The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Robust Detection of Link Communities in Large Social Networks by Exploiting Link Semantics

Di Jin,[1] Xiaobao Wang,[1] Ruifang He,[1] Dongxiao He,[1] Jianwu Dang,[1,2] Weixiong Zhang[3,4]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China, [2] School of Information Science, Japan Advanced Institute of Science and Technology, Japan, [3] College of Math and Computer Science, Institute for Systems Biology, Jianghan University, Wuhan 430056, China, [4] Department of Computer Science and Engineering, Washington University, St. Louis, MO 63130, USA

{jindi, wxbxmt, rfhe, hedongxiao}@tju.edu.cn, jdang@jaist.ac.jp, weixiong.zhang@wustl.edu

Abstract

Community detection has been extensively studied for various applications, focusing primarily on network topologies. Recent research has started to explore node contents to identify semantically meaningful communities and interpret their structures using selected words. However, links in real networks typically have semantic descriptions, e.g., comments and emails in social media, supporting the notion of communities of links. Indeed, communities of links can better describe multiple roles that nodes may play and provide a richer characterization of community behaviors than communities of nodes. The second issue in community finding is that most existing methods assume network topologies and descriptive contents to be consistent and to carry the compatible information of node group membership, which is generally violated in real networks. These methods are also restricted to interpret one community with one topic. The third problem is that the existing methods have used top ranked words or phrases to label topics when interpreting communities. However, it is often difficult to comprehend the derived topics using words or phrases, which may be irrelevant. To address these issues altogether, we propose a new unified probabilistic model that can be learned by a dual nested expectation-maximization algorithm. Our new method explores the intrinsic correlation between communities and topics to discover link communities robustly and extract adequate community summaries in sentences instead of words for topic labeling at the same time. It is able to derive more than one topical summary per community to provide rich explanations. We present experimental results to show the effectiveness of our new approach, and evaluate the quality of the results by a case study.

Introduction

Social networking has become increasingly important for connecting people of diverse background. It is prevalent over the internet for geographically dispersed users. As a result, large quantities of network data, particularly in social sciences, have been accumulated. Analysis of such large quantities of data is able to help reveal underlying social structures and discern behavior and future trends.

Graph is the simplest form of a social network. It represents basic units as nodes and relationships between them as links. A growing interest in social networks has revived graph mining algorithms. An important problem in analyzing social networks is the problem of community detection (Girvan and Newman 2002). The primary objectives of this problem are to identify groups of nodes with common functions and to discover meaningful functional structures of such groups of nodes. A group of nodes form a dense region of closely related entities in a graph, and thus constitute a community. Finding such communities is an effective means to social network analysis, e.g., personalized recommendations and recognition of abnormal activities.

Most community detection methods use exclusively information of network topologies, as reviewed in (Fortunato and Hric 2016). These have hierarchical clustering (Girvan and Newman 2002), modularity-based methods (Newman and Girvan 2004), spectral optimization algorithms (Li et al. 2015), Markov dynamic algorithms (Rosvall et al. 2014), and statistical inference methods (He et al. 2015).

However, content information, e.g., descriptions of interactions among the entities in a network, may also provide useful information on network communities. Indeed, it has been shown in recent work that the use of node contents can significantly improve the quality of the resulting communities. The methods along this line have topic model-based methods (Zhao et al. 2012), generative models (Yang, McAuley, and Leskovec 2013), and heuristic methods (Ruan, Fuhry, and Parthasarathy 2013). Information of network topologies and node contents are also complementary to each other; if one is missing or inaccurate, the other can be used to make up for the missing or noisy data.

More importantly, node contents can also be used to discover interpretable community descriptions to help reveal the latent functions of individual communities. Communities with functional descriptions are desirable in practice and have attracted attention lately, e.g., in the latest works in (Wang et al. 2016; He et al. 2017).

While some progress has been made, the results of existing methods are far from satisfactory in several aspects. We first observe that a group of users in a social network may interact over more than one topic of common interest and subsequently form a community. They often exchange text messages, which are naturally represented by the links connecting them in the network. These links may cover more than one topic. That is, a community may have more than one topic, as illustrated in Figure 1(a). This observation helps reveal several limitations of existing methods.

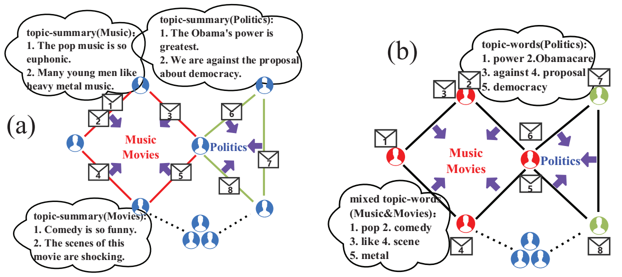

Figure 1: Illustration of a social media network, where nodes represent users and links represent messages among users. (a) The network is structured by our method that considers link contents and topologies. The red and green links form two communities. A community can have more than one topic and each topic is represented by an extracted summary that is more informative than individual words. (b) The network is structured by existing algorithms that combine node contents and topologies. The red and green nodes form two communities. Per community is related to one topic and each topic is represented by top-ranked words.

First, supported by our observation above, contents on links carry more information of community structures than contents on nodes. However, the existing methods only exploit node contents. In order to apply an existing algorithm to networks with messages on links, we may combine all messages sent by a user as the content of the node (Figure 1(b)). This conversion of link content to node content may lose information, e.g., addressees in Figure 1(b), and reduce the effectiveness of the method because the user may exchange messages with others in different communities, some of which may not even be consistent.

Second, it is usually assumed in the existing methods that network structures and node contents have the same information of node group memberships (i.e., communities and topical clusters being the same), which is often violated. For example, social relations in Twitter often directly reflect users' groups, whereas user-generated contents may be diverse. Therefore, when node contents do not match well with community structures, these algorithms' performance may deteriorate.

Third, the existing methods aim at finding one topic for one community, despite that communities in real social networks may have multiple topics, providing limited interpretability of resulting communities. For example, there are red and green user communities in Figure 1(b). The red community naturally has two topics, which are difficult to distinguish by the existing methods. Instead, these methods will interpret this community by a mixture of these two topics, which is difficult to comprehend.

Finally, the existing methods use individual words or short phrases to summarize communities, even though the text messages exchanged among users are typically complete sentences that have more information than individual words. It may not be straightforward to understand communities using a few words. Take Figure 1(b) as an example. It is difficult to appreciate the listed topics without knowing how the words used are related. It becomes worse when the top-ranked terms for the topics are also overlapped, e.g., the mixed topic of red community.

Although link contents are more informative than node contents and the former have unique characteristics that are missing in the latter (Ahn, Bagrow, and Lehmann 2010), no method has been developed to use both network topologies and link contents for finding link communities.

We developed a new probabilistic model for finding link communities with informative explanations. We developed a dual nested expectation-maximization (EM) algorithm to exploit network topologies, link contents and their intrinsic correlations. We like to highlight that our method addresses the four main problems discussed above. It does not assume that topologies and contents share the same community memberships, is able to interpret a community by more than one topic, and uses whole sentences to summarize communities, as illustrated in Figure 1(a).

The Model and Method

We first design a unified model for finding link communities and extracting their summaries by tightly integrating topologies and link contents. We learn the model via a dual nested EM algorithm. We then summarize the method and analyze its complexity. Table 1 shows the notations used.

The Probabilistic Model

Consider a network $G = (N, E)$ of n nodes $\{v_1, \cdots, v_n\}$ and e (undirected and unweighted) links that belong to a given number c of link communities. The network structure is represented by an adjacency matrix A = $(a_{ij})_{n\times n}$ with $a_{ij} = 1$ if a link exists between nodes v_i and v_j, or 0, otherwise. The link content is represented by a document-term matrix X = $(x_{ij,k})_{e\times m}$ with $x_{ij,k} = 1$ if the content of link <*i,j*> contains the kth word w_k of the dictionary, or 0 otherwise. We have the content of each link to be text of sentences, e.g., emails or messages, which are texts between corresponding users.

Table 1: The notations used in the paper

Signs	Descriptions
A, X	Adjacency matrix and document-term matrix
N, E	The set of nodes and edges
n, e	The number of nodes and edges
l, m	The number of sentences and words of the dictionary
c, k	The number of communities and topical clusters
w_k	The kth word of the dictionary
<*i,j*>	The link between nodes v_i and v_j
a_{i*}	One endpoint i of a possible link <*i,j*>, i.e., a half of the link
y_b	Empirical distribution over words specific to bth sentence
μ_b	The number of words in bth sentence
z_{ij}	Community associated with link <*i,j*>
$g_{ij,k}$	Topic associated with link-term pair <<*i,j*>, w_k>
$s_{ij,k}$	Sentence associated with link-term pair <<*i,j*>, w_k>
τ	Multinomial distribution over communities
ω_r	Multinomial distribution over nodes specific to rth community
ψ_r	Multinomial distribution over topics specific to rth community
φ_t	Multinomial distribution over sentences specific to tth topic

The Problems: Given a network G, our objectives are to 1) partition G into c link communities and k topical clusters based on network topologies and link content together, 2) explore the correlation between communities and topical clusters to combine the two to interpret each community using more than one topic, and 3) extract understandable topical summaries to describe communities. Our method can deal with cases of $c \neq k$ where the numbers of communities and topical clusters are different; nevertheless, for clarity and simplicity, we focus on the case of $c = k$.

We like to note *that each of these three problems is technically challenging and has attracted much attention individually; here, we consider to solve them altogether.*

To solve these issues, we divide link set E into c communities, called *link communities*. We partition E into c groups using link contents to form *topical clusters*, where each cluster is labeled by a topic with summary description, i.e., *topical summary*, composed of some top ranked sentences rather than words, extracted from the text on links. Meanwhile, we derive the intrinsic correlation between communities and topical clusters, and utilize this correlation to improve the results of community and topical cluster by incorporating the results of the other. We also utilize this correlation to interpret a link community using more than one dominate topical summary.

We cast the problems into building a unified generative model, illustrated in Figure 2. This model is specified by three types of quantities. The first is the observed quantities, which include the adjacency matrix A, document-term matrix X, and empirical distribution of observed sentences over words Y = $(y_{bk})_{l\times m}$, where $y_{bk} = p(x_{ij,k} = 1|s_{ij,k} = b)$ denotes the empirical probability that the bth sentence contains the kth word w_k. Therefore, $y_{bk} = 1/\mu_b$ if sentence b has the kth word, or 0 otherwise.

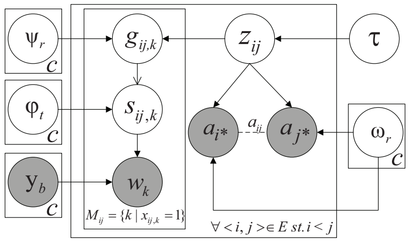

Figure 2: A schematic diagram of the integrative generative model for jointly solving three key community discovery problems. The symbols and notations used are summarized in Table 1.

The second type is the latent quantities, which include community assignments z, where z_{ij} denotes the label of the community which link <*i,j*> belongs to, the topic assignment g where $g_{ij,k}$ is the label of the topic which link-term pair <<*i,j*>, w_k> belongs to, and the sentence assignments s where $s_{ij,k}$ is the label of the sentence which <<*i,j*>, w_k> is expected to belong to.

The third type is model parameters, which include τ = $(\tau_r)_{1\times c}$, where $\tau_r = p(z_{ij} = r)$ is the prior probability that any link belongs to the rth community; $\Omega = (\omega_{rj})_{c\times n}$, where $\omega_{rj} = p(a_{j*} \mid z_{ij} = r)$ denotes the probability that the rth community selects node v_j as one of the two endpoints when it generates a link; $\Psi = (\psi_{rt})_{c\times c}$, where $\psi_{rt} = p(g_{ij,k} = t \mid z_{ij} = r)$ denotes the probability that the rth community selects the tth topic; and $\Phi = (\varphi_{tb})_{c\times l}$, where $\varphi_{tb} = p(s_{ij,k} = b \mid g_{ij,k} = t)$ is the probability that the tth topic selects the bth sentence. Then, the generative process of this model is:

For each node v_i:
  For each node v_j with $a_{ij} = 1$ and $i < j$
    a. Draw community assignment $z_{ij} \sim$ *Multinomial* (τ)
    b. Draw $a_{i*} \sim$ *Multinomial* ($\omega_{z_{ij}}$)
    c. Draw $a_{j*} \sim$ *Multinomial* ($\omega_{z_{ij}}$)
    d. For each of the kth term with $x_{ij,k} = 1$:
      i. Draw topic assignment $g_{ij,k} \sim$ *Multinomial* ($\psi_{z_{ij}}$)
      ii. Draw sentence $s_{ij,k} \sim$ *Multinomial* ($\varphi_{g_{ij,k}}$)
      iii. Draw term $w_k \sim$ *Multinomial* ($y_{s_{ij,k}}$)

Then, the likelihood that G is generated by the model is

$$P\left(\mathrm{A},\mathrm{X} \mid \tau,\Omega,\Psi,\Phi,\mathrm{Y}\right) = \sum\nolimits_{\mathrm{z,g,s}} P\left(\mathrm{A},\mathrm{X},\mathrm{z},\mathrm{g},\mathrm{s} \mid \tau,\Omega,\Psi,\Phi,\mathrm{Y}\right)$$
$$= \sum\nolimits_{\mathrm{z,g,s}} P(\mathrm{z} \mid \tau) P(\mathrm{A} \mid \mathrm{z},\Omega) P(\mathrm{g} \mid \Psi,\mathrm{z}) P(\mathrm{s} \mid \Phi,\mathrm{g}) P\left(\mathrm{X} \mid \mathrm{s},\mathrm{Y}\right) \quad (1)$$
$$= \sum\nolimits_{\mathrm{z,g,s}} \left(\begin{array}{l} \prod\limits_{i<j} \left(\tau_{z_{ij}}\right)^{a_{ij}} \prod\limits_{i<j} \left(\omega_{z_{ij},i}\omega_{z_{ij},j}\right)^{a_{ij}} \prod\limits_{i<j}\prod\limits_{k=1}^{m} \left(\psi_{z_{ij},g_{ij,k}}\right)^{x_{ij,k}} \\ \times \prod\limits_{i<j}\prod\limits_{k=1}^{m} \left(\varphi_{g_{ij,k},s_{ij,k}}\right)^{x_{ij,k}} \prod\limits_{i<j}\prod\limits_{k=1}^{m} \left(y_{s_{ij,k},k}\right)^{x_{ij,k}} \end{array} \right)$$

subject to $\sum_{r=1}^{c}\tau_r = 1$, $\sum_{j=1}^{n}\omega_{rj} = 1$ for $r = 1\ldots c$, $\sum_{t=1}^{c}\psi_{rt} = 1$ for $r = 1\ldots c$, and $\sum_{b=1}^{l}\varphi_{tb} = 1$ for $t = 1\ldots c$.

Eq. (1) has five parts. The first two are the fitting to network topology, the third is a set of induced probabilities of generating topics g in communities z with distribution Ψ, and the last two parts are the fitting to link contents.

Recall the problems to be solved. First, the last two parts of (1), for fitting to link contents, are similar to topic model. But here we introduce an additional layer of *hierarchy*, i.e., instead of making topics be distributions over terms or words, we assume that topic models are mixtures of some existing base language models. Here we use sentence language models as the base language model. One benefit of this assumption is that each topic is then represented by some selected sentences, instead of expressed by keywords. So it includes more language information for finding topical clusters, and we can use top ranked sentences as the topical summary to represent the topic of this cluster.

Furthermore, we treat link communities and topical clusters separately (i.e., communities may not correspond to topics), and use the correlation matrix Ψ, which can be regarded as a matrix of probabilities for transitions from communities to topical clusters. As a result, our model tightly integrates the structural and content features of the network. We can use the correlation matrix Ψ, combined with the topics derived, to interpret each community with more than one topical summary. Moreover, through the function of the induced projection of this correlation matrix, our model can also be more robust by incorporating the network topology and link contents, especially when communities and topical clusters do not match well.

Model Inference

We need to maximize the likelihood in (1) to best fit the given data. Since maximizing (1) directly is difficult, we instead maximize its log likelihood

$$L = \log \sum_{\mathrm{z,g,s}} P(\mathrm{z}\mid\tau)P(\mathrm{A}\mid\mathrm{z},\Omega)P(\mathrm{g}\mid\Psi,\mathrm{z})P(\mathrm{s}\mid\Phi,\mathrm{g})P(\mathrm{X}\mid\mathrm{s},\mathrm{Y}) \quad (2)$$

We adopt an expectation-maximization (EM) algorithm to maximize (2). By applying Jensen's inequality to (2), we obtain the expected log likelihood

$$\begin{aligned} L &\geq \bar{L} \\ &= \sum_{\mathrm{z}} \log \frac{\sum_{\mathrm{g,s}} P(\mathrm{z}\mid\tau)P(\mathrm{A}\mid\mathrm{z},\Omega)P(\mathrm{g}\mid\Psi,\mathrm{z})P(\mathrm{s}\mid\Phi,\mathrm{g})P(\mathrm{X}\mid\mathrm{s},\mathrm{Y})}{q(\mathrm{z})} \quad (3) \\ &= \sum_{i<j}^{n}\sum_{r=1}^{c} q_{ij,r}\left(\begin{array}{l} a_{ij}\left(\log\tau_r + \log\omega_{ri} + \log\omega_{rj}\right) + \\ \sum_{k=1}^{m} x_{ij,k}\log\left(\sum_{t=1}^{c}\psi_{rt}\sum_{b=1}^{l}\varphi_{tb}y_{bk}\right) - \log q_{ij,r} \end{array}\right) \end{aligned}$$

where q(z) is a distribution over community assignments z such that $\sum_{\mathrm{z}} q(\mathrm{z}) = 1$, $q_{ij,r} = \sum_{\mathrm{z}} q(\mathrm{z})\delta_{z_{ij},r}$ is the marginal probability within q(z) that link <i,j> belongs to the rth link community, and δ_{rs} is the Kronecker delta.

The maximum of $\bar{L}$ with respect to all of the possible choices of distribution q(z) will be obtained when $\bar{L} = L$, which, following Jensen's inequality, is when

$$q(\mathrm{z}) = \frac{\sum_{\mathrm{g,s}} P(\mathrm{z}\mid\tau)P(\mathrm{A}\mid\mathrm{z},\Omega)P(\mathrm{g}\mid\Psi,\mathrm{z})P(\mathrm{s}\mid\Phi,\mathrm{g})P(\mathrm{X}\mid\mathrm{s},\mathrm{Y})}{\sum_{\mathrm{z,g,s}} P(\mathrm{z}\mid\tau)P(\mathrm{A}\mid\mathrm{z},\Omega)P(\mathrm{g}\mid\Psi,\mathrm{z})P(\mathrm{s}\mid\Phi,\mathrm{g})P(\mathrm{X}\mid\mathrm{s},\mathrm{Y})} \quad (4)$$

Thus, the maximization of likelihood L with respect to τ, Ω, Ψ and Φ is equivalent to maximization of its lower bound $\bar{L}$ with respect to both q(z) (making $\bar{L} = L$) and the parameters. The EM algorithm for this double maximization is to repeatedly maximize with respect to first q(z) (i.e., the E-step) and then τ, Ω, Ψ and Φ (i.e., the M-step), proved to monotonically converge to local maxima.

For the E-step, we need to make $\bar{L} = L$. From (3) we can get the optimal $q_{ij,r}$'s by using

$$\begin{aligned} q_{ij,r} &= \sum_{\mathrm{z}} q(\mathrm{z})\delta_{z_{ij},r} = P\left(z_{ij} = r \mid \mathrm{A},\mathrm{X},\tau,\Omega,\Psi,\Phi,\mathrm{Y}\right) \\ &= \frac{\left(\tau_r\omega_{ri}\omega_{rj}\right)^{a_{ij}}\prod_{k=1}^{m}\left(\sum_{t=1}^{c}\sum_{b=1}^{l}\psi_{rt}\varphi_{tb}y_{bk}\right)^{x_{ij,k}}}{\sum_{r=1}^{c}\left(\tau_r\omega_{ri}\omega_{rj}\right)^{a_{ij}}\prod_{k=1}^{m}\left(\sum_{t=1}^{c}\sum_{b=1}^{l}\psi_{rt}\varphi_{tb}y_{bk}\right)^{x_{ij,k}}} \quad (5) \end{aligned}$$

However, the M-step is nontrivial because the expected log-likelihood $\bar{L}$ has two latent quantities, i.e., g and s.

M-Step with a Dual Nested EM Process

Now we need to maximize $\bar{L}$ in (3) over the parameters but a fixed $q_{ij,r}$. Maximization of τ and Ω is straightforward. Differentiating with respect to τ_r, subject to the normalization condition $\sum_{r=1}^{c}\tau_r = 1$, gives

$$\tau_r = \sum_{i,j=1}^{n} q_{ij,r}a_{ij} / 2e \quad (6)$$

Subsequently computing the derivative, setting the result to zero and satisfying $\sum_{j=1}^{n}\omega_{rj} = 1$ for $r = 1\ldots c$, the maximum with respect to ω_{rj} is obtained for

$$\omega_{rj} = \sum_{i=1}^{n} q_{ij,r}a_{ij} / \sum_{i,j=1}^{n} q_{ij,r}a_{ij} \quad (7)$$

Maximization with respect to Ψ and Φ is tricky. Only the second term in (3) depends on Ψ and Φ. But direct differentiation of this term yields an equation difficult to solve. We then apply Jensen's inequality to (3) again to have

$$\begin{aligned} &\sum_{i,j=1}^{n}\sum_{r=1}^{c} q_{ij,r}\sum_{k=1}^{m} x_{ij,k}\log\left(\sum_{t=1}^{c}\psi_{rt}\sum_{b=1}^{l}\varphi_{tb}y_{bk}\right) \\ &\qquad \geq \sum_{i,j=1}^{n}\sum_{r=1}^{c} q_{ij,r}\sum_{k=1}^{m} x_{ij,k}\sum_{t=1}^{c} p_{rk}^{t}\log\left(\psi_{rt}\sum_{b=1}^{l}\varphi_{tb}y_{bk} / p_{rk}^{t}\right) \end{aligned} \quad (8)$$

where p_{rk}^{t} may follow any distribution, with $\sum_{t=1}^{c} p_{rk}^{t} = 1$. Here we ignore the terms in $\bar{L}$ which can be regarded as constants with respect to Ψ and Φ.

The exact equality, and hence the maximum of the right-hand side, is achieved when

$$p_{rk}^{t} = \psi_{rt}\sum_{b=1}^{l}\varphi_{tb}y_{bk} / \sum_{t=1}^{c}\psi_{rt}\sum_{b=1}^{l}\varphi_{tb}y_{bk} \quad (9)$$

Thus, by the same argument as before, we can maximize the left-hand side of (8) by iteratively maximizing the right-hand side with respect to p_{rk}^{t} using (9) and with respect to Ψ and Φ by differentiation. Differentiation of (8), subject to $\sum_{t=1}^{c}\psi_{rt} = 1$, for $r = 1\ldots c$ and setting it to zero, the maximum with respect to Ψ is reached when

$$\psi_{rt} = \sum\nolimits_{k=1}^{m} p_{rk}^{t} \sum\nolimits_{i,j=1}^{n} q_{ij,r} x_{ij,k} \,/\, \sum\nolimits_{i,j=1}^{n} q_{ij,r} K_{ij} \qquad (10)$$

where $K_{ij} = \sum_{k=1}^{m} x_{ij,k}$.

However, when differentiating with respect to Φ, we still have latent quantities, i.e., sentences s. Again, we use a nested EM process, and apply Jensen's inequality to (8):

$$\begin{aligned} &\sum_{i,j=1}^{n}\sum_{r=1}^{c} q_{ij,r} \sum_{k=1}^{m} x_{ij,k} \sum_{t=1}^{c} p_{rk}^{t} \log\left(\sum_{b=1}^{l} \varphi_{tb} y_{bk}\right) \geq \\ &\qquad \sum_{i,j=1}^{n}\sum_{r=1}^{c} q_{ij,r} \sum_{k=1}^{m} x_{ij,k} \sum_{t=1}^{c} p_{rk}^{t} \sum_{b=1}^{l} h_{tk}^{b} \log\left(\varphi_{tb} y_{bk} / h_{tk}^{b}\right) \end{aligned} \qquad (11)$$

where h_{tk}^{b} can be any distribution, subject to $\sum_{b=1}^{l} h_{tk}^{b} = 1$. Here we ignore the terms in right-hand of (8) which can be regarded as constants with respect to Φ.

The exact equality, and hence the maximum of the right-hand side, is achieved when

$$h_{tk}^{b} = \varphi_{tb} y_{bk} \,/\, \sum\nolimits_{b=1}^{l} \varphi_{tb} y_{bk} \qquad (12)$$

As before, we can maximize the left-hand side of (11) by repeatedly maximizing the right-hand side with respect to h_{tk}^{b} using (12) and with respect to Φ by differentiation.

Similarly, differentiating with respect to φ_{tb} , subject to $\sum_{b=1}^{l} \varphi_{tb} = 1$ for *t* = 1…*c*, gives

$$\varphi_{tb} = \frac{\sum_{k=1}^{m} h_{tk}^{b} \sum_{r=1}^{c} p_{rk}^{t} \sum_{i,j=1}^{n} q_{ij,r} x_{ij,k}}{\sum_{k=1}^{m} \sum_{r=1}^{c} p_{rk}^{t} \sum_{i,j=1}^{n} q_{ij,r} x_{ij,k}} \qquad (13)$$

Then, the optimal Ψ and Φ can be calculated by iterating (9) to (13) from a random initial seed until convergence.

The Dual Nested EM Algorithm

We now summarize the dual nested EM algorithm proposed above in Algorithm 1. When we have the optimal $q_{ij,r}$'s, Ψ and Φ, we can use $q_{ij,r}$'s, where $q_{ij,r}$ is the posterior probability that link <*i,j*> belongs to the *r*th community, to find the communities of links. We can then use Φ, where φ_t is the distribution of sentences specific to topic *t*, to derive a summary of topic for each topical cluster. We can further use the correlation matrix Ψ, where ψ_r is the distribution of the topical clusters specific to the *r*th community, to find the dominant topical summary for each community.

We now turn to the complexity of the algorithm by taking into account of the sparsity of the data matrices A, X and Y. First, the time to update $q_{ij,r}$'s, τ and Ω once via (5), (6) and (7) is ce+flc^2, 2*ec* and 2*ec* respectively, on networks of *e* links, *c* communities as well as *c* topical clusters, where $f = \sum_{i<j}^{n} K_{ij}$ is the number of terms and $K_{ij} = \sum_{k=1}^{m} x_{ij,k}$ the number of terms of link <*i,j*>. Then, the time to compute p_{rk}^{t} 's and Ψ once via (9) and (10) is lmc^2 and $2fc^2$, respectively. The time to compute h_{tk}^{b} 's and Φ once via (12) and (13) is *cml* and $2flc^2$; the time to compute the likelihood function once is $2flc^2$. The overall time complexity of the algorithm is O(ce+flc^2), which is nearly linear on large sparse networks with link contents.

Alg. 1: Dual nested EM algorithm

```
Input: A, X and Y; Output: $q_{ij,r}$'s, Ψ and Φ
Make an initial guess (for instance at random) for
the values of τ, Ω, Ψ and Φ
For $t_1$ = 1: $T_1$ //main EM
  Update marginal probabilities $q_{ij,r}$'s using (5)
  Update τ and Ω using (6) and (7)
  For $t_2$ = 1: $T_2$ //nested EM
    Update $p_{rk}^{t}$ 's and Ψ using (9) and (10)
    For $t_3$ = 1: $T_3$ //dual nested EM
      Update $h_{tk}^{b}$ 's and Φ using (12) and (13)
```

Experiments

We evaluated our algorithm in comparison with eight state-of-the-art community detection methods on two types of real datasets and on a case study.

Datasets

The following two datasets were used in our experiments.

Enron Email Dataset (Qi, Aggarwal, and Huang 2012):

This dataset is about a legal investigation of the Enron corporation financial scandal in 2000. It contains a large number of emails among employees. The dataset totally has 200,399 messages from 158 senior managers of Enron. Each manager may have more than one email account. We use this dataset to construct a network of email accounts (nodes in the network) linked by their email messages (edges). This means that the links are associated with the content of email messages. A useful characteristic of this dataset is that its subsets have been annotated by students at UC Berkeley, focusing on business-related emails and California Energy Crises. This was very useful for evaluation purposes. We selected a subset of 1,557 emails with thematic features labeled by 11 categories (i.e., regulations, internal projects, company image, political influence, California energy crisis, internal company policy and operations, alliances and partnership, legal advice, talking points, meetings, trip reports) as did in (Zhao et al. 2012). If there is an email between two accounts, a link labeled by the category of this email is introduced. (Multiple emails between two accounts are taken as multi-links, also supported by our model.) After removing common stop words and stemming, the dataset contains 974 nodes, 1,557 links, 31,563 sentences, as well as 15,382 words.

The Reddit Datasets (Wang, Lai, and Philip 2014):

We used three Reddit datasets corresponding to three days respectively, from August 25 to August 27 in 2012. Each Reddit dataset contains the threads of 3 sub-forums (i.e., Movies, Politics and Science) in www.reddit.com. Each user can submit a post and the other users may comment on the post by replying to the post. If one user v_i reports a post, and another user v_j comments on this post, there will be a link between v_i and v_j. Link <*i*,*j*> is then labeled by the post's category *z* which can be also regarded as the category of this post-comment pair. After preprocessing by removing common stop words and stemming, the dataset of August 25 contains 1,314 users, 1,339 links, 3,273 sentences and 4,616 words; that of August 26 contains 1,590 users, 1,714 links, 3,952 sentences and 5,055 words; and that of August 27 contains 2,143 users, 2,290 links, 5,794 sentences as well as 6,635 words.

Performance Metrics

As mentioned earlier, the datasets we used are associated with class labels in addition to link contents. These class labels are associated with links. To compare the effectiveness of our method with baselines equitably (see next subsection), we need to convert class labels on links into class labels on nodes by assuming that the two nodes of a link belong to the same community as did in (Wang, Lai, and Philip 2014). The converted class labels are overlapped. So we need metrics for overlapping communities of nodes.

Two metrics were used in experiments, the F-score and Jaccard similarity (Yang, McAuley, and Leskovec 2013), i.e., we evaluated the detected communities C with ground-truth communities C^* by $\frac{1}{2|C^*|}\sum_{C_i^* \in C^*} max_{C_j \in C} \delta\left(C_i^*, C_j\right) + \frac{1}{2|C|}\sum_{C_j \in C} max_{C_i^* \in C^*} \delta(C_i^*, C_j)$, where $\delta(C_i^*, C_j)$ is a similarity measure (F-score or Jaccard) between C_i^* and C_j.

Community Detection Results

We compared our method with eight state-of-the-art methods. We included two topology-based methods, 1) LMBP (He et al. 2015) and 2) BigCLAM (Yang and Leskovec 2013). LMBP is similar to the network topology part of our method, which is to detect link communities without information of link contents. BigCLAM uses topology structures to partition nodes into overlapping communities. We also selected two content-based approaches: 3) SMR (Hu et al 2014) and 4) GibbsLDA (Phan and Nguyen 2007), which clusters texts on links to form link communities. Finally, we considered four approaches: 5) Circles (McAuley and Leskovec 2012), 6) CESNA (Yang, McAuley, and Leskovec 2013), 7) SCI (Wang et al. 2016) and 8) NEMBP (He et al. 2017) that use both network topology and node contents. Circles and CESNA are algorithms for detecting overlapping community while SCI and NEMBP are for non-overlapping ones. To compare with these four methods, we need an equivalent way for modeling the content at the nodes as opposed to the links for the same scenario. For Enron dataset, the content at a node is the concatenation of all emails sent by the participant corresponding to the node. For Reddit datasets, the content at a node is the union of comments posted by the user.

The programs of all methods compared were obtained from their authors, and we used their default parameters. For our method, we set the numbers of iterations to T_1 = 40, T_2 = 10 and T_3 = 3, making the method converge to stable results. Besides, because all of these algorithms converge to local minima, we ran each algorithm 20 times and report the result with the highest likelihood. The number of communities for each of the networks considered, which is required by all methods compared as an input parameter, was set to the ground-truth of the number of communities.

In order to compare all the methods in the same framework, we transform the results of LMBP, SMR, GibbsLDA as well as our method into communities of nodes by considering that an induced community for a set of links is the set of vertices corresponding to the end points of all edges in the community. So the results of all methods compared are communities of nodes and then the transformed class labels on nodes are introduced to evaluate the effectiveness of all of the methods using F-score and Jaccard index.

Our method outperforms all of the methods compared in terms of the two performance metrics F-score and Jaccard (Table 2). The results can be summarized as follows.

Table 2: Comparison of methods in F-score and Jaccard. R8.25 denotes Reddit of August 25, 2012. Bold denotes best result.

Metrics (%)	Datasets	Datasets								
		LMBP	BigCLAM	SMR	LDA	Circles	CESNA	SCI	NEMBP	Ours
F-score	Enron	43.69	18.90	44.14	36.88	45.22	30.15	37.65	36.59	**50.29**
	R8.25	52.60	20.36	42.42	35.54	50.23	34.88	46.93	51.84	**54.19**
	R8.26	48.66	24.29	47.40	40.20	51.08	33.96	49.16	45.41	**53.47**
	R8.27	51.94	17.81	43.03	45.23	52.55	27.81	49.48	50.78	**55.48**
Jaccard	Enron	30.21	10.92	31.06	23.24	32.11	20.21	25.96	25.66	**37.89**
	R8.25	37.62	12.63	30.21	22.49	36.09	25.93	33.65	37.08	**40.74**
	R8.26	32.52	16.33	35.99	26.50	37.28	21.53	34.64	31.57	**42.02**
	R8.27	35.44	10.58	29.96	30.28	38.20	17.15	34.87	37.30	**40.48**

(1) Our algorithm outperforms both the topology-based and the content-based methods. In particular, our algorithm is better than LMBP, which is similar to the part of fitting network topology of our method. It means that link contents often contain useful information to find communities.

(2) The algorithms that combine link contents and topology information perform better than that combine node contents and topologies. Besides, the methods that utilize node contents and topologies may sometimes be worse than the topology-based or the content-based algorithms, as the results on Reddit of August 25, 2012 show (Table 2). This may be because in networks of social media, contents are more naturally associated with links than nodes. But the algorithms that combine node contents and topologies, e.g., Circles and SCI, concatenate the link contents together to create node contents. This conversion often reduces their effectiveness, due to loss of information when mixing contents from diverse links. As a result, this may lower the discriminative power of methods to different communities.

(3) The superior performance of our method may also be due to another factor. It is often believed that users tend to communicate frequently over certain topical interests and

then form a community, and our model better depicted this phenomenon in three ways. 1) We considered the interactive information in networks as the content on links and search for link communities; 2) we did not assume that the topologies and contents share the same group memberships, making it flexible and robust; and 3) we used text sentences directly in base language model which may introduce rich information to enhance the results.

A Case Study

We carried out a case study to further evaluate the effectiveness of our method in discovering structurally and topically meaningful communities with more than one topical summary. We used the Reddit dataset of August 27, 2012 as an example and SCI as the baseline algorithm to contrast ours, since no other method has been developed to use summary topics for community interpretation. The true communities of every comment in Reddit are labeled by one of the 3 sub-forums, i.e., Movies, Politics and Science.

We used the default parameters of SCI and set the number of community to the number of 3 true communities. After one run of SCI, it found three communities, each of which had one topic represented by top-ranked words. We show in Figure 3 one example of semantic words for one community from SCI. For this community, we select top ten words. The community vaguely shows that this covers a group of "Politics" and a group of "Bank". Note that while words "scandal" and "libor" are all related to bank, they are also a part of "Politics" topic.

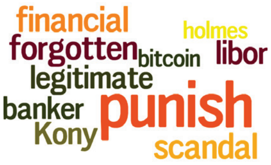

Figure 3: Word clouds for a community discovered by the baseline SCI. Top ten words of this sampled community are shown here. More relevant a word, larger it is in the figure.

However, the above result has two problems. First, the three communities in the Reddit network are in fact tightly connected. For instance, there exist many links connecting the "Movies" and "Politics" communities via threads discussing *political movies*. The single community topic in Figure 3 also has words "Kony" and "Holmes" that are related to "Movies". It means that the topic learned by SCI is a mixture of topics "Politics" and "Movies". Therefore, each community in this network may contain more than one topic that the users discussed, which cannot be discovered by SCI. As a result, mixing multiple topics into one is not an ideal way to interpret a community, especially when it has more than one topic. More importantly, after obtaining a list of top-ranked words, SCI may still require a manual summarization of the words to derive a more accurate description of a topic, because it is nontrivial to appreciate this topic without knowing how these words are related.

Remarkably, our new method found one community with two dominant topics and 2 communities with one dominant topic each. Here we only show the community with two topics as an example (which roughly corresponds to the community found by SCI shown in Figure 3). As shown in Table 3, the "Politics" members discuss topics on, e.g., *options of gays*, *OSHA fine*, and *AAP policy statement*, and the "Movies" members discuss various movie related subjects, e.g., the scenes in *Die Welle, Elf*, and *Torn Curtain*; the result even give a link to knife fight scene on YouTube. Obviously, the results from our method are easier to understand than those from SCI. This clearly showed that it is indeed beneficial to use more than one topic with summary description to characterize a community.

Table 3: An example of a community discovered by our method with two topics which are represented by some top ranked sentences, as summaries.

Topic name	The top sentences of each topic
Politics	Mere hoping that all straight folks everywhere one day realize that antigay ravers come in just two flavors…and assholes who are attempting to compensate for and/or draw attention away from their own moral shortcomings.
	An OSHA fine of only $2000 most likely means the company did all they could to reasonably protect the worker, but the worker…
	I'll start with another quote from the AAP policy statement: Systematic evaluation of English-language peer-reviewed literature from 1995 through 2010 indicates …
	In the UK people are being arrested for insulting people over Twitter, rioters are thrown in jail for 10 years for stealing shoes.
	One of my other favorite examples was how, after BP had exploded a deluge of oil all over the Gulf of Mexico, they somehow managed to lean on a variety of local gov't agencies to block off sections of the beach to prevent even just reporting of the disaster.
Movies	Die Welle - A film about "a high school teacher's unusual experiment to demonstrate to his students what life is like under a dictatorship".
	Until moviegoers stop handing over hard earned money for broad, unoriginal movies and we see more studio push backs and delays which result in major financial losses.
	As to my own thoughts on the current state of the industry major American films have been steadily declining in quality and content since 1975.
	Elf- yes, there are moments where he seems to be the stereotypical Will… that it (for the most part) is just about the voice, so the performance is all in that.
	In all seriousness, though, Torn Curtain's killing of Gromek was insanely brutal for its time, and is still pretty nasty - though the woman dashing about for blunt objects is a little bit Looney Toons.
	Here's a link to the end knife fight scene: https://www.youtube.com/watch?v=1izw6ZsPSp0 EDIT: That clip is dubbed and sounds real shitty, but the point is there haha

In addition, using the matrix multiplication of Φ and Y, our method can also derive the multinomial distribution of each topic over words, and then return the top-ranked words as topic words. The word-clouds of the two topics of this community which correspond to that in Table 3 are shown in Figure 4. We may roughly understand this community with a quick read of the figure than reading topical summaries. Then, a better way of our method may be to consider interpreting topics in a *hierarchical way*. For each topic, we first offer only the topic words, which may let users quickly grasp the overall meaning of this topic. If users are still unclear about the topic with individual words, they may move onto topical summaries for better understanding. Besides, the two topics in this community, which are "Politics" and "Movies", are more pure and comprehensible than the only one topic derived by SCI.

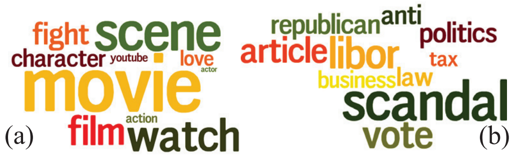

Figure 4: Word clouds for a community with two topics inferred by our method, which can be taken as an auxiliary of summaries.

Conclusion and Discussions

We proposed a new probabilistic model for link community detection that explores link contents, and developed a dual nested EM algorithm for learning the model. The new algorithm was developed particularly to address the four critical problems in the current research on community detection, i.e., 1) text contents in social networks are often associated with links rather than with nodes, and hence may form multiple communities of links; 2) network topologies and link contents may not share the same information of community memberships, so that a community may have more than one topic; 3) when contents do not match well with network communities, the results of detected communities deteriorate even with this additional node contents; and 4) it is desirable to have more contextual summaries to better understand the topics of a community. For these issues, our novel method was designed to discover link communities, extract summaries for topic labeling, and explore the intrinsic correlation of communities and topics, all at the same time. By exploiting this correlation, our model can not only combine network topologies and contents to accurately identify link communities, but also interpret each community using more than one topical summary if necessary, providing richer explanations. We evaluated the new method on two types of real networks, where it outperformed eight state-of-the-art existing methods.

In this paper we focused mainly on how to design the model more accurately, while one can get the number of communities c via model selection when c is unknown. We considered cases where the number of communities c and topics k are the same because the datasets we used set those two to be equal. Our method is also suitable when $c \neq k$.

Acknowledgment

The work was supported by Natural Science Foundation of China (61772361, 61502334, 61673293, 61472277). Correspondence should be addressed to R.H. (rfhe@tju.edu.cn).

References

Ahn, Y.; Bagrow, J.P.; and Lehmann, S. 2010. Link communities reveal multiscale complexity in networks. *Nature*. 466: 761-764.

Fortunato, S; and Hric, D. 2016. Community detection in networks: A user guide. *Phys. Rep.* 695: 1-44.

Girvan, M.; and Newman, M. E. J. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99: 7821-7826.

He, D.; Feng, Z.; Jin, D.; Wang X.; and Zhang W. 2017. Joint Identification of Network Communities and Semantics via Integrative Modeling of Network Topologies and Node Contents. In *Proceedings of AAAI'17*, 116-124. Palo Alto, California, USA: AAAI Press.

He, D.; Liu, D.; Jin, D.; and Wang Z. 2015. A stochastic model for detecting heterogeneous link communities in complex networks, In *Proceedings of AAAI'15*, 130-136. Palo Alto, California, USA: AAAI Press.

Hu, H.; Lin, Z.; Feng. J.; and Zhou J. 2014. Smooth representation clustering. In *Proceedings of CVPR'14*, 3834-3841. Piscataway, NJ, USA: IEEE Press.

Li, Y.; He, K.; Bindel, D.; and Hopcroft, J. E. 2015. Uncovering the small community structure in large networks: A local spectral approach. In *Proceedings of WWW'15*, 658-668. New York, NY, USA: ACM Press.

McAuley J. and Leskovec J. 2012. Learning to discover social circles in ego networks. In *Proceedings of NIPS'12*, 539-547, Massachusetts, USA: MIT Press.

Newman, M. E. J.; and Girvan, M. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69(2): 026113.

Phan, X.-H. and Nguyen, C.-T. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). http://gibbslda.sourceforge.net/.

Qi, G.; Aggarwal, C. C.; and Huang, T. 2012. Community Detection with Edge Content in Social Media Networks. In *Proceedings of ICDE'12*, 534-545. Piscataway, NJ, USA: IEEE Press.

Rosvall, M.; Esquivel, A. V.; Lancichinetti, A.; West, J. D.; and Lambiotte. R. 2014. Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*. 5(1): 4630.

Ruan, Y.; Fuhry, D.; and Parthasarathy, S. 2013. Efficient community detection in large networks using content and links. In *Proceedings of WWW'13*, 1089-1098. New York, NY, USA: ACM Press.

Wang, X.; Jin, D.; Cao, X.; Yang L.; and Zhang W. 2016. Semantic community identification in large attribute networks. In *Proceedings of AAAI'16*, 172-178. Palo Alto, California, USA: AAAI Press.

Wang, C.; Lai, J.; and Philip, S. Y. 2014. NEIWalk: Community discovery in dynamic content-based networks. *IEEE Trans. Knowl. Data Eng.* 26(7): 1734-1748.

Yang, J.; and Leskovec, J. 2013. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of WSDM'13*, 587-596. New York, NY, USA: ACM Press.

Yang, J.; McAuley, J.; and Leskovec, J. 2013. Community detection in networks with node attributes. In *Proceedings of ICDM'13*, 1151-1156. Piscataway, NJ, USA: IEEE Press.

Zhao, Z.; Feng, S.; Wang, Q.; Huang, J. Z.; Williams, G. J.; and Fan, J. 2012. Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*. 26: 164-173.