. Modeling communities and topics by different latent variables are indeed more flexible and can capture more information that previous models can not (directly) obtain, such as the topic distribution (interests) of a community.

6. CONCLUSION AND FUTURE WORK

In this paper, a mutual enhanced infinite community-topic model MEI is proposed to simultaneously detect communities and their topics in text-augmented social networks. To automatically determine the numbers of communities and topics, Dirichlet Process mixture model (DPM) and Hierarchical Dirichlet Process mixture model (HDP) are used to model the community part and topic part respectively of our model. Gibbs sampling based approach is used to estimate the model parameters. To make our model more general, tuning parameters $\lambda$ is introduced to weight the affect of community part and topic part on the sampling of community assignments of users.

In the experimental section, the generalization performance of MEI is compared with counterpart baseline models in terms of perplexity $c$ and $MRK_t$ on a subset of DBLP data. Due to the original
### TABLE 3. Top nine communities detected by MEI

<table>
<thead>
<tr>
<th>Community 1</th>
<th>Community 2</th>
<th>Community 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLUES</strong></td>
<td><strong>Web Search and Mining Group</strong></td>
<td><strong>Web Search and Mining Group</strong></td>
</tr>
<tr>
<td>Ryen W. White</td>
<td>16</td>
<td>Lei Zhang</td>
</tr>
<tr>
<td>Wei Pan</td>
<td>10</td>
<td>Eugene Agichtein</td>
</tr>
<tr>
<td>Jun Yang</td>
<td>9</td>
<td>Deepak Agarwal</td>
</tr>
<tr>
<td>C. M. Jermaine</td>
<td>7</td>
<td>Yue Pan</td>
</tr>
<tr>
<td>Luis L. Perez</td>
<td>7</td>
<td>Flavio Junqueira</td>
</tr>
<tr>
<td>topic 6</td>
<td>0.185119</td>
<td>topic 6</td>
</tr>
<tr>
<td>topic 21</td>
<td>0.110548</td>
<td>topic 68</td>
</tr>
<tr>
<td>topic 74</td>
<td>0.076333</td>
<td>topic 36</td>
</tr>
<tr>
<td>topic 26</td>
<td>0.070192</td>
<td>topic 5</td>
</tr>
<tr>
<td>community 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 6</td>
<td>0.163911</td>
<td>topic 32</td>
</tr>
<tr>
<td>topic 4</td>
<td>0.117675</td>
<td>topic 41</td>
</tr>
<tr>
<td>topic 27</td>
<td>0.106119</td>
<td>topic 67</td>
</tr>
<tr>
<td>topic 66</td>
<td>0.084056</td>
<td>topic 37</td>
</tr>
<tr>
<td>community 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 6</td>
<td>0.157995</td>
<td>topic 21</td>
</tr>
<tr>
<td>topic 7</td>
<td>0.128299</td>
<td>topic 6</td>
</tr>
<tr>
<td>topic 17</td>
<td>0.123548</td>
<td>topic 81</td>
</tr>
<tr>
<td>topic 84</td>
<td>0.109979</td>
<td>topic 62</td>
</tr>
<tr>
<td>topic 21</td>
<td>0.090288</td>
<td>topic 32</td>
</tr>
</tbody>
</table>

### TABLE 4. Twelve topics selected from those detected by MEI

<table>
<thead>
<tr>
<th>topic 3</th>
<th>topic 4</th>
<th>topic 6</th>
<th>topic 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>aqualogic platform</td>
<td>fast association</td>
<td>web search</td>
<td>academic search</td>
</tr>
<tr>
<td>platform</td>
<td>0.097897</td>
<td>association</td>
<td>0.102931</td>
</tr>
<tr>
<td>aqualogic</td>
<td>0.083921</td>
<td>fast</td>
<td>0.080890</td>
</tr>
<tr>
<td>access</td>
<td>0.069946</td>
<td>factorization</td>
<td>0.066196</td>
</tr>
<tr>
<td>event</td>
<td>0.055971</td>
<td>discovering</td>
<td>0.051502</td>
</tr>
<tr>
<td>time</td>
<td>0.048983</td>
<td>opinion</td>
<td>0.051502</td>
</tr>
<tr>
<td>topic 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temporal modeling</td>
<td>social</td>
<td>0.215475</td>
<td>mining</td>
</tr>
<tr>
<td>temporal</td>
<td>0.074775</td>
<td>networks</td>
<td>0.170116</td>
</tr>
<tr>
<td>causal</td>
<td>0.065434</td>
<td>browsing</td>
<td>0.060498</td>
</tr>
<tr>
<td>clustering</td>
<td>0.060763</td>
<td>aware</td>
<td>0.051048</td>
</tr>
<tr>
<td>classification</td>
<td>0.056093</td>
<td>network</td>
<td>0.041598</td>
</tr>
<tr>
<td>topic 32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>search queries</td>
<td>sponsored advertising</td>
<td>semantic community</td>
<td>video</td>
</tr>
<tr>
<td>search</td>
<td>0.132072</td>
<td>advertising</td>
<td>0.119155</td>
</tr>
<tr>
<td>queries</td>
<td>0.061498</td>
<td>sponsored</td>
<td>0.093263</td>
</tr>
<tr>
<td>document</td>
<td>0.056606</td>
<td>ad</td>
<td>0.093263</td>
</tr>
<tr>
<td>analysis</td>
<td>0.051715</td>
<td>rare</td>
<td>0.062193</td>
</tr>
<tr>
<td>time</td>
<td>0.049619</td>
<td>series</td>
<td>0.051836</td>
</tr>
</tbody>
</table>
measures of perplexity and MRK can not capture the mutual effect between community and topic, we modify them by incorporating community information into the computation of perplexity and topic information into the computation of MRK, resulting in two novel measures $\text{perplexity}_c$ and $\text{MRK}_t$. Comparing the performance of MEI in terms of original measures and our modified ones, the mutual enhancement effect between community and topic is validated. From the parameter study, we observe that the performance of MEI model is averagely improved by 3.7% and 15.5% in $\text{perplexity}_c$ and $\text{MRK}_t$ respectively by setting relatively low $\lambda$. By putting the same weight on community part and topic part of MEI, experimental results show that MEI performs significantly better than the baseline models in terms of both $\text{perplexity}_c$ and $\text{MRK}_t$, which indicates the superiority of MEI in summarizing and predicting the unseen texts and links. Moreover, it is validated that MEI can detect the appropriate number of communities and topics automatically, and scales more or less the same as some existing models. Finally, from further investigation into several communities and topics detected by MEI, it is found that MEI could discover meaningful communities and topics in our selected data.

In the future, we will further investigate the power of discriminating community and topic when modeling text-augmented social networks, and study how the model can benefit applications like text classification, expert search and resource recommendation.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China under Grants 70771043, 61173170, National High Technology Research and Development Program of China under Grant 2007AA01Z403, and Innovation Fund of Huazhong University of Science and Technology under Grant 2011TS135 and 2010MS068. We sincerely thank the anonymous reviewers and editors for their very comprehensive and constructive comments.

REFERENCES


APPENDIX A. DERIVATION OF GIBBS SAMPLING EQUATIONS

There are mainly four distributions involved in MEI model, that is Dirichlet distribution, Multinomial distribution, Beta Distribution, and Bernoulli distribution. Dirichlet is conjugate prior of Multinomial, and Beta is that of Bernoulli.

The probability density function of Dirichlet distribution is given by the following formula.

\[
f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma \left( \sum_{i=1}^{K} \alpha_i \right)}{\Gamma \left( \sum_{i=1}^{K} \alpha_i \right)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}
\]

According to the probability theory, integrating out the probability density function produces probability one, i.e.

\[
\int f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) dx = 1
\]

After some mathematical manipulations, we have the following equation.

\[
\int \prod_{i=1}^{K} x_i^{\alpha_i - 1} dx = \frac{\Gamma(\alpha)}{\Gamma(\alpha + \beta)}
\]

Similarly, we have the following result derived from Beta distribution.

\[
x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}
\]

The probability mass function of Multinomial distribution and Bernoulli distribution are expressed as Equation A.3 and A.4 respectively.

\[
f(x_1, ..., x_K; p_1, ..., p_K) = \frac{\Gamma \left( \sum_{i=1}^{K} x_i + 1 \right)}{\prod_{i=1}^{K} \Gamma(x_i + 1)} \prod_{i=1}^{K} p_i^{x_i}
\]

\[
f(a; p) = p^a(1-p)^{1-a}
\]

It is worthy to point out that there are some restrictions for the variables and parameters in the above probability distributions. For example, the input variable to the Dirichlet distribution \(x_1, ..., x_K\) are restricted to a \(K-1\) simplex space such that \(\sum_{i=1}^{K} x_i = 1\). The restrictions for other variables or parameters can be found in the literatures about probability theory.

According to the independence assumption of MEI shown by the graphical representation of MEI (Figure 2) and Chinese Restaurant Process (CRP) (Section 3.4), \(p(k_i = g | k_{-i}, z_i, z_{-i}, r)\) can be factored as follows.

\[
p(k_i = g | k_{-i}, z_i, z_{-i}, r) \propto p(k_i = g | k_{-i}) \times [p(z_i | k_i = g, z_{-i}, k_{-i})]^\lambda [p(r_i | k_i = g, r_{-i}, k_{-i})]^{1-\lambda}
\]

where \(\lambda\) and \(1-\lambda\) are the weights of topic part and community part for the sampling of community membership of users respectively.

\[
p(k_i = g | k_{-i})\text{ represents the distribution of community assignment of user} \ i \text{ conditioned on the community assignment of all the other users without any more evidences. The distribution can be formulated}
\]
as follows, which has been explained by using CRP in Section 3.4.

\[
p(k_i = g|k_{-i}) = \begin{cases} 
\frac{C_{g-1}}{M - 1 + \alpha} & C_{g-1} > 0 \\
C_{g-1} > 0 & \text{Otherwise} 
\end{cases} \tag{A.6}
\]

If \(C_{g-1} > 0\), \(p(z_i|k_i = g, z_{-i}, k_{-i})\) is derived as follows.

\[
p(z_i|k_i = g, z_{-i}, k_{-i}) = \begin{cases} 
p(z_i, z_{-i}|k_i, k_{-i}) & \text{otherwise} 
\end{cases} \tag{A.7}
\]

From the above, we have

\[
p(z_i|k_i = g, z_{-i}, k_{-i}) = \begin{cases} 
\prod_{l=1}^{T} \Gamma(C_{g'l} + C_{i'lg} + \beta) \Gamma(C_{i'l} + T\beta) & C_{g'-1} > 0 \\
\Gamma(T\beta) \prod_{l=1}^{T} \Gamma(C_{i'l} + T\beta) & \text{otherwise} 
\end{cases} \tag{A.7}
\]

Next, we derive the probability \(p(r_i|k_i = g, r_{-i}, k_{-i})\). Similarly, we first suppose \(C_{g'} > 0\).

\[
p(r_i|k_i = g, r_{-i}, k_{-i}) = \begin{cases} 
p(r_i, r_{-i}|k_i, k_{-i}) & \text{otherwise} 
\end{cases} \tag{A.7}
\]

In the above derivation, the second line is obtained by applying the product rule of Bayesian theorem, and the third one is obtained by applying the sum rule of Bayesian theorem. The fourth line to the fifth line is obtained by using Equation A.3 and the probability density function of Dirichlet distribution. The last line is obtained by applying Equation A.1 two times.

Otherwise, if \(g\) is a new community, \(p(z_i|k_i = g, z_{-i}, k_{-i})\) is derived as

\[
p(z_i|k_i = g, z_{-i}, k_{-i}) = \begin{cases} 
p(z_i, z_{-i}|g, k_{-i}) & \text{otherwise} 
\end{cases} \tag{A.7}
\]

Since there are not any users except \(i\) in community \(g\), \(C_{g'i}^{\alpha} \) disappears in the above equation.
By combining Eqn. A.5, A.6, A.7 and A.8, the community sampling equation (Equation 3) is obtained.

Similarly, the sampling equations for tables and topics associated with each word can be derived as that for communities as above. The main techniques used are Bayesian theorem and the usage of Eqn. A.1, A.2, A.3 and A.4.