Community Detection in Social Networks Considering Topic Correlations

Yingkui Wang,[1] Di Jin,[1,*] Katarzyna Musial,[2] Jianwu Dang[1,3]
[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China,
[2]Advanced Analytics Institute, School of Software, University of Technology Sydney, Australia,
[3]School of Information Science, Japan Advanced Institute of Science and Technology, Japan
{ykwang, jindi}@tju.edu.cn, katarzyna.musial-gabrys@uts.edu.au, jdang@jaist.ac.jp

Abstract

Network contents including node contents and edge contents can be utilized for community detection in social networks. Thus, the topic of each community can be extracted as its semantic information. A plethora of models integrating topic model and network topologies have been proposed. However, a key problem has not been resolved that is the semantic division of a community. Since the definition of community is based on topology, a community might involve several topics. To achieve better community detection results and to better understand the fundamental community semantics, we investigate the correlations of different topics in community detection model. This work models the formation of each edge assuming that users are more likely to communicate with each other when they are in the same community and their topics are closely correlated. A Topic Correlations based Community Detection (TCCD) model is proposed, which can learn community structure and semantic interpretation of each community. Our model is evaluated on two real datasets and is compared with four state-of-the-art methods. Experimental results show that TCCD significantly improves the accuracy of community detection. Finally, a case study shows that TCCD can detect the topic correlations inside a community. And we can infer better semantic interpretation of each community.

Introduction

In recent years, research in the area of community detection in networks has become a hot topic (Newman 2006; Fortunato and Hric 2016). This is due to the fact that communities play a very important role in a network and they enable to understand and interpret networks functions and characteristics. Community is defined as a group of nodes that are densely connected internally (Girvan and Newman 2002). Recent community detection methods not only detect community structures but also identify semantics of communities (He et al. 2017a; Zhang et al. 2018). It is significant to understand the innate character of communities as we can learn what users are interested in, what they care about in a community, and how the topic of communities evolves.

In real social networks, e.g., Weibo, Twitter, and Facebook, users interact with each other talking about different topics. Networks are created based on a large amount of heterogeneous and complex contents, such as microblogs, tweets, and posts. This type of information is considered as node contents or link contents depending on whether it is connected with nodes or links. To understand what topics are connected with a given community, such contents need to be analyzed and used as integral part of community detection process. Approaches that integrate network topologies and node contents have been proposed (Mcauley and Leskovec 2014; Pei, Chakraborty, and Sycara 2015). Recent studies begin to investigate community level diffusion, i.e., modeling diffusion patterns of topics across different communities (Hu et al. 2015). The work in (Cai et al. 2017), for the first time, formalizes the concept of community profiling, which is to characterize the intrinsic nature and extrinsic behavior of a community. Community structure is also incorporated into network embedding methods (Tu et al. 2018).

However, several issues have not been well resolved by existing methods. By analyzing a large number of social networks, beyond the observation that a community might focus on several topics (Jin et al. 2018), we further found that there are correlations between the topics, which significantly affect community structures. Users focusing on a topic might have great interests in interacting with others focusing on a different topic, which means that these two topics are highly correlated. While there are also opposite situations between two topics, which means that the topics are minorly correlated. Take paper co-authorship network as example. Suppose that the topics in a network include *Machine learning*, *Image processing* and *Data mining*. The correlation between *Machine learning* and *Image processing* is closer than that between *Image processing* and *Data mining*. To characterize the correlations among latent topics, some studies have made great efforts, such as correlated topic models (Blei and Lafferty 2006; Chen et al. 2013) which replace Dirichlet distribution with logistic normal distribution in LDA, and topic embedding (Li et al. 2016; Jiang et al. 2013; He et al. 2017b; Li et al. 2018a) which represents topics in a low-dimension space. However, none of existing works consider the factor above in community detection. Existing methods are limited to resolve following issues corresponding to the observation.

First, topics are inferred from network contents including node contents and edge contents. While edge contents are

*Corresponding author

responsible for the formation of correlations between topics, node contents make no contributions since they are isolated documents that never generate edges. So, node contents and edge contents should be both considered to infer topics. Meanwhile, they should be utilized in different components corresponding to topics and topic correlations respectively. But, none of existing methods integrate node contents and edge contents into a unified model to infer topics and their correlations.

Second, according to our observation, even though two users focus on two different topics, they are still more likely to interact with each other when those two topics are highly correlated. So, topic correlations have significant effects on the generation of edges and further affect community structures. While, existing methods assume that interactions always occur between users who share the same topics in a community, which ignores the principle of generating edges according to topic correlations.

Third, understanding the fundamental semantics of communities is a challenge. Currently, most methods only detect the topics of communities as their semantics. Then they use top-ranked words to represent topics. The work in (Jin et al. 2018) uses whole sentences to interpret communities. In fact, community semantics are far beyond above aspects. Since a community might focus on several topics, what is the mechanism of the composition of topics inside communities? This question leads us to understand communities in a natural way. But, existing work cannot resolve the question.

Based on above discussions, we propose a generative model for community detection which consists of three components. The first part generates all user contents based on users community memberships and their topics. The second generates all link contents based on two endpoint users community memberships and their topics. The third generates each directed link with users community memberships and topic correlations together. Beyond existing work, our work for the first time interprets the mechanism of the composition of topics inside communities and understand communities in a natural way.

Topic Correlations based Community Detection

Problem Formulation

The notations used in this paper are summarized in Table 1.

Definition 1. A **social network** is defined by $G = (U, E, D)$, where U is a set of users. A user is presented by $u \in U$. E is a set of directed links and D is a set of documents published by users. A directed link is denoted by (i, j) that is from user i to user j. We allow multiple edges to exist between two users. If user i replies to user j multiple times, then there will be multiple edges from user i to user j. Each link is associated with a document and W_{iq} denotes the word list of the *q*-th link document of user i.

Definition 2. A user i's **community membership** is defined by a $|C|$ dimensional vector π_i. $|C|$ is the number of communities. For a community c, element $\pi_{i,c}$ represents the probability of belonging to community c.

Table 1: Notations

Notations	Descriptions
U, K, C, T	set of users, topics, communities and time stamps
W	word set of vocabulary
D_i	posts not on links sent by user i
$E_i, e_{ii'}$	links sent by user i, directed link from user i to i'
W_{ij}, W_{iq}	word list of the *j*-th post, the *q*-th link of user i
W_{ijl}, W_{iqr}	the *l*-th and the *r*-th word of W_{ij}, and W_{iq}
π_i	multinomial distribution over communities of user i
θ_c	multinomial distribution over topics of community c
ϕ_k	multinomial distribution over words of topic k
ψ_{kc}	multinomial distribution over time of topic k and community c
c_{ij}, g_{iq}	community indicator of post and link
z_{ij}, y_{iq}	topic indicator of post and link
t_{ij}, t_{iq}	time stamp of post and link
$\eta_{gy,g'y'}$	the probability of forming a link between community g with topic y and community g' with topic y'
I_i	user i's posting preference
$\alpha, \beta, \epsilon, \rho$	Dirichlet priors

Definition 3. A **topic** k is defined by a $|W|$ dimensional vector ϕ_k following a multinomial distribution over vocabulary. For a word w, the element ϕ_{kw} represents the probability of belonging to topic k. The number of topics is $|K|$.

Definition 4. **Topic distribution** of a community c is defined by a $|K|$ dimensional vector θ_c. An element θ_{ck} represents the probability of belonging to topic k.

Definition 5. **Time stamp distribution of community and topic** are defined by a $|T|$ dimensional vector ψ_{kc}, $c \in C$, $k \in K$.$|T|$ is the number of time stamps. It is a multinomial distribution over time stamps.

Definition 6. **Topic correlation** $\eta_{gy,g'y'}$ defines the correlation between two topics in different communities. It reflects the tendency of forming a link between user i who is in community g and focus on topic y and user j who is in community g' and focus on topic y'.

Model Structure

We design a generative model to properly generate network topology, link contents and node contents. The probabilistic graphical model of TCCD is shown in Fig.1. It includes three main components: a) User post component; b) Link content component; c) Link component.

User post component. Take a forum network for example, the posts submitted by users are considered as node contents or link contents. Those posts that are never replied by others are processed as node contents. This component has no relation with network topology. But a user's posts have deep relation with his latent community membership and

community topics. We generate them on the basis of serval latent factors, i.e., users community membership distribution, community-topic distribution and topic-word distribution.

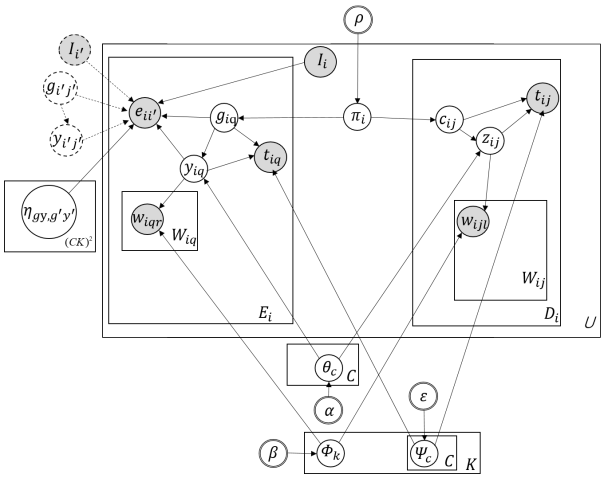

Figure 1: The graphical representation of our model. The notations are summarized in Table 1.

Link content component. This component generates all link contents of anetwork. The basic idea is that link contents reflect what topics the two users are talking about. Though the link structure of the network plays the key role for community detection, the link contents also provide rich information to the forming of community structure and community topics. It is the basic principle that users in the same community and interested in the same topics are more likely to interact.

Link component. Suppose that user i replies to a post of user i' at time stamp t. Then we formulate the generation of the directed link $E^t_{ii'}$ from user i to user i'. Let I_i and $I_{i'}$ represent out-going tendency and in-coming tendency of users i and i' respectively. They are calculated by $\frac{(out-degree)_i}{(degree)_i}$ and $\frac{(in-degree)_{i'}}{(degree)_{i'}}$. User's in-degree indicates how many times others reply him while out-degree indicates how many times he replies to others. The value of $\eta_{gy,g'y'}$ represents the tendency of forming a link from user i to user i' where g, y, g' and y' equal to g_{iq}, y_{iq}, $g_{i'q}$ and $y_{i'q}$ respectively as described in link contents generation component part. In order to integrate $\eta_{gy,g'y'}$, I_i and $I_{i'}$, we define ω_{ij} as follows:

$$\omega_{ij} = \eta_{gy,g'y'} + I_i \cdot I_{i'}. \tag{1}$$

Finally, we utilize sigmoid function to generate this link.

$$\begin{aligned} &P(E^t_{ii'} = 1 | I_i, I_{i'}, g_i, g_{i'}, y, y', \eta) \\ =&\sigma(\omega_{ij}) \\ =&1/(1+\mathrm{e}^{-\omega_{ij}}). \end{aligned} \tag{2}$$

It is hard to infer a Gibbs sampler for this model. So, we adopt Pólya-Gamma distribution to model a logistic function (Polson, Scott, and Windle 2013).

$$\frac{1}{1+\mathrm{e}^{-\omega_{ij}}} = \frac{1}{2}\int_0^\infty \varphi(\omega_{ij}, \xi_{ij}) P(\xi_{ij}) d\xi_{ij}, \tag{3}$$

where $\varphi(\omega_{ij}, \xi_{ij}) = e^{(\omega_{ij} - \xi_{ij}\omega_{ij}^2)/2}$ and $\xi_{ij} \sim PG(1, 0)$. Then we derive a joint probability distribution:

$$P(E^t_{ii'} = 1, \xi_{ij}) = \frac{1}{2}\varphi(\omega_{ij}, \xi_{ij}) P(\xi_{ij}|1, 0) \tag{4}$$

Generative process.We summarize above generative process as follows.

1. For each topic $k = 1, 2, ..., K$,
   (a) Sample the words distribution from a Dirichlet prior: $\phi_k \mid \beta \sim Dir(\beta)$;
   (b) For each community $c = 1, 2, ..., C$,
      i. Sample the distribution over time stamps from a Dirchlet prior: $\psi_{kc} \mid \epsilon \sim Dir(\epsilon)$
2. For each community $c = 1, 2, ..., C$,
   (a) Sample the distribution over topics from a Dirichlet prior: $\theta_c \mid \alpha \sim Dir(\alpha)$;
3. For each user $i = 1, 2, ..., U$,
   (a) Sample his community distribution from a Dirichlet prior: $\pi_i \mid \rho \sim Dir(\rho)$;
   (b) For each post $j = 1, 2, ...$,
      i. Sample community indicator from a Multinomial distribution: $c_{ij} \mid \pi_i \sim Mul(\pi_i)$;
      ii. Sample topic indicator from a Multinomial distribution: $z_{ij} \mid \theta_{c_{ij}} \sim Mul(\theta_{c_{ij}})$;
      iii. For each word $l = 1, 2, ...$,
         • Sample word from a Multinomial distribution: $w_{ijl} \mid \phi_{z_{ij}} \sim Mul(\phi_{z_{ij}})$;
      iv. Sample time stamp $t_{ij} \mid \psi_{z_{ij}c_{ij}} \sim Mul(\psi_{z_{ij}c_{ij}})$;
   (c) For each link $q = 1, 2, ...$,
      i. Sample community indicator from a Multinomial distribution: $g_{iq} \mid \pi_i \sim Mul(\pi_i)$;
      ii. Sample topic indicator from a Multinomial distribution: $y_{iq} \mid \theta_{g_{iq}} \sim Mul(\theta_{g_{iq}})$;
      iii. Sample the link from i to i': $E^t_{ii'} \mid I_i, I_{i'}, g_{iq}, g_{i'q}, y_{iq}, y_{i'q}, \eta \sim Ber(\sigma(\eta_{g_{iq}y_{iq},g_{i'q}y_{i'q}} + I_i \cdot I_{i'}))$;
      iv. For each word $r = 1, 2, ...$,
         • Sample word from a Multinomial distribution: $w_{iqr} \mid \phi_{y_{iq}} \sim Mul(\phi_{y_{iq}})$;
      v. Sample time stamp $t_{iq} \mid \psi_{y_{iq}g_{iq}} \sim Mul(\psi_{y_{iq}g_{iq}})$;

Model Inference

Under this model, given observed data $\{U, E, D\}$, our target is to infer $\{c, z, g, y\}$ and to estimate $\{\pi, \theta, \phi, \psi\}$ and parameter η. The full posterior distribution of TCCD is:

$$\begin{aligned} &P(\pi, \theta, \phi, \psi, \eta, c, z, g, y, \xi | U, E, D, \rho, \alpha, \beta, \varepsilon, I, t) \\ &\propto P(\pi|\rho)P(\theta|\alpha)P(\phi|\beta)P(\psi|\varepsilon)P(c, g|\pi) \\ &\cdot P(z|c, \theta)P(w_d|z, \phi)P(t_d|c, z, \psi) \\ &\cdot P(y|g, \theta)P(w_e|y, \phi)P(t_e|g, y, \psi) \\ &\cdot P(e, \xi|I, \eta, g, y). \end{aligned} \tag{5}$$

The normalizing constant is difficult to calculate, thereby we adopt collapsed Gibbs sampling (Griffiths and Steyvers 2004; Li et al. 2018b) for approximate inference.

Approximate Inference

Marginalizing out $\{\pi, \theta, \phi, \psi\}$ in Eq. (5), we get:

$$\begin{aligned} & P(c, z, g, y|.) \\ & \propto \int P(\pi|\rho)P(c, g|\pi)d\pi \\ & \cdot \int P(\theta|\alpha)P(z|c, \theta)P(y|g, \theta)d\theta \\ & \cdot \int P(\phi|\beta)P(w_d|z, \phi)P(w_e|y, \phi)d\phi \\ & \cdot \int P(\psi|\varepsilon)P(t_d|c, z, \psi)P(t_e|g, y, \psi)d\psi \\ & \cdot P(e, \xi)). \end{aligned} \tag{6}$$

The first integral in Eq. (6) is calculated as follows.

$$\begin{aligned} & \int P(\pi|\rho)P(c, g|\pi)d\pi \\ & = \int (\prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}} \prod_{c=1}^{|C|} \pi_{ic}^{\rho-1})(\prod_{i=1}^{|U|} \prod_{j=1}^{|D_i|} \prod_{c=1}^{|C|} \pi_{ic}^{n_j^{(c)}}) \\ & \cdot (\prod_{i=1}^{|U|} \prod_{q=1}^{|E_i|} \prod_{c=1}^{|C|} \pi_{ig}^{n_q^{(g)}})d\pi \\ & = \prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{C}} \int \prod_{i=1}^{|U|} \prod_{c=1}^{|C|} \pi_{ic}^{n_i^{(c)}+\rho-1} d\pi \\ & = \prod_{i=1}^{|U|} \frac{\Gamma(|C|\rho)}{(\Gamma(\rho))^{|C|}} \cdot \frac{\prod_{c=1}^{|C|} \Gamma(n_i^{(c)} + \rho)}{\Gamma(n_i^{(\cdot)} + |C|\rho)}, \end{aligned} \tag{7}$$

where $n_i^{(c)}$ is the number of posts and links assigned to community c of user i. $n_i^{(\cdot)}$ denotes the total number of posts and links assigned to all communities of user i.

For the second integral in Eq. (6),

$$\begin{aligned} & \int P(\theta|\alpha)P(z|c, \theta)P(y|g, \theta)d\theta \\ & = \int (\prod_{c=1}^{|C|} \frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{K}} \prod_{k=1}^{|K|} \theta_{ck}^{\alpha-1}) \\ & \cdot \prod_{i=1}^{|U|} \prod_{j=1}^{|D_i|} \prod_{k=1}^{|K|} \theta_{ck}^{n_{jc}^{(k)}} \prod_{i=1}^{|U|} \prod_{q=1}^{|E_i|} \prod_{k=1}^{|K|} \theta_{gk}^{n_{qg}^{(k)}} d\theta \\ & = \prod_{c=1}^{|C|} \frac{\Gamma(|K|\alpha)}{(\Gamma(\alpha))^{|K|}} \cdot \frac{\prod_{k=1}^{|K|} \Gamma(n_{Dc}^{(k)} + n_{Ec}^{(k)} + \alpha)}{\Gamma(n_{Dc}^{(\cdot)} + n_{Ec}^{(\cdot)} + |K|\alpha)}, \end{aligned} \tag{8}$$

where $n_{Dc}^{(k)}$ and $n_{Ec}^{(k)}$ are the number of posts and the number of links assigned to community c with topic k respectively. $n_{Dc}^{(\cdot)}$ and $n_{Ec}^{(\cdot)}$ denote total number of posts and total number of links assigned to community c integrating all topics respectively.

For the third integral in Eq. (6),

$$\begin{aligned} & \int P(\phi|\beta)P(w_d|z, \phi)P(w_e|y, \phi)d\phi \\ & = \int (\prod_{k=1}^{|K|} \frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}} \prod_{w=1}^{|W|} \phi_{kw}^{\beta-1}) \prod_{i=1}^{|U|} \prod_{j=1}^{|D_i|} \prod_{w=1}^{|W|} \phi_{zw}^{n_{jz}^{(w)}} \\ & \cdot \prod_{i=1}^{|U|} \prod_{q=1}^{|E_i|} \prod_{w=1}^{|W|} \phi_{yw}^{n_{qy}^{(w)}} d\phi \\ & = \prod_{k=1}^{|K|} \frac{\Gamma(|W|\beta)}{(\Gamma(\beta))^{|W|}} \cdot \frac{\prod_{w=1}^{|W|} \Gamma(n_{Dk}^{(w)} + n_{Ek}^{(w)} + \beta)}{\Gamma(n_{Dk}^{(\cdot)} + n_{Ek}^{(\cdot)} + |W|\beta)}, \end{aligned} \tag{9}$$

where $n_{Dk}^{(w)}$ and $n_{Ek}^{(w)}$ denote the number of times of word w assigned to topic k in posts and link contents respectively. $n_{Dk}^{(\cdot)}$ and $n_{Ek}^{(\cdot)}$ denote total number of times of word w assigned to topic k in posts and link contents respectively.

The last integral in Eq. (6) is calculated as follows.

$$\begin{aligned} & \int P(\psi|\varepsilon)P(t_d|c, z, \psi)P(t_e|g, y, \psi)d\psi \\ & = \int \prod_{c=1}^{|C|} \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \prod_{t=1}^{|T|} \psi_{ck}^{\varepsilon-1} \prod_{i=1}^{|U|} \prod_{j=1}^{|D_i|} \prod_{t=1}^{|T|} \psi_{cz}^{n_{jcz}^{(t)}} \\ & \cdot \prod_{i=1}^{|U|} \prod_{q=1}^{|E_i|} \prod_{t=1}^{|T|} \psi_{gy}^{n_{qgy}^{(t)}} d\psi \\ & = \prod_{c=1}^{|C|} \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \\ & \cdot \int \prod_{c=1}^{|C|} \prod_{k=1}^{|K|} \prod_{t=1}^{|T|} \psi_{ck}^{n_{Dck}^{(t)} + n_{Eck}^{(t)} + \varepsilon - 1} d\psi \\ & = \prod_{c=1}^{|C|} \prod_{k=1}^{|K|} \frac{\Gamma(|T|\varepsilon)}{(\Gamma(\varepsilon))^{|T|}} \cdot \frac{\prod_{t=1}^{T} \Gamma(n_{Dck}^{(t)} + n_{Eck}^{(t)} + \varepsilon)}{\Gamma(n_{Dck}^{(\cdot)} + n_{Eck}^{(\cdot)} + |T|\varepsilon)}, \end{aligned} \tag{10}$$

where $n_{Dck}^{(t)}$ and $n_{Eck}^{(t)}$ are number of posts and number of links assigned to community c with topic k at time stamp t respectively. $n_{Dck}^{(\cdot)}$ and $n_{Eck}^{(\cdot)}$ denote total number of posts and total number of links assigned to community c with topic k integrating all time stamps respectively.

For each post d_{ij} sent by user i, we sample its community membership $c_{ij} = c$ and topic $z_{ij} = k$.

$$\begin{aligned} & P(c_{ij} = c|c_{\neg ij}, z_{ij} = k, t_{ij} = t, g, y, .) \\ & = \frac{P(c, z, g, y)}{P(c_{\neg ij}, z, g, y)} \\ & = \frac{n_{i,\neg ij}^{(c)} + \rho}{n_{i,\neg ij}^{(\cdot)} + |C|\rho} \cdot \frac{n_{c,\neg ij}^{(k)} + \alpha}{n_{c,\neg ij}^{(\cdot)} + |K|\alpha} \\ & \cdot \frac{n_{ck,\neg ij}^{(t)} + \varepsilon}{n_{ck,\neg ij}^{(\cdot)} + |T|\varepsilon}, \end{aligned} \tag{11}$$

where $n_{i,\neg ij}^{(c)}$ is number of posts and links assigned to community c sent by user i excluding current post d_{ij}. $n_{c,\neg ij}^{(k)}$ is number of posts and links assigned to community c with topic k excluding current post d_{ij}. $n_{ck,\neg ij}^{(t)}$ means the number of occurrence of time stamp t generated by community c and topic k. All dots denote marginal count, e.g., $n_{i,\neg ij}^{(\cdot)}$ denotes the total number of posts and links assigned to all communities.

$$\begin{aligned} & P(z_{ij} = k|z_{\neg ij}, c_{ij} = c, t_{ij} = t, g, y, .) \\ & = \frac{P(z, c, g, y)}{P(z_{\neg ij}, c, g, y)} \\ & = \frac{n_{c,\neg ij}^{(k)} + \alpha}{n_{c,\neg ij}^{(\cdot)} + |K|\alpha} \\ & \cdot \frac{\Pi_{w=1}^{|W|} \Pi_{q=0}^{n_{ij}^{(w)}-1} (n_{k,\neg ij}^{(w)} + q + \beta)}{\Pi_{q=0}^{n_{ij}^{(\cdot)}-1} (n_{k,\neg ij}^{(\cdot)} + q + \beta)} \\ & \cdot \frac{n_{ck,\neg ij}^{(t)} + \varepsilon}{n_{ck,\neg ij}^{(\cdot)} + |T|\varepsilon}, \end{aligned} \tag{12}$$

where $n_{ij}^{(w)}$ denotes number of occurrence of word w appearing in the post d_{ij}. $n_{k,\neg ij}^{(w)}$ denotes number of times of

word w assigned to topic k with post d_{ij} excluded. $n^{(\cdot)}_{k,\neg ij}$ is calculated over all words excluding post d_{ij}. Suppose that $e_{ii'}$ is the q-th link sent by user i. We sample user i's community membership g_{iq} and topic y_{iq} according to its contents.

$$\begin{aligned} &P(g_{ij}=c|g_{\neg ij}, y_{ij}=k, t_{ij}=t, c, z, .) \\ &= \frac{P(g,c,z,y)}{P(g_{\neg ij},c,z,y)} \\ &= \frac{n^{(c)}_{i,\neg ij}+\rho}{n^{(\cdot)}_{i,\neg ij}+|C|\rho} \cdot \frac{n^{(k)}_{c,\neg ij}+\alpha}{n^{(\cdot)}_{c,\neg ij}+|K|\alpha} \\ &\cdot \frac{n^{(t)}_{ck,\neg ij}+\varepsilon}{n^{(\cdot)}_{ck,\neg ij}+|T|\varepsilon} \cdot \varphi(\omega_{ij},\xi_{ij}). \end{aligned} \tag{13}$$

$$\begin{aligned} &P(y_{ij}=k|y_{\neg ij}, g_{ij}=c, t_{ij}=t, c, z, .) \\ &= \frac{P(y,c,z,g)}{P(y_{\neg ij},c,z,g)} \\ &= \frac{n^{(k)}_{c,\neg ij}+\alpha}{n^{(\cdot)}_{c,\neg ij}+|K|\alpha} \\ &\cdot \frac{\Pi_{w=1}^{|W|}\Pi_{q=0}^{n^{(w)}_{ij}-1}(n^{(w)}_{k,\neg ij}+q+\beta)}{\Pi_{q=0}^{n^{(\cdot)}_{ij}-1}(n^{(\cdot)}_{k,\neg ij}+q+\beta)} \\ &\cdot \frac{n^{(t)}_{ck,\neg ij}+\varepsilon}{n^{(\cdot)}_{ck,\neg ij}+|T|\varepsilon} \cdot \varphi(\omega_{ij},\xi_{ij}). \end{aligned} \tag{14}$$

At last we sample ξ_{ij}.

$$P(\xi_{ij}|.) \propto e^{-\frac{1}{2}\xi_{ij}\omega_{ij}^2} P(\xi_{ij}|1,0) = PG(1,\omega_{ij}). \tag{15}$$

Parameter Estimation

We obtain above samples by running Gibbs sampler for adequate iterations. Then, we estimate as follows:

$$\hat{\pi}_{ic} = \frac{n^{(c)}_i+\rho}{n^{(\cdot)}_i+|C|\rho}. \tag{16}$$

$$\hat{\theta}_{ck} = \frac{n^{(k)}_c+\alpha}{n^{(\cdot)}_c+|K|\alpha}. \tag{17}$$

$$\hat{\phi}_{kw} = \frac{n^{(w)}_k+\beta}{n^{(\cdot)}_k+|W|\beta}. \tag{18}$$

$$\hat{\psi}_{kc,t} = \frac{n^{(t)}_{ck}+\varepsilon}{n^{(\cdot)}_{ck}+|T|\varepsilon}. \tag{19}$$

For parameter η, we aggregate all community and topic pairs w.r.t all links.

Algorithm Summarization and Time Complexity

Our inference procedure is shown in Alg.1.

T denotes the iterations for convergence. For steps 3-6, we sample community indicator and topic indicator for posts of all users. The total number of users is $|U|$ and $|D_i|$ is the number of posts of user i. In line 5, because all counters (e.g., how many times a user is assigned to a community) are recorded in memory, equation 7 takes constant time for a specific community. In line 6, to compute the second fraction of equation 8, it takes $\Theta(|W|)$ for a specific topic, where $|W|$ is the size of vocabulary. There are $|K|$ topics. So, steps 3-6 takes $\Theta(|U| \times |D| \times |C| + |U| \times |D| \times |K| \times |W|)$. In steps 7-10, it computes community indicator, topic indicator and ξ_{ij} for all links. The number of all links is $|E|$. Equation 9 and equation 11 takes constant time. Equation 10 takes $\Theta(|W|)$. So, steps 7-10 takes $\Theta(|E|\times|C|+|E|\times|K|\times|W|)$. For steps 11-12, we calculate η. It takes $\Theta(|E|)$. Based on the above discussions, the complexity is linear to the data size. As data size grows bigger, the efficiency turns to be lower. Since our key target is to evaluate the accuracy of community detection, we leave parallel implementation of TCCD as our future work.

Algorithm 1 Inference for TCCD

1: Initialize $\alpha, \beta, \epsilon, \rho, \eta$;
2: **for** $iter = 1 : T$ **do**
3: **for** each user $i \in U$ **do**
4: **for** each post $d_{ij} \in D_i$ **do**
5: Sample community indicator c_{ij} according to Eq. (11);
6: Sample topic indicator z_{ij} according to Eq. (12);
7: **end for**
8: **end for**
9: **for** each link $e \in E$ **do**
10: Sample community indicator g_{ij} according to Eq. (13);
11: Sample topic indicator y_{ij} according to Eq. (14);
12: Sample ξ_{ij} according to Eq. (15);
13: **end for**
14: **for** each link $e \in E$ **do**
15: Update η by aggregating community and topic of two endpoint users;
16: **end for**
17: **end for**

Experiments

We evaluate our model on two real datasets and compare it with four state-of-the-art baselines. All experiments are implemented on a computer with Intel 4.2GHz CPUs and 32GB RAMs.

Datasets

To accurately evaluate community detection results of TCCD and other baselines, we choose two real datasets with ground-truth: Reddit dataset and DBLP dataset (Wang, Lai, and Philip 2014).

Reddit data covers period from August 25, 2012 to August 31, 2012. It includes three sub-forums: Science, Movie and Politics. So, there are three communities. The threads sent by users are used as node contents. We choose one day as a time snap. DBLP dataset is a paper co-authorship network consisting of publications in three research fields from year 2001 to 2011. So, the number of communities is three. We choose one year as a time snap. For Reddit dataset, we remove users who does not have any posts. For DBLP dataset,

we remove authors who publish less than four papers. After removing stop words and stemming by PreTexT2, the statistics of the two datasets are summarized in Table 2.

Table 2: Summarization of datasets with ground-truth

	#users	#links	#user posts	#words
Reddit	23,820	51,149	3,925	14,370
DBLP	24,241	209,351	68,702	7,769

Baselines

We choose four state-of-the-art baselines to evaluate our model's accuracy:

- Community Level Diffusion (COLD) (Hu et al. 2015). It generates documents and links based on the same latent community membership factor. It assigns a time stamp vector to each document to identify temporal topics of communities.
- Community Profiling and Detection (CPD) (Cai et al. 2017). This model integrates friendship relations, diffusion links and individual preference to identify community profiling.
- Poisson Mixed-Topic Link Model (PMTLM) (Zhu et al. 2013). It combines the LDA model and Poisson distribution to generate the text of each node and the links between them.
- Community Role Model (CRM) (Han and Tang 2015). It assigns a role to each user and model friendship links and diffusion links in networks based on users' community assignment.

Metrics

To evaluate the accuracy of the community detection outcomes, we use generalized normalized mutual information (GNMI) (Wu, Xiong, and Chen 2009), F-score and Jaccard index as metrics. F-score is the harmonic mean of precision and recall: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. Jaccard index is defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$(i.e., measuring the similarity of sample set A and B).

Comparison with Baselines

Table 3 shows the result comparisons between baselines and TCCD on two datasets respectively. TCCD outperforms all baselines for all metrics.

For Reddit dataset, there are 3,925 isolated posts. They don't appear on links. Our model separates these isolated posts and link posts, such that they do not participate in the formulation of network topology. Results show that our model achieves 42% GNMI improvement, 3% F-score improvement and 3% Jaccard improvement over the second-best baseline on Reddit. For DBLP dataset, table 3 shows that our model achieves 38% GNMI improvement, 4% F-score improvement and 6% Jaccard improvement over the second-best baseline on DBLP.

Table 3: Experimental results comparations on Reddit and DBLP

Metrics (%)	Datasets	Methods				
		COLD	CPD	CRM	PMTLM	Ours
NMI	Reddit	16.24	13.12	38.47	41.61	**59.07**
	DBLP	31.77	25.56	20.99	14.94	**44.23**
F-score	Reddit	59.81	80.49	69.96	64.30	**82.95**
	DBLP	74.90	78.66	70.50	72.01	**81.50**
Jaccard	Reddit	44.82	70.68	56.49	57.36	**72.64**
	DBLP	60.45	65.26	55.39	56.28	**69.10**

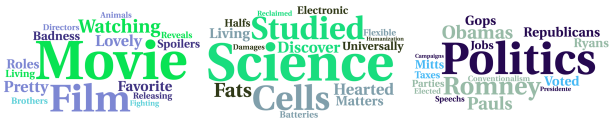

Figure 2: Word clouds of three topics: *Movie*, *Science* and *Politics*.

TCCD is more capable to process those networks with isolated user posts, which is a common case in social networks. Even for those networks without isolated node contents such as paper co-authorship networks, TCCD also outperforms all baselines for all metrics.

A Case Study

In this section, we analyze four parameters modeled in TCCD on Reddit dataset. They are topic distribution of communities, word distribution of topics, temporal topics of each community, and topic correlations respectively. (i.e., $\{\theta, \phi, \psi, \eta\}$).

- Word distribution of each topic
  Word clouds of three topics are illustrated in Fig.2. It shows that each topic we detected is meaningful (i.e.,*Movie*, *Politics* and *Science*).
- Topic distribution of each community
  In Fig.3, three doughnut charts represent three communities (i.e., *Movie*, *Politics* and *Science*). Each color on doughnuts denotes one topic. As it shows, topic *Movie* and *Politics* are dominant in community *Movie* and *Politics* respectively. But for community *Science*, though the topic *Science* is dominant, there are 35 percentage of posts talking about *Politics* and 16 percentage of posts talking about *Movie*.
- Time distribution of community and topic
  In Fig.3, three plots around each community illustrate temporal variations of corresponding topics. Even though each community has a dominant topic, it also includes some discussions about other two topics. The temporal pattern of the dominant topic denotes how most of users in a community focus on the topic at different timestamps.
- Topic correlations
  Fig.4 shows the topic correlations with respect to com-

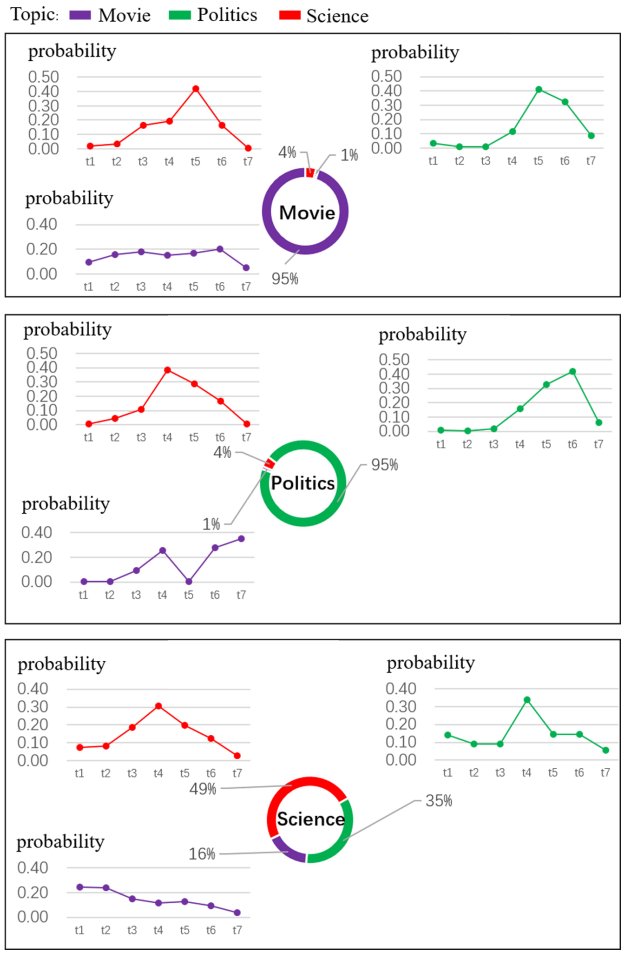

Figure 3: Topic distribution of three communities and time distribution of community and topic. Colors represent topics. Circles represent communities.

munities. Fig.4(a) is the topic correlations inside community *Movie*, which represents how users in this community interact with each other. The circle with the biggest weight from topic *Movie* to itself shows that users in community *Movie* are more likely to talk about *Movie* topic. The weights of other links are small, which shows that users focusing on *Science* and *Politics* also interact with each other but with much less intense communication. Fig.4(b) is the topic correlations inside community *Politics*. We can see that there are more users talking about topic *Politics*. The interactions between other topic pairs also exist but with small weights. Fig.4(c) is the topic correlations inside community *Science*. In Reddit dataset, the number of posts in *Science* is less than that in other two communities. So, the interactions are sparser than other two communities. Even so, there are more users talking about topic *Science*. Fig.4(d) represents the topic correlations between community *Science* and *Politics*. As it shows, users in the two communities talk about all topics. The reason is that there are 35 percentage of posts in community *Science* talking about *Politics* and 16 percentage of posts talking about *Movie*. The topic correlations between other community pairs also exists with probability smaller than 0.01. So, we don't present them, which means that users in these two different communities seldom interact with each other.

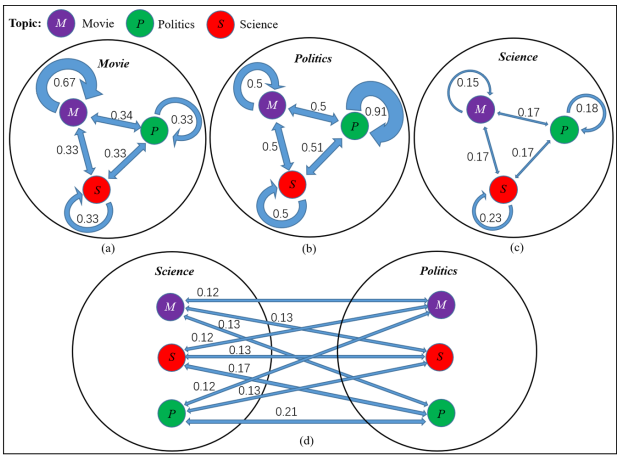

Figure 4: Topic correlations with respect to communities. Big circles with labels represent communities (i.e. *Movie*, *Politics* and *Science*). Solid circles with colors represent topics. The weighted arc with different width represents topic correlations. Figure (a) to (c) represent the topic correlations inside communities: *Movie*, *Politics* and *Science*. Figure (d) is the topic correlations between community *Science* and *Politics*.

Parameter Initiation

We set $|C|$ and $|K|$ to real value according to ground-truth. For η, we can initiate it at random. For Dirichlet hyperparameters, we run TCCD under different values. The results show that TCCD is not sensitive to Dirichlet hyperparameters, thus we set them to fixed values (i.e., $\rho = 0.01$, $\alpha = 0.001$, $\beta = 0.1$, $\epsilon = 0.001$). For the threshold for determining overlapping communities, we test its values from $1/|C|$ to 0.5 with step 0.1. The experiments show that $1/|C|$ is the best value. For each user, we choose those communities with probabilities bigger than the threshold as his real communities.

Conclusion and Discussions

In this paper, we found that there are correlations between topics, which significantly affect community structures. The observation reveals that existing methods are limited to resolve three key issues: a) How to process node contents and edge contents to infer topics and topic correlations in a unified model; b) How to generate network topology considering the influence of topic correlations; c) How to detect the composition of topics inside communities to understand community semantics. We proposed a generative model (TCCD) for community detection which consists of three components, i.e., user post component, link content

component and link component. The experimental results show that our model improves the accuracy of community detection and can detect all topics efficiently. It resolves all above key issues. Our work for the first time interprets the mechanism of the composition of topics inside communities to understand community semantics in a natural way. It can also reveal how the popularity of a topic changes over time in a community.

We also found that topic correlations in different communities are not consistent. It is the true reflection of communications of users in communities. Because most users in a community mainly focus on primary topics. For the topics that are talked about by few users , the correlations are small. In fact, the fundamental reason is the limitation of community definition which is only based on network topology.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2018YFC0809800) and the Natural Science Foundation of China (No. 61772361, 61572353).

References

Blei, D., and Lafferty, J. 2006. Correlated topic models. *Advances in neural information processing systems* 18:147.

Cai, H.; Zheng, V. W.; Zhu, F.; Chang, K. C.-C.; and Huang, Z. 2017. From community detection to community profiling. *Proceedings of the VLDB Endowment* 10(7):817–828.

Chen, J.; Zhu, J.; Wang, Z.; Zheng, X.; and Zhang, B. 2013. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, 2445–2453.

Fortunato, S., and Hric, D. 2016. Community detection in networks: A user guide. *Physics Reports* 659:1–44.

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12):7821–7826.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Han, Y., and Tang, J. 2015. Probabilistic community and role model for social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 407–416. ACM.

He, D.; Feng, Z.; Jin, D.; Wang, X.; and Zhang, W. 2017a. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In *AAAI*.

He, J.; Hu, Z.; Berg-Kirkpatrick, T.; Huang, Y.; and Xing, E. P. 2017b. Efficient correlated topic modeling with topic embedding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 225–233. ACM.

Hu, Z.; Yao, J.; Cui, B.; and Xing, E. 2015. Community level diffusion extraction. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1555–1569. ACM.

Jiang, D.; Leung, K. W.-T.; Ng, W.; and Li, H. 2013. Beyond click graph: Topic modeling for search engine query log analysis. In *International Conference on Database Systems for Advanced Applications*, 209–223. Springer.

Jin, D.; Wang, X.; He, R.; He, D.; Dang, J.; and Zhang, W. 2018. Robust detection of link communities in large social networks by exploiting link semantics. In *AAAI Conference on Artificial Intelligence*.

Li, S.; Chua, T.-S.; Zhu, J.; and Miao, C. 2016. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 666–675.

Li, X.; Li, C.; Chi, J.; Ouyang, J.; and Li, C. 2018a. Dataless text classification: A topic modeling approach with document manifold. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 973–982. ACM.

Li, X.; Li, C.; Chi, J.; and Ouyang, J. 2018b. Variance reduction in black-box variational inference by adaptive importance sampling. In *IJCAI*, 2404–2410.

Mcauley, J., and Leskovec, J. 2014. Discovering social circles in ego networks. *Acm Transactions on Knowledge Discovery from Data* 8(1):1–28.

Newman, M. E. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103(23):8577–8582.

Pei, Y.; Chakraborty, N.; and Sycara, K. 2015. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *International Conference on Artificial Intelligence*, 2083–2089.

Polson, N. G.; Scott, J. G.; and Windle, J. 2013. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association* 108(504):1339–1349.

Tu, K.; Cui, P.; Wang, X.; Yu, P. S.; and Zhu, W. 2018. Deep recursive network embedding with regular equivalence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2357–2366. ACM.

Wang, C.-D.; Lai, J.-H.; and Philip, S. Y. 2014. Neiwalk: community discovery in dynamic content-based networks. *IEEE transactions on knowledge and data engineering* 26(7):1734–1748.

Wu, J.; Xiong, H.; and Chen, J. 2009. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 877–886. ACM.

Zhang, G.; Jin, D.; Gao, J.; Jiao, P.; Fogelman-Soulié, F.; and Huang, X. 2018. Finding communities with hierarchical semantics by distinguishing general and specialized topics. In *IJCAI*, 3648–3654.

Zhu, Y.; Yan, X.; Getoor, L.; and Moore, C. 2013. Scalable text and link analysis with mixed-topic link models. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 473–481.