MEI: Mutual Enhanced Infinite Generative Model for Simultaneous Community and Topic Detection

Dongsheng Duan, Yuhua Li*, Ruixuan Li, Zhengding Lu, and Aiming Wen

School of Computer Science and Technology,
Huazhong University of Science and Technology Wuhan 430074, P. R. China
{duandongsheng,wenaiming}@smail.hust.edu.cn
{idcliyuhua,rxli,zdlu}@hust.edu.cn

Abstract. Community and topic are two widely studied patterns in social network analysis. However, most existing studies either utilize textual content to improve the community detection or use link structure to guide topic modeling. Recently, some studies take both the *link emphasized* community and *text emphasized* topic into account, but community and topic are modeled by using the same latent variable. However, community and topic are different from each other in practical aspects. Therefore, it is more reasonable to model the community and topic by using different variables. To discover community, topic and their relations simultaneously, a *mutual enhanced infinite generative model* (MEI) is proposed. This model discriminates the community and topic from one another and relates them together via community-topic distributions. Community and topic can be detected simultaneously and can be enhanced mutually during learning process. To detect the appropriate number of communities and topics automatically, Hierarchical/Dirichlet Process Mixture model (H/DPM) is employed. Gibbs sampling based approach is adopted to learn the model parameters. Experiments are conducted on the co-author network extracted from DBLP where each author is associated with his/her published papers. Experimental results show that our proposed model outperforms several baseline models in terms of perplexity and link prediction performance.

Keywords: social network analysis, community detection, topic modeling, mutual enhanced infinite generative model, dirichlet process, gibbs sampling.

1 Introduction

Social network is usually modeled as a graph where nodes represent users and edges represent links among users. However, in many applications we encounter not only the link structure but also user generated textual contents in a social network, such as papers published by authors in a co-author network. With the

* Corresponding author.

T. Elomaa, J. Hollmén, and H. Mannila (Eds.): DS 2011, LNCS 6926, pp. 91–106, 2011.

widespread of social networking sites and micro-blogs, social networks with user generated textual contents become ubiquitous and have been growing rapidly, which makes it a hot topic to mine latent patterns from them. Community and topic are two important patterns receiving much attention recently. Overall speaking, communities are densely linked sub-graphs with relatively sparse links to the outside in the social network and topics are semantic coherent clusters of correlated words in the user generated textual contents. In the following, we review the related works and show the key points of our work.

There are large volume of works about *link emphasized* community detection [3] and *text emphasized* topic modeling [7,1]. However, these works consider only one aspect (either links or contents) of the social network. Community detection algorithms discover communities by summarizing the link structure but ignoring the textual contents, while topic modeling regards the textual contents as a collection of separated documents without links among them.

In the most recent studies, some models and algorithms are put forward to combine the links and contents. However, these works combine link structure and textual content either for community detection [18,4,11] or for topic modeling [9,5,8,15] only. Although [12] proposes Pairwise LinkLDA and Link-PLSA-LDA that take both the link structure and textual content into a unified generative model, both community and topic are modeled by using the same latent variable. In contrast, we explicitly discriminate community and topic from each other by modeling them via different variables. Our model is motivated by the common sense that a community can be interested in more than one topics and a topic can be interested by more than one communities. In addition, community and topic are related together via the community-topic distributions in our model.

Group-topic model [17] also discriminates group from topic. In group-topic model the textual contents are associated with events while in our model the textual contents are associated with users. Moreover, the textual contents in group-topic model is not generated by users but distributed by governments or official organizations. Group-topic model can not be directly applied to solve the community and topic detection issues in our problem setting.

Moreover, most previous works for community detection or topic modeling require the number of latent classes to be specified in advance. However, the appropriate number of latent classes are difficult to estimate in prior. To alleviate this problem, we leverage non-parametric Bayesian approach, Hierarchical/Dirichlet Process Mixture model (H/DPM), to automatically select the number of communities and topics. Since community and topic can be enhanced by each other via community-topic distributions and the number of both communities and topics are allowed to grow infinitely, our model is named *mutual enhanced infinite generative model* (MEI for short).

The main contributions of this paper are summarized as follows.

– A generative model MEI is proposed to model the social network with textual contents. This model explicitly discriminates community and topic from each other through modeling them by different latent variables, and relates them together via community-topic distributions. Moreover, non-parametric

Bayesian approach H/DPM is employed to select the appropriate number of communities and topics automatically. Compared to the previous models, our model captures both the difference and correlation between community and topic.
– The performance of MEI is evaluated and compared with several baseline community or topic models through experiments on the co-author network extracted from DBLP. Experimental results show MEI outperforms the baseline models in terms of both perplexity and link prediction performance.

The rest of the paper is organized as follows. Section 2 proposes the mutual enhanced infinite generative model. Section 3 gives out the Gibbs Sampling based learning algorithm. The experimental settings and results are reported in section 4 and we conclude this paper and show some future works in section 5.

2 Mutual Enhanced Infinite Generative Model

In this section, mutual enhanced infinite generative model is proposed. Firstly, we present finite version of the model. Secondly, 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 number of communities and topics grow infinitely by using Chinese Restaurant Process metaphor.

2.1 Mutual Enhanced Generative Model

The graphical representation of the finite version of the proposed model, i.e. *mutual enhanced generative model* (ME), is shown in Figure 1. This model is actually a combination of Stochastic Block Model (SBM) [14] and Latent Dirichlet Allocation (LDA) [1]. Concretely, SBM is a generative model for the link emphasized community structure of the social network. It uses community-pair specific Binomial distributions to model the presence and absence of edges between pairs of users. LDA is a generative model for the textual contents associated with users. It models the generating process of words by using two Multinomial distributions, i.e user-specific topic distribution and topic-specific word distribution. The detailed generative process of ME for link structure and textual contents are very similar to SBM and LDA. 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 generative model, the community variable k is mainly used to model the link structure of the social network while the topic variable z is used to model the textual content associated with users. Although community and topic 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, in which there are much denser links than to the outside. According to the homophyly phenomenon [10] in the social network, users from the same community tend to be interested in similar topics. Therefore, concatenating the

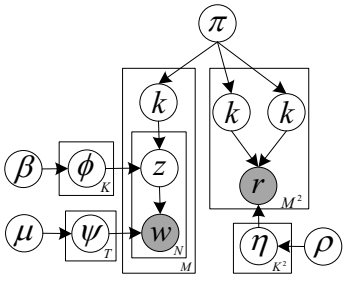

Fig. 1. Graphical representation of ME

textual contents associated with users from the same community together can benefit the topic detection results. On the other hand, users with the similar topics tend to be from the same community, thus topics can also be leveraged to improve the community detection results. The mutual enhanced process is controlled by community-topic distribution, which naturally correlates community and topic together.

In Figure 1, the number of communities and topics are fixed to be K and T respectively. However, it is usually a difficult task to specify the number of communities and topics in prior. Fortunately, Hierarchical/Dirichlet Process Mixture model (H/DPM) allows the number of latent classes to grow infinitely in a probabilistic model, and they are widely used to select the appropriate number of latent classes in mixture models automatically. We employ H/DPM into ME, which forms *mutual enhanced infinite generative model* denoted as MEI for short. Next, we describe MEI in detail.

2.2 Mutual Enhanced Infinite Generative Model

Based on ME, MEI utilizes H/DPM to select the number of both communities and topics automatically. More precisely, DPM and HDP are used for modeling *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 2 illustrates the community part and topic part of ME in a graphical view.

Following [13], DPM model for the community part is defined as

$$\begin{aligned} z_i|\theta_i &\sim f(z_i;\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, H_0) \end{aligned} \tag{1}$$

where G is a Dirichlet Process with base measure H_0 and concentration α, $f(z_i;\theta_i)$ is a multinomial distribution with parameters $\theta_i = \phi_{k_i} = \{\phi_{k_i 1}, \ldots, \phi_{k_i T}\}$ where k_i denotes the community assignment of user i and T is the number of topics. For the purpose of simplicity, the base measure H_0 is assumed to follow a symmetric Dirichlet prior with parameter β.

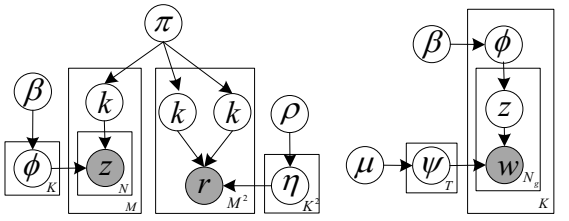

Fig. 2. Graphical representation of (a) community part and (b) topic part

Following [16], HDP mixture model for topic part is defined as

$$\begin{aligned} w_{gn}|\theta_{gn} &\sim f(w_{gn};\theta_{gn}) \\ \theta_{gn}|G_g &\sim G_g \\ G_g|G_0 &\sim DP(\alpha_0, G_0) \\ G_0 &\sim DP(\gamma, H) \end{aligned} \tag{2}$$

where w_{gn} is the n-th word in community g, G_g is a Dirichlet Process with concentration α_0 and base measure G_0 which is drawn from an overall Dirichlet Process with concentration γ and base measure H, $f(w_{gn};\theta_{gn})$ is a multinomial distribution with parameters $\theta_{gn} = \psi_{z_{gn}} = \{\psi_{z_{gn}1}, \ldots, \psi_{z_{gn}V}\}$, 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 is defined as a symmetric Dirichlet distribution with parameter μ.

H/DPM defined above are formal enough to be understood. In the following, we resort to the Chinese Restaurant Process to describe the model in a more comprehensible way.

2.3 Chinese Restaurant Process Metaphor

Chinese restaurant process [16] can be used to explain how DPM and HDP allow the number of communities and topics to grow infinitely. For the DPM 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 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 $\frac{C_{gt}}{\alpha_0+N_g-1}$, which is proportional to the number C_{gt} of customers already sitting at that table, and chooses a new table with probability $\frac{\alpha_0}{\alpha_0+N_g-1}$. For each table, the waiter serves an existing dish l with probability $\frac{D_l}{\gamma+D}$, which is proportional to the number D_l of tables already served that dish, and serves a new dish with probability $\frac{\gamma}{\gamma+D}$.

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 number of communities and topics to grow infinitely.

3 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 finally the settings for some hyper-parameters are discussed.

3.1 Sampling Equations

Inspired by the Gibbs sampling equation for DPM [13] and HDP [16], we list the sampling equations for our model. The detailed derivation of these equations is omitted due to space limit. Notice that in the following *table* refers to that in HDP not DPM, since in DPM a table is just a community.

Sampling equation of the user community variables for each user i.

$$p(k_i = g|k_{-i}, z_i, \boldsymbol{z}_{-i}, \boldsymbol{r}, \alpha, \beta, \rho) \propto$$
$$\begin{cases} \frac{C_g}{M-1+\alpha} \frac{\prod_{l=1}^{T} \Gamma(C_{gl}^{-i}+C_{il}+\beta)}{\Gamma(\sum_{l=1}^{T}(C_{gl}^{-i}+C_{il})+T\beta)} \frac{\Gamma(\sum_{l=1}^{T} C_{gl}^{-i}+T\beta)}{\prod_{l=1}^{T}\Gamma(C_{gl}^{-i}+\beta)} \\ \times \prod_{h=1}^{K} \frac{(C_{gh}+\rho)\Gamma(\overline{C}_{gh}+\rho)\Gamma(C_{gh}^{-i}+\overline{C}_{gh}^{-i}+2\rho)}{\Gamma(C_{gh}+\overline{C}_{gh}+2\rho)\Gamma(C_{gh}^{-i}+\rho)\Gamma(\overline{C}_{gh}^{-i}+\rho)} & C_g^{-i} > 0 \\ \frac{\alpha}{M-1+\alpha} \frac{\Gamma(T\beta)}{\prod_{l=1}^{T}\Gamma(\beta)} \frac{\prod_{l=1}^{T}\Gamma(C_{il}+\beta)}{\Gamma(\sum_{l=1}^{T} C_{il}+T\beta)} \prod_{h=1}^{K} \frac{\Gamma(C_{ih}+\rho)\Gamma(\overline{C}_{ih}+\rho)\Gamma(2\rho)}{\Gamma(C_{ih}+\overline{C}_{ih}+2\rho)\Gamma(\rho)\Gamma(\rho)} & Otherwise \end{cases} \tag{3}$$

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

$$p(t_{gn} = t|t_{-gn}, w_{gn} = v, w_{-gn}, z, \alpha_0, \gamma, \mu) \propto$$
$$\begin{cases} \frac{C_{gt}}{\alpha_0+N_g-1} \frac{C_{lv}+\mu}{\sum_{v'=1}^{V} C_{lv'}+V\mu} & C_{gt} > 0 \\ \frac{\alpha_0}{\alpha_0+N_g-1} \sum_{l=1}^{T} \frac{D_l}{\gamma+D} \frac{C_{lv}+\mu}{\sum_{v'=1}^{V} C_{lv'}+V\mu} + \frac{\gamma}{\gamma+D}\frac{\Gamma(\mu)}{\Gamma(V\mu)} & Otherwise \end{cases} \tag{4}$$

Sampling equation of the table topic variables for a new table t^{new} when the word w_{gn} is sampled to that table.

$$p(z_{gt^{new}} = l|z_{-gt^{new}}, w_{gn} = v, w_{-gn}, \gamma, \mu) \propto \begin{cases} \frac{D_l}{\gamma+D} \frac{C_{lv}+\mu}{\sum_{v'=1}^{V} C_{lv'}+V\mu} & D_l > 0 \\ \frac{\gamma}{\gamma+D}\frac{\Gamma(\mu)}{\Gamma(V\mu)} & Otherwise \end{cases} \tag{5}$$

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

$$p(z_{gt} = l|z_{-gt}, w_{gt}, w_{-gt}, \gamma, \mu) \propto$$
$$\begin{cases} \frac{D_l}{\gamma+D} \frac{\prod_{v=1}^{V}\Gamma(C_{lv}^{-gt}+C_{tv}+\mu)}{\Gamma(\sum_{v=1}^{V}(C_{lv}^{-gt}+C_{tv})+V\mu)} \frac{\Gamma(\sum_{v=1}^{V} C_{lv}^{-gt}+V\mu)}{\prod_{v=1}^{V}\Gamma(C_{lv}^{-gt}+\mu)} & D_l^{-gt} > 0 \\ \frac{\gamma}{\gamma+D} \frac{\prod_{v=1}^{V}\Gamma(C_{tv}+\mu)}{\Gamma(\sum_{v=1}^{V} C_{tv}+V\mu)} \frac{\Gamma(V\mu)}{\prod_{v=1}^{V}\Gamma(\mu)} & Otherwise \end{cases} \tag{6}$$

In all the above sampling equations, $\Gamma(\cdot)$ represents Gamma function and $C_{\cdot\cdot}$ and $D_{\cdot}$ represent the number of samples during Gibbs sampling and superscript '$-$' represents excluding the instances sampled currently. For example, C_{lv}^{-gt} is number of topic l is assigned to word v excluding the words on the table t of community g, and D_l^{-gt} is the number of tables assigned to topic l except the table t in community g.

3.2 Parameter Estimation Algorithm

Once the community assignment for each object and the topic assignment for each word are sampled, the model parameters $\{\phi_g\}_{g=1}^{K}$, $\{\psi_l\}_{l=1}^{T}$ and $\{\eta_{gh}\}_{g=1,h=1}^{K,K}$ can be estimated by using Maximum Likelihood Estimation (MLE) as follows.

$$\phi_{gl} = \frac{C_{gl} + \beta}{\sum_{l'=1}^{T} C_{gl'} + T\beta} \quad \psi_{lv} = \frac{C_{lv} + \mu}{\sum_{v'=1}^{V} C_{lv} + V\mu} \quad \eta_{gh} = \frac{C_{gh} + \rho}{C_{gh} + \overline{C}_{gh} + 2\rho} \tag{7}$$

Based on all the above analysis and formulas, the learing algorithm for MEI is summarized in Algorithm 1.

Algorithm 1. Parameter Estimation Algorithm for MEI

Input: A social network $\boldsymbol{r} = \{r_{ij} | 1 \le i, j \le M\}$ and user generated textual contents $w_i = \{w_{i1}, \ldots, w_{iN_i}\}$ for $1 \le i \le M$

Output: The community assignment for each object k_i and the parameters $\{\phi_g\}_{g=1}^{K}$, $\{\psi_l\}_{l=1}^{T}$ and $\{\eta_{gh}\}_{g=1,h=1}^{K,K}$

1 Initialize the community of each object, the table and the topic associated with each word randomly;
2 **repeat**
3 Sample the community of each object through Eqn. 3;
4 Sample the table of each word through Eqn. 4; If t is a new table, then sample the topic of that table through Eqn. 5;
5 Sample the topic of each table through Eqn. 6;
6 **until** *Reaches Iteration Number* ;
7 Estimate model parameters $\{\phi_g\}_{g=1}^{K}$, $\{\psi_l\}_{l=1}^{T}$ and $\{\eta_{gh}\}_{g=1,h=1}^{K,K}$ through Eqn. 7;

3.3 Hyper-parameter Setting

In the MEI model, there are some hyper-parameters, including the concentration parameters of Dirichlet Process, α, α_0, γ, and Dirichlet prior parameters β, μ, and Beta prior parameter ρ. For the Dirichlet prior parameters β and μ and ρ, 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 method proposed in [16] and [2]. Those methods assume that the concentration parameters have Gamma priors and sample them with the help of one or two auxiliary variables. Particularly, in our model α,

α_0 and γ are supposed to have $Gamma(1, 0.1)$, $Gamma(1, 1)$, $Gamma(1.0, 0.1)$ as priors respectively and set the iteration number for sampling these hyper-parameters to be 20.

4 Experiments

4.1 Evaluation Criterion

To evaluate the performance of MEI model, two widely used criterion, perplexity and link prediction performance, are used. Since MEI simultaneously models textual content and link structure, perplexity is used to measure how well MEI models the textual content and link prediction performance is used to measure how well MEI models the link structure.

Perplexity. Perplexity [6] is a widely used measure to evaluate the generalization performance of a probability model. Lower perplexity value indicates better generalization performance. For MEI, the perplexity for a set of held-out users' generated textual contents $\{\boldsymbol{w}^{test}\}$ with M^{test} users and N_i^{test} words for each user $i(1 \leq i \leq M)$ is calculated as

$$perplextity(\boldsymbol{w}^{test}) = \exp\{-\frac{\sum_{i=1}^{M^{test}} \sum_{n=1}^{N_i^{test}} \ln p(w_{in})}{\sum_{i=1}^{M^{test}} N_i^{test}}\} \tag{8}$$

$$p(w_{in}) = \sum_{g=1}^{K} p^{test}(k_i = g) \sum_{l=1}^{T} \varphi_{gl} \psi_{lw_{in}}$$

In the above equation, if user i in the test data is also in the training data and it is assigned to community g, 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 π_g.

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

$$p(r_{ij}^{test} = 1 | w_i, w_j) = \sum_{g=1}^{K} \sum_{h=1}^{K} p^{test}(g|w_i) p^{test}(h|w_j) \eta_{gh} \tag{9}$$

where $p^{test}(g|w_i)$ is proportional to $p^{test}(g) \prod_{n=1}^{N_i^{test}} \sum_{l=1}^{T} \varphi_{gl} \psi_{lw_{in}}$ where $p^{test}(g)$ is estimated as π_g.

Like [12], the performance of link prediction is evaluated by a rank value as follows. For each user i, the probabilities that i links to all the other users can be computed, then the links can be 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, MRK for short, to evaluate the link prediction performance.

4.2 Baseline Models

In this subsection, some baseline models are listed for the comparison purpose with MEI model. These baselines include,

SBM: SBM only models the community structure in a social network but ignores the user generated textual contents.

LDA: LDA only models the topics shared by a set of user generated documents but ignore the link structure between users. LDA simply regards each user as a community, which means LDA does not consider communities at all.

LinkLDA: LinkLDA models both textual contents and link structure. 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.

In the following experiments, we compare the performance of MEI model with the above baselines. To compare all the models on equal footing, we also make the non-parametric Bayesian inference for the baseline models. There are also some other works for analyzing social network with textual content, but the above methods are directly related to MEI.

4.3 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 textual content associated with each author is the concatenating of all the titles of his/her published papers. 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 devide 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 testing 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.

4.4 Performance Study

In this subsection, the practical performance of MEI is studied and compared with baseline models in terms of perplexity and MRK value. For each experiment, taking E2006 as an example, we train the models using the data from the year

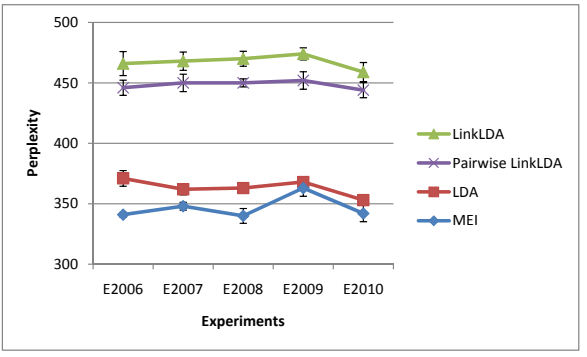

Fig. 3. The mean of perplexity output by different models for each experiment where vertical bars indicate the standard deviations

2007 to 2010 and compute the perplexity and MRK value on the data of year 2006 using the learned model. Due to the stochastic nature of Gibbs Sampling based learning algorithm, each experiment is performed for five times and then the mean value and standard deviation of perplexities and MRK values are compared among different models.

Perplexity Comparison Result. The comparison results of perplexity produced by each model in each experiment is illustrated in Figure 3. SBM does not model the user generated textual contents, therefore the perplexity of SBM does not make any sense thus is omitted in the figure.

As Figure 3 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' publication. 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 performs even worse in terms of perplexity, since the topic detected by them is decentralized by the links and ordered node pairs respectively.

MRK Comparison Result. The comparison results of MRK value produced by each model in each experiment is illustrated in Figure 4. LDA does not model the link structure of the social network, therefore the MRK value of LDA is not shown.

As Figure 4 shows, MEI significantly outperforms all the baselines in terms of MRK value, which indicates its superior link prediction performance. SBM performs the worst for link prediction as it only uses the link information. For an unknown user, SBM does not know which community the user likely or unlikely to belong to, and simply assigns the user to each community with the probability

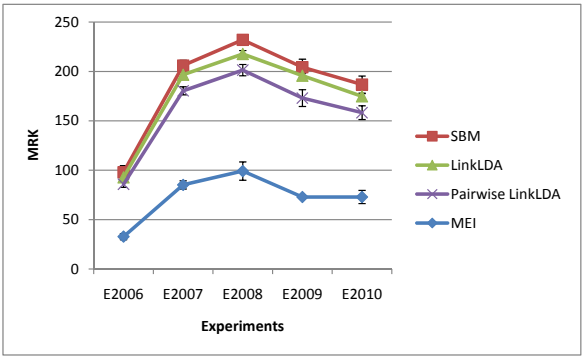

Fig. 4. The mean of MRK output by different models for each experiment where vertical bars indicate the standard deviations

proportional to the size of the community. LinkLDA and Pairwise LinkLDA performs 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, which also means 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 gains much better link prediction performance through discriminating community and topic explicitly and relates them together through community-topic distributions. As a brief conclusion, MEI has the best perplexity and link prediction performance among all the compared models.

4.5 Select the Number of Communities and Topics

Since DPM and HDP is leveraged in the model, MEI can automatically select 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 number of communities and topics in the dataset, thus the number of both two latent classes is set to 1 as the

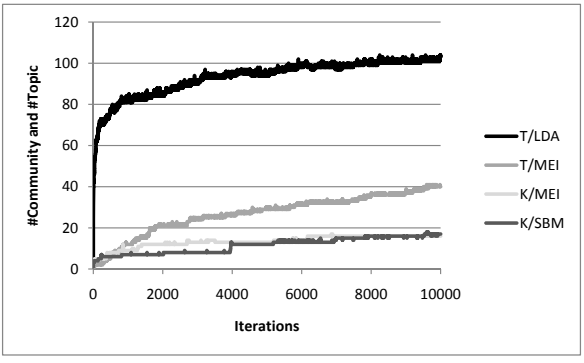

Fig. 5. Number of communities and topics versus iteration time when initializing the number of both communities and topics to be 1

initial value. Figure 5 shows how the number 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 number 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 number of communities and topics versus iteration time is recorded in Figure 6. 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 number of communities and topics to some degree. Although the number of communities and topics detected by the models are not consistent under different initializations, the convergence directions are the same. Theoretically, we believe that both initializations converge to the same number of communities and topics when performing infinite iterations. Gibbs Sampling indeed converges slowly and 10000 iteration time can be not so sufficient to simulate the complicated joint probability distribution.

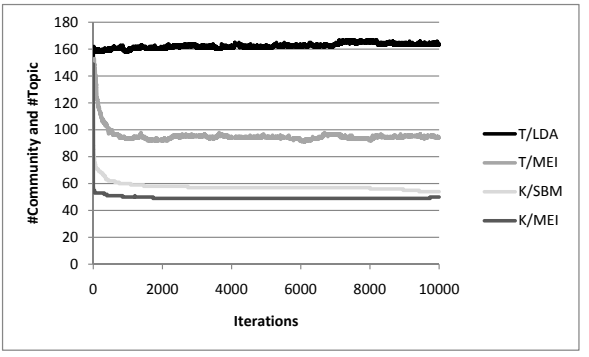

Fig. 6. Number of communities and topics versus iteration time when initializing the number of both communities and topics to be 150

4.6 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. Under this initialization, MEI discovers 50 communities and 94 topics within 10000 Gibbs sampling iterations. Table 1 shows top 9 communities selected from totally 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. Table 2 shows 12 topics involved in the selected communities. Each topic is shown with top 5 words 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 topic 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

Table 1. Top 9 communities detected by MEI

Community 1		Community 2		Community 3	
CLUES		Web Search and Mining Group		Web Search and Mining Group	
Ryen W. White	16	Lei Zhang	18	Hang Li	17
Wei Fan	10	Eugene Agichtein	11	Xin Li	13
Jun Yang	9	Deepak Agarwal	10	Hsiao-Wuen Hon	12
C. M. Jermaine	7	Yue Pan	10	Olivier Chapelle	12
Luis L. Perez	7	Flavio Junqueira	10	Vanessa Murdock	12
topic 6	0.185119	topic 6	0.223021	topic 6	0.188899
topic 21	0.110548	topic 68	0.111897	topic 14	0.175721
topic 27	0.085984	topic 32	0.085841	topic 70	0.071025
topic 74	0.076333	topic 36	0.083541	topic 27	0.068828
topic 26	0.070192	topic 5	0.075112	topic 23	0.066632
Community 4		Community 5		Community 6	
Data Mining		Performance Advertising Group		Web Research Group	
Wei Wang	17	Vanja Josifovski	20	R. A. Baeza-Yates	11
Shenghuo Zhu	16	Andrei Z. Broder	17	Jian Huang	10
Zhaohui Zheng	14	Tao Li	13	Rong Jin	9
Kai Yu	11	Raghu Ramakrishnan	13	Kevin C.-C. Chang	9
Gordon Sun	11	Susan T. Dumais	12	Jun Wang	8
topic 6	0.163911	topic 32	0.159506	topic 6	0.206593
topic 4	0.117675	topic 41	0.133465	topic 27	0.123165
topic 21	0.106119	topic 6	0.122072	topic 9	0.109261
topic 30	0.105068	topic 67	0.084639	topic 34	0.092376
topic 66	0.084056	topic 37	0.061041	topic 32	0.080458
Community 7		Community 8		Community 9	
Social Network Mining		Data Mining Research Group		Data Mining	
Jie Tang	15	Jiawei Han	38	Philip S. Yu	29
Zhong Su	14	Xuemin Lin	14	Jeffrey Xu Yu	13
Sihem Amer-Yahia	13	Hong Cheng	13	S. Vaithyanathan	10
Peter J. Haas	12	Xifeng Yan	11	R. Krishnamurthy	8
Kevin S. Beyer	9	Rui Li	11	Xiaofang Zhou	7
topic 6	0.157995	topic 21	0.228537	topic 6	0.164801
topic 7	0.128299	topic 6	0.150502	topic 3	0.136818
topic 17	0.123548	topic 81	0.095561	topic 27	0.121271
topic 84	0.100979	topic 62	0.083616	topic 21	0.120235
topic 21	0.090288	topic 32	0.070079	topic 1	0.065305

Table 2. Twelve topics selected from those detected by MEI

topic 3		Topic 4		Topic 6		topic 7	
aqualogic platform		fast association		web search		academic search	
platform	0.097897	association	0.102931	web	0.153212	extraction	0.109227
aqualogic	0.083921	fast	0.080890	search	0.115681	multi	0.067249
access	0.069946	factorization	0.066196	data	0.071979	engine	0.058853
event	0.055971	discovering	0.051502	information	0.055271	metasearch	0.050458
time	0.048983	opinion	0.051502	user	0.042418	arnetminer	0.042062
topic 9		topic 17		Topic 21		Topic 27	
temporal modeling		social networks		graph mining		efficient query	
modeling	0.130821	social	0.215475	mining	0.134176	query	0.133652
temporal	0.074775	networks	0.170116	model	0.054413	efficient	0.049922
causal	0.065434	browsing	0.060498	approach	0.048602	retrieval	0.046164
clustering	0.060763	aware	0.051048	large	0.048074	system	0.040797
classification	0.056093	network	0.041598	graph	0.047546	data	0.040261
Topic 32		topic 41		Topic 68		topic 81	
search queries		sponsored advertising		semantic community		video	
search	0.132072	advertising	0.119155	semantic	0.150721	video	0.076348
queries	0.061498	sponsored	0.093263	community	0.085431	discriminative	0.061094
document	0.056606	ad	0.093263	question	0.065341	customer	0.053467
analysis	0.051715	rare	0.062193	models	0.060318	advertising	0.045839
time	0.049619	series	0.051836	score	0.050273	reachability	0.045839

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 model can not obtain, such as the topic distribution (interests) of a community.

5 Conclusion and Future Work

In this paper, mutual enhanced infinite generative model MEI is proposed for social network analysis. To automatically select the number of communities and topics, Hierarchical/Dirichlet Process mixture model are leveraged in our model. Gibbs sampling based approach is used to estimate the model parameters. In the experimental section, the perplexity and link prediction performance of MEI are studied and compared with counterpart baseline models on the DBLP data. Experimental results show that MEI performs better than the baseline models in terms of perplexity and link prediction performance. Moreover, it is validated that MEI can detect the appropriate number of communities and topics automatically. Finally, from the further investigation into several communities and topics detected by MEI, it is found that MEI really discovers meaningful communities and topics. In the future, we will further investigate the power of discriminating community and topic when modeling social network with textual contents and study how the model can benefit other applications, such as text classification, expert search and resource recommendation. To understand the model more deeply, we will also investigate the time consumption and scalability of the learning algorithm for MEI.

Acknowledgments. This work is supported by National Natural Science Foundation of China under Grant 70771043, 60873225, 60773191, National High Technology Research and Development Program of China under Grant 2007AA01Z403, Natural Science Foundation of Hubei Province under Grant 2009CDB298, Wuhan Youth Science and Technology Chenguang Program under Grant 200950431171, Open Foundation of State Key Laboratory of Software Engineering under Grant SKLSE20080718, and Innovation Fund of Huazhong University of Science and Technology under Grants 2010MS068 and Q2009021.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
2. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588 (1994)

3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (2010)
4. Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., Han, J.: On community outliers and their efficient detection in information networks. In: KDD, pp. 813–822 (2010)
5. Guo, Z., Zhang, Z.M., Zhu, S., Chi, Y., Gong, Y.: Knowledge discovery from citation networks. In: ICDM, pp. 800–805 (2009)
6. Heinrich, G.: Parameter estimation for text analysis. Technical report, University of Leipzig (2008)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
8. Li, H., Nie, Z., Lee, W.-C., Giles, C.L., Wen, J.-R.: Scalable community discovery on textual data with relations. In: WWW, pp. 101–110 (2008)
9. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. JAIR 30, 249–272 (2007)
10. McPherson, M., Lovin, L.S., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27(1), 415–444 (2001)
11. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: CIKM, pp. 1203–1212 (2008)
12. Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint latent topic models for text and citations. In: KDD, pp. 542–550 (2008)
13. Neal, R.M.: Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics 9(2), 249–265 (2000)
14. Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455), 1077–1087 (2004)
15. Sun, Y., Han, J., Gao, J., Yu, Y.: Itopicmodel: Information network-integrated topic modeling. In: ICDM, pp. 493–502 (2009)
16. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
17. Wang, X., Mohanty, N., Mccallum, A.: Group and topic discovery from relations and text. In: LinkKDD, pp. 28–35 (2005)
18. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining link and content for community detection: A discriminative approach. In: KDD, pp. 927–935 (2009)