Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Joint Identification of Network Communities and Semantics via Integrative Modeling of Network Topologies and Node Contents

Dongxiao He,[1] Zhiyong Feng,[2] Di Jin,[1] Xiaobao Wang,[2] Weixiong Zhang[3,4]

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China, [2]School of Computer Software, Tianjin University, Tianjin 300072, China, [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
{hedongxiao, zyfeng, jindi, xiaobao}@tju.edu.cn, weixiong.zhang@wustl.edu

Abstract

The objective of discovering network communities, an essential step in complex systems analysis, is two-fold: identification of functional modules and their semantics at the same time. However, most existing community-finding methods have focused on finding communities using network topologies, and the problem of extracting module semantics has not been well studied and node contents, which often contain semantic information of nodes and networks, have not been fully utilized. We considered the problem of identifying network communities and module semantics at the same time. We introduced a novel generative model with two closely correlated parts, one for communities and the other for semantics. We developed a co-learning strategy to jointly train the two parts of the model by combining a nested EM algorithm and belief propagation. By extracting the latent correlation between the two parts, our new method is not only robust for finding communities and semantics, but also able to provide more than one semantic explanation to a community. We evaluated the new method on artificial benchmarks and analyzed the semantic interpretability by a case study. We compared the new method with eight state-of-the-art methods on ten real-world networks, showing its superior performance over the existing methods.

1. Introduction

Complex systems, which are best represented as networks, are often organized in functional modules that directly or indirectly interact with one another. For example, large companies are typically organized in units with designated functions, and proteins in the cell typically form complexes to exert their functions. Therefore, it is essential to identify *communities* or modules in networks of interest, where nodes within a community are densely connected (Fortunato 2010). Identification of communities of a network helps understand how the system is organized and how individual modules function. It is equally important to identify the underlying semantics of communities so as to explain the meaning or extract the functions of the communities, i.e., to functionally annotate the communities.

Most conventional community detection algorithms only use network topologies. The premise is that functional communities have structural signatures (Yang and Leskovec 2014). Many community detection methods have been proposed (Fortunato 2010; Fortunato and Hric 2016) based on various assumptions and using different techniques, including hierarchical clustering (Girvan and Newman 2002; Newman 2004), modularity optimization (Blondel et al. 2008; Duch and Arenas 2005), spectral partition (White and Smyth 2005), Markov dynamics (Rosvall and Bergstrom 2008), and statistical inference (Karrer and Newman 2011; He et al. 2015; Yang and Leskovec 2013).

Node contents, particularly the attributes of nodes, have been recently used in finding communities. It is believed that individuals or objects with similar attributes or features are likely to belong to the same community. Different from network structures that specify node connectivities, node contents provide semantic information of nodes and underlying network. Such semantics may capture deep knowledge of the nature of communities and are orthogonal and complementary to structural information. When used together, missing structural information may be compensated by content information, and vice versa, to improve community detection. Indeed, methods using these two types of information have been proposed (Balasubramanyan and Cohen 2011; Sun, Aggarwal and Relation 2012; Xu et al. 2012; Yang et al. 2009; Mcauley and Leskovec 2014; Pei, Chakraborty and Sycara 2015; Ruan, Fuhry and Parthasarathy 2013).

In addition to improving community detection, node contents may also provide semantic descriptions of communities. Such descriptions may help explain why certain nodes belong to a community, or help reveal the functions or characteristics of communities. Community semantics certainly make network analysis valuable. It has been proposed recently to use structural and content information to identify communities and derive descriptions (Pool, Bonchi and Leeuwen 2014; Yang, McAuley and Leskovec 2013; Liu et al. 2015; Wang et al 2016).

However, these newly developed methods have at least three serious problems. First, they typically assume that network structures and node contents share the same information of node community memberships, which is often nota the case in practice. For example, the social relationships in Twitter often reflect user groups directly, while users may generate messages of diverse contents (Pei, Chakraborty and Sycara 2015), so that contents and community structures may not align at all. When node contents do not match well with the underlying community structures, these algorithms perform poorly. Second, they assume one topic per community, an assumption that does not hold in practice. In social networks again, users tend to twit frequently over more than one topic, so that a community may better be characterized by multiple topics. Focusing on one topic for a community limits the applicability of the existing methods. Third, the existing methods handle topologies and node contents separately. As a result, they need to balance the effects of the two on community detection, which is difficult to achieve.

We introduced a generative model for jointly identifying communities and deriving their semantic description at the same time. The model accommodates two sets of variables, one for communities and the other for description, which are implicitly correlated. To train the model, we developed an effective method to combine a nested expectation-maximization (EM) algorithm and a belief propagation process, which is named as NEMBP. The learning process models and explores the hidden correlation between the two parts of the model to improve community detection and description extraction.

2. The Model

We aim at developing a novel generative model for undirected and unweighted networks with node contents (Figure 1). An attributed network G of n nodes and m node attributes can be represented by an adjacency matrix A = $(a_{ij})_{n \times n}$ and an attribute matrix X = $(x_{ik})_{n \times m}$, where $a_{ij} = 1$ if an edge exists between nodes v_i and v_j, or 0 otherwise, and $x_{ik} = 1$ if node v_i has the kth attribute w_k, or 0 otherwise.

Our objective is twofold: 1) to partition the nodes separately into communities and clusters of contents and 2) to seek the best association between the two so as to best annotate communities using semantics from content clusters.

In particular, we divide the set of nodes V into c *network communities* such that the nodes within a community are densely connected and nodes in different communities are sparsely linked, and partition V into c *content clusters* such that the nodes in a cluster share common attributes, which are named as *semantic topics* in topic modeling (Blei, Ng and Jordan 2003). Ideally, we want to associate a community with at least one content cluster as its semantics, i.e., we seek the best interpretation of the communities using the content clusters and their topics. In the process, we search for communities and content clusters at the same time using the association between the two derived so far as a guide.

We fold the two objectives into a unified model, which is specified by three types of quantities. The first is the set of observed quantities, including the adjacency matrix A and the attribute matrix X. The second is the set of latent quantities, including the community memberships z where z_i is the label of the community that node v_i belongs to, and the topic (or content) assignments g where g_{ik} is the label of topic that the node-attribute pair $<v_i, w_k>$ specifies. The third is the set of model parameters: 1) $\pi = (\pi_r)_{1 \times c}$, where $\pi_r = p(z_i = r)$ is the probability that the node v_i belong to the rth community; 2) $\Theta = (\theta_{rs})_{c \times c}$, where $\theta_{rs} = p(z_i = r, z_j = s)$ is the probability that a pair of nodes in the rth and sth communities is connected; 3) $\mathrm{H} = (\eta_{rs})_{c \times c}$, where $\eta_{rs} = p(g_{ik} = s \mid z_i = r)$ is the probability that node v_i is in the sth content cluster given that it belongs to the rth community (independent of attribute w_k); and 4) $\mathrm{B} = (\beta_{sk})_{c \times m}$, where $\beta_{sk} = p(x_{ik} = 1 \mid g_{ik} = s)$ is the probability that the sth topic uses the kth attribute, which is independent of node v_i.

Note that we use the shared latent variables (g) for nodes and attributes, in order to make the content clusters (denoted by H) to be associated with the topics consisting of attributes (denoted by B), so that a cluster has a topic to represent its semantic. Besides, we use $d_i d_j \theta_{rs}$ instead of θ_{rs} to denote the probability $p(z_i = r, z_j = s)$, where d_i is the degree of node v_i. This corresponds to the degree-corrected stochastic blockmodel (DCSBM) (Karrer and Newman 2011) that can better describe network communities.

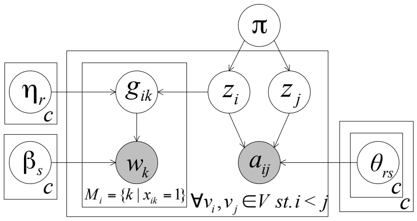


Figure 1: *A sketch of the model of communities and semantics. The right part is the fitting of the model (with latent communities z, their prior π and block matrix $\Theta = (\theta_{rs})_{c \times c}$) to the observed network data in adjacency matrix $A = (a_{ij})_{n \times n}$. The upper-left part is the prior for generating topics g with distribution $H = (\eta_{rs})_{c \times c}$ under communities z. The lower-left part is the fitting of the mod-*

el (with g and topic-attribute distribution B = $(\beta_{sk})_{c\cdot m}$) to the given content information (in attribute matrix X = $(x_{ik})_{n\cdot m}$).

The model is sketched in Figure 1 and can be generated in the following.

For each node v_i:

(a) Draw community assignment $z_i \sim Multinomial(\pi)$
(b) For each node v_j with $j > i$:
Draw edge $a_{ij} \sim Bernoulli(d_i d_j \theta_{z_i z_j})$
(c) For each of the kth attribute with $x_{ik} = 1$:
i. Draw topic assignment $g_{ik} \sim Multinomial(\eta_{z_i})$
ii. Draw attribute $w_k \sim Multinomial(\beta_{g_{ik}})$

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

$$P(\mathrm{A},\mathrm{X}|\pi,\Theta,\mathrm{H},\mathrm{B})$$

$$= \sum_{\mathrm{z},\mathrm{g}} P(\mathrm{z}\,|\,\pi)P(\mathrm{A}\,|\,\Theta,\mathrm{z})P(\mathrm{g}\,|\,\mathrm{H},\mathrm{z})P(\mathrm{X}\,|\,\mathrm{B},\mathrm{g}) \tag{1}$$

$$= \sum_{\mathrm{z},\mathrm{g}} \left(\begin{array}{l} \prod_{i=1}^{n} \pi_{z_i} \prod_{i<j} (d_i d_j \theta_{z_i z_j})^{a_{ij}} (1 - d_i d_j \theta_{z_i z_j})^{1-a_{ij}} \\ \times \prod_{i=1}^{n} \prod_{k=1}^{m} (\eta_{z_i g_{ik}})^{x_{ik}} \prod_{i=1}^{n} \prod_{k=1}^{m} (\beta_{g_{ik} k})^{x_{ik}} \end{array} \right)$$

subject to $\sum_{r=1}^{c} \pi_r = 1$, $\sum_{s=1}^{c} \eta_{rs} = 1$, and $\sum_{k=1}^{m} \beta_{sk} = 1$.

Eq. (1) has four parts. The first two parts are the fitting of the model to network structures, the third part is the prior probability of generating content clusters (with their topics) g under communities z with distribution H. The fourth part is the fitting to contents. Generally, the fitting to structures and fitting to contents are dominant in the likelihood, and the prior helps improve the overall model fitting.

The latent correlation matrix H is a matrix of probabilistic transitions from communities to content clusters and is critical for finding communities and content clusters. When the communities and content clusters at hand match well, H will be close to an identity matrix. We may use the correlations in H, along with the topics derived, to interpret the communities. Even if the communities and content clusters do not match well, we may still utilize the correlations in H to improve the community result through the projection from the clusters to communities. On the other hand, if the communities and content clusters do not match at all, H will be nearly homogenous in that its values are nearly the same so that no correspondence between communities and clusters will emerge. This may occur when there exists indeed a disparity between network structures and node contents in the data. For this case, we may ignore the contents and return communities as the only result.

Besides, in the case when the contents are too noisy to form clusters regardless if they match with communities. The fitting to the contents will have little effect on the likelihood, so that the prior and the fitting to network structures are dominant. Therefore, the communities depend mainly on network topologies, and the correlation matrix H will be almost an identity one as it can help maximize the likelihood. Likewise, when the network has a poor community structure, the priori and the fitting to contents will be main factors on the likelihood and H will also be an identity matrix. Then, the communities and content clusters will be the same, which depend on contents only.

This is a probabilistic model with two parts (for communities and clusters of contents/topics) that are linked through latent associations. As such, it does not require a parameter to balance the two parts like some existing methods do. Finding these associations is a central piece of model training.

3. Training the Model

The model is trained through a nested expectation-maximization (EM) algorithm with an inference process of belief propagation.

3.1 Fitting the Model to Data

Given the observed data, we aim to find the model parameters π, Θ, H and B to maximize the likelihood in (1). Since this is difficult, we instead maximize its logarithm:

$$L = \log \sum_{\mathrm{z},\mathrm{g}} P(\mathrm{z}\,|\,\pi)P(\mathrm{A}\,|\,\Theta,\mathrm{z})P(\mathrm{g}\,|\,\mathrm{H},\mathrm{z})P(\mathrm{X}\,|\,\mathrm{B},\mathrm{g}) \tag{2}$$

Since maximizing (2) is still nontrivial, we adopt an EM algorithm (Dempster, Laird and Rubin 1977). By applying the Jensen's inequality to (2), we have the expected log likelihood $\bar{L}$.

$$L \geq \bar{L} = \sum_{\mathrm{z}} q(\mathrm{z}) \log \frac{\sum_{\mathrm{g}} P(\mathrm{z}\,|\,\pi)P(\mathrm{A}\,|\,\Theta,\mathrm{z})P(\mathrm{g}\,|\,\mathrm{H},\mathrm{z})P(\mathrm{X}\,|\,\mathrm{B},\mathrm{g})}{q(\mathrm{z})}$$

$$= \sum_{i=1}^{n} \sum_{r=1}^{c} q_i^r \left(\log \pi_r + \sum_{k=1}^{m} (x_{ik} \log \sum_{s=1}^{c} \eta_{rs} \beta_{sk}) \right) \tag{3}$$

$$+ \sum_{i<j} \sum_{r,s=1}^{c} q_{ij}^{rs} \left(\begin{array}{l} a_{ij} (\log d_i + \log d_j + \log \theta_{rs}) \\ +(1 - a_{ij}) \log(1 - d_i d_j \theta_{rs}) \end{array} \right) - \sum_{\mathrm{z}} q(\mathrm{z}) \log q(\mathrm{z})$$

where q(z) is a distribution over community memberships z such that $\sum_{\mathrm{z}} q(\mathrm{z}) = 1$, $q_i^r = \sum_{\mathrm{z}} q(\mathrm{z}) \delta_{z_i r}$ is the marginal probability within q(z) that node v_i belongs to community r, $q_{ij}^{rs} = \sum_{\mathrm{z}} q(\mathrm{z}) \delta_{z_i r} \delta_{z_j s}$ is the joint marginal probability that nodes v_i and v_j belong to communities r and s, respectively, and δ_{rs} is the Kronecker delta.

The maximum of $\bar{L}$ with respect to possible choices of distribution q(z) is achieved when $\bar{L} = L$. Following Jensen's inequality, this is when

$$q(\mathrm{z}) = \frac{\sum_{\mathrm{g}} P(\mathrm{z}\,|\,\pi)P(\mathrm{A}\,|\,\Theta,\mathrm{z})P(\mathrm{g}\,|\,\mathrm{H},\mathrm{z})P(\mathrm{X}\,|\,\mathrm{B},\mathrm{g})}{\sum_{\mathrm{z},\mathrm{g}} P(\mathrm{z}\,|\,\pi)P(\mathrm{A}\,|\,\Theta,\mathrm{z})P(\mathrm{g}\,|\,\mathrm{H},\mathrm{z})P(\mathrm{X}\,|\,\mathrm{B},\mathrm{g})} \tag{4}$$

Thus, maximization of L with respect to π, Θ, H and B to obtain the best parameters is equivalent to maximization of its lower bound $\bar{L}$ with respect to q(z) (making $\bar{L} = L$) and the parameters. The EM algorithm for this dual maximization is to repeatedly maximize q(z) (i.e., the E-step) first and then π, Θ, H and B (i.e., the M-step), which can be proven to monotonically converge to a local maximum.

3.1.1 The E-Step with Belief Propagation

It is possible to infer the optimal $q(\mathrm{z})$ using (4) in the E-step, but this amounts to computing all possible c^n community assignments z in the denominator, which is infeasible for all but some small networks. The standard way around this problem is to approximate the distribution $q(\mathrm{z})$ by an importance sampling using Markov chain Monte Carlo.

Here, we instead use a recently proposed method based on a fast belief propagation (BP) (Decelle et al. 2011; Martin, Ball and Newman 2016). We first define the BP message $\psi_r^{i\to j}$ to follow the marginal probability that node v_i belongs to community r in absence of node v_j. We then derive the BP equations that are a good approximation for large sparse networks to speedup the computation.

In the approximation, if there is an edge between v_i and v_j (i.e., $a_{ij} = 1$) the BP equation for the message $\psi_r^{i\to j}$ is:

$$\psi_r^{i\to j} = \frac{\pi_r}{Z^{i\to j}} \prod_{k=1}^{m}\left(\sum_{s=1}^{c}\eta_{rs}\beta_{sk}\right)^{x_{ik}} \exp(-h_r^i) \prod_{k\in\partial_i/j}\sum_{s=1}^{c}\psi_s^{k\to i}\theta_{sr} \qquad (5)$$

where $Z^{i\to j}$ is the normalization coefficient to ensure $\sum_{r=1}^{c}\psi_r^{i\to j} = 1$, and ∂_i the neighbor set of v_i. But for the non-neighboring v_j's of node v_i (i.e., $a_{ij} = 0$), as an approximation we force them to share the same message ψ_r^i :

$$\psi_r^{i} = \frac{\pi_r}{Z^{i}} \prod_{k=1}^{m}\left(\sum_{s=1}^{c}\eta_{rs}\beta_{sk}\right)^{x_{ik}} \exp(-h_r^i) \prod_{k\in\partial_i}\sum_{s=1}^{c}\psi_s^{k\to i}\theta_{sr} \qquad (6)$$

where Z^i is for normalization so that $\sum_{r=1}^{c}\psi_r^{i} = 1$, and

$$h_r^i = \sum_{k=1}^{n}\sum_{s=1}^{c} d_k d_i \theta_{sr}\psi_s^k \qquad (7)$$

The BP equations (5) and (6) can be solved by iteration. Upon convergence, we have the one-node marginal probability $q_i^r = \psi_r^i$, and the two-node marginal probability

$$\begin{aligned} q_{ij}^{rs} &= P\left(z_i = r, z_j = s \mid a_{ij}\right) \\ &= \frac{P\left(z_i = r, z_j = s\right) P\left(a_{ij} \mid z_i = r, z_j = s\right)}{\sum_{rs} P\left(z_i = r, z_j = s\right) P\left(a_{ij} \mid z_i = r, z_j = s\right)} \end{aligned} \qquad (8)$$

where $P(z_i = r, z_j = s) = \psi_r^{i\to j}\psi_s^{j\to i}$ and $P(a_{ij}|z_i = r, z_j = s) = (d_i d_j \theta_{rs})^{a_{ij}}(1 - d_i d_j \theta_{rs})^{1-a_{ij}}$. As to be shown shortly, we only need to calculate q_{ij}^{rs}'s with $a_{ij} = 1$.

3.1.2 The M-Step with a Nested EM Procedure

We now consider maximizing $\bar{L}$ in (3) with q_i^r and q_{ij}^{rs} fixed. To maximize over π and Θ, we differentiate $\bar{L}$ with respect to π_r, subject to the condition $\sum_{r=1}^{c}\pi_r = 1$, gives

$$\pi_r = \frac{1}{n}\sum_{i=1}^{n} q_i^r \qquad (9)$$

Ignoring the terms beyond the first order, we substitute $\sum_{i<j,r,s} q_{ij}^{rs}(1 - a_{ij})\log(1 - d_i d_j \theta_{rs}) \simeq -\frac{1}{2}\sum_{i,j,r,s} q_{ij}^{rs} d_i d_j \theta_{rs}$ into (3). Taking derivative and setting the result to zero, the maximum with respect to θ_{rs} is

$$\theta_{rs} = \frac{\sum_{i,j=1}^{n} a_{ij} q_{ij}^{rs}}{\sum_{i,j=1}^{n} d_i d_j q_{ij}^{rs}} \simeq \frac{\sum_{i,j=1}^{n} a_{ij} q_{ij}^{rs}}{\left(\sum_{i=1}^{n} d_i q_i^r\right)\left(\sum_{j=1}^{n} d_j q_j^s\right)} \qquad (10)$$

where we use $q_{ij}^{rs} \simeq q_i^r q_j^s$ in the denominator because for a large sparse network the community assignments of distant nodes are not correlated.

However, maximizing $\bar{L}$ with respect to H and B is complicated as it is contained by latent variables g. Again, we use the EM algorithm and apply Jensen's inequality to (3),

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{r=1}^{c} q_i^r \sum_{k=1}^{m}\left(x_{ik} \log \sum_{s=1}^{c}\eta_{rs}\beta_{sk}\right) \\ &\geq \sum_{i=1}^{n}\sum_{r=1}^{c} q_i^r \sum_{k=1}^{m}\left(x_{ik} \sum_{s=1}^{c} t_{rk}^s \log(\eta_{rs}\beta_{sk} / t_{rk}^s)\right) \end{aligned} \qquad (11)$$

where t_{rk}^s can be any distribution subject to $\sum_{s=1}^{c} t_{rk}^s = 1$. Eq. (11) ignores the items in $\bar{L}$ in (3) which can be regarded as the constant with respect to H and B. The exact equality of (11), and hence the maximum of the right-hand side, is achieved when

$$t_{rk}^s = \eta_{rs}\beta_{sk} / \sum_{s=1}^{c}\eta_{rs}\beta_{sk} \qquad (12)$$

As before, we can maximize the left-hand side of (11) by repeatedly maximizing the right-hand side with respect to t_{rk}^s using (12) and with respect to H and B by differentiation. Differentiating the right-hand side of (11) with respect to η_{rs}, subject to $\sum_{s=1}^{c}\eta_{rs} = 1$ for $r = 1\ldots c$, results in

$$\eta_{rs} = \sum_{i=1}^{n}\sum_{k=1}^{m} x_{ik} t_{rk}^s q_i^r / \sum_{i=1}^{n} q_i^r K_i \qquad (13)$$

where $K_i = \sum_{k=1}^{m} x_{ik}$. Similarly, differentiating with respect to β_{sk}, subject to $\sum_{k=1}^{m}\beta_{sk} = 1$ for $s = 1\ldots c$, gives

$$\beta_{sk} = \sum_{i=1}^{n}\sum_{r=1}^{c} x_{ik} q_i^r t_{rk}^s / \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{r=1}^{c} x_{ik} q_i^r t_{rk}^s \qquad (14)$$

Subsequently, optimal H and B can be derived by iterating through (12), (13) and (14) until convergence.

3.2 Summary and Complexity Analysis

The nested EM algorithm with belief propagation, named as NEMBP, is in Algorithm 1. In our experiments we set iteration steps $L_1 = 100$ and $L_2 = L_3 = 20$. Small numbers of iterations of $L_2 = L_3 = 20$ are adequate for obtaining reasonable results in the early stage of the method and sufficient for convergence to local optima in the late stage.

Algorithm 1: Nested EM with BP (NEMBP)

```
Input: A, X and c
Output: $q_i^r$'s, H and B
Initialize π, Θ, H and B randomly
For $l_1$ = 1: $L_1$ //main EM
    For $l_2$ = 1: $L_2$ //belief propagation
        Update beliefs via (7), (5) and (6)
    Get one-node marginal probabilities via $q_i^r = \psi_r^i$
    Calculate two-node marginal probabilities via (8)
    Update π and Θ via (9) and (10)
    For $l_3$ = 1: $L_3$ //nested EM
```

Update t_{rk}^{s}'s, H and B via (12), (13) and (14)

At a local optimum, we use: 1) q_i^r's (where q_i^r is the marginal posterior probability that node v_i belongs to community r) to identify the final network communities, 2) B (where β_{sk} is the probability that topic s selects the kth attribute) to extract the semantic for each content cluster, and 3) the correlation matrix H (where η_r is the distribution of the content clusters and their topics over community r) for finding the dominant topics for each community.

On sparse networks, the new method NEMBP is efficient. The time to update all messages once via (5) to (7) is $\sum_{i=1}^{n} d_i c^2(d_i + K_i) < 2ec^2(d_{\max} + K_{\max})$, $(2e+f)c^2$, and $2nc+c^2$, respectively, on a network of n nodes, e edges, and c communities, where d_i is the degree of node v_i, $d_{\max} = \max(d_1,\ldots,d_n)$, $K_i = \sum_{k=1}^{m} x_{ik}$ the number of attributes of v_i, $f = \sum_{i=1}^{n} K_i$ the number of attributes in contents, and $K_{\max} = \max(K_1,\ldots,K_n)$. The time to compute the two-node marginal probabilities once via (8) is ec^2. (Note that, we only need to consider $\psi_r^{i\to j}$'s and q_{ij}^{rs}'s with $a_{ij} = 1$.) The time to calculate π and Θ once via (9) and (10) is $nc+2c^2(e+n)$. The time to calculate t_{rk}^{s}'s, H and B once via (12) to (14) is $mc^2+ec^2+ fc^2$. The time to compute the full likelihood once is $2nc+2ec^2+fc^2$. Finally, the total complexity is no more than $O((e(d_{\max}+K_{\max})+f)c^2)$, which is nearly linear for large sparse networks. On a workstation (Intel(R) Xeon(R) CPU E3-1225 v3 @3.2GHz 3.2GHz processor with 16 Gbytes of main memory) running MatLab, NEMBP finished in 24 seconds on the "PubMed" dataset with 19,729 nodes (see Table 1).

4. Evaluation and Applications

We evaluated the new method in three different ways, on artificial benchmarks, on an online music system to assess the interpretability of communities, and on 10 real networks with comparison to eight state-of-the-art methods.

4.1 Artificial Benchmarks

The first benchmark we used was the Newman's model (Girvan and Newman 2002) for random networks. The networks have 128 nodes divided into 4 communities where each node has on average z_{in} edges (i.e., internal degree) connecting to nodes of the same community and z_{out} edges (i.e., external degree) to nodes of other communities, and $z_{in} + z_{out} = 16$. Note that p_{in} ($= z_{in}/32$) $> p_{out}$ ($= z_{out}/96$), so that the internal degrees are more likely greater than the external degrees. We generated a $4h$-dimensional binary attributes (i.e., x_i) for each node v_i to form 4 content clusters of nodes, corresponding to the 4 network communities. To be specific, for every node in the sth cluster, we use a binomial distribution with mean $\rho_{in} = h_{in}/h$ to generate a h-dimensional binary vector as its $((s-1) \times h + 1)$-th to $(s \times h)$-th attributes, and generated the rest attributes using a binomial distribution with mean $\rho_{out} = h_{out}/(3h)$. Since $\rho_{in} > \rho_{out}$, the h-dimensional attributes are associated with the sth cluster with a higher probability, whereas the rest $3h$ attributes are irrelevant. In our experiments, we set $4h = 200$ and the average number of attributes w_k with $x_{ik} = 1$ for each node v_i to $h_{in} + h_{out} =16$.

In the first experiment, we set $z_{out} = h_{out} = 8$ and generated networks with topologies and contents sharing the same community memberships. We then randomly selected a proportion (p_{mis}) of nodes and swapped their attribute vectors. The larger p_{mis} is, the more content clusters mismatch with network communities. We varied p_{mis} from 0 to 1 with an increment of 0.1, and tested our new method NEMBP. As shown in Figure 2, when p_{mis} is small, i.e., content clusters match well with communities, NEMBP significantly outperforms the degree-corrected stochastic blockmodel (DCSBM) (Karrer and Newman 2011). (DCSBM can be regarded as a variant of our method using topologies only.) Even when p_{mis} is large, NEMBP can still utilize the content information to improve the results. When node contents are not informative of communities (i.e., $p_{mis} = 1$), they can be ignored, and the final result will be similar to that from DCSBM. Besides, the NMI (normalized mutual information) accuracy (Danon et al. 2005) of NEMBP is also greater than that of SCI method that uses both topologies and contents (Wang et al. 2016). This also showed that NEMBP can better utilize information of mismatched contents.

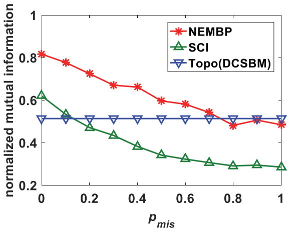
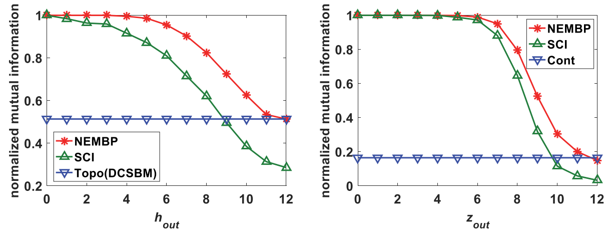


Figure 2: *The NMI accuracies of 3 methods compared on random networks as a function of the fraction (p_{mis}) of nodes with mismatched community and contents memberships. Topo (DCSBM) is a variant to our method using topologies only. SCI uses topologies and contents.*



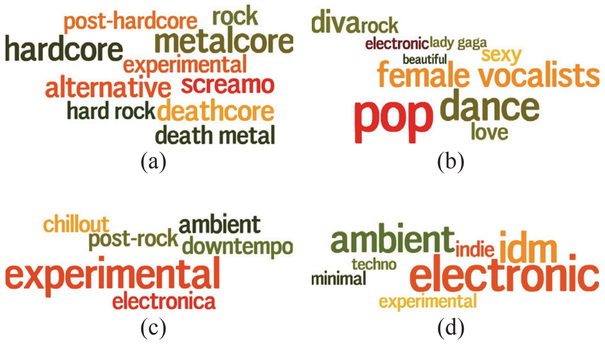
(a) (b)

Figure 3: *NMI accuracies of 3 methods on random networks as a function of **(a)** the number (h_{out}) of irrelevant or noise attributes and **(b)** the average outside-community degree (z_{out}) of nodes. Cont is our method (NEMBP) using contents information only. Each point in the figure is averaged over 50 problem instances.*

We now consider the situation when node contents have poor cluster structures. For simplicity we set $p_{mis} = 0$ to have the topologies and contents share the same membership, although $0 < p_{mis} << 1$ will not affect the results. In our experiment, we first set $z_{out} = 8$, and varied h_{out} from 0 to 12 with an increment 1. The larger h_{out} is, the less structures in node contents. When $h_{out} = 12$ (i.e., $\rho_{in} = \rho_{out}$), the contents do not have any cluster structure. In this special case our method NEMBP did not perform worse than the baseline method DCSBM which only uses topology information, whereas SCI performed much worse (Figure 3(a)). The results also showed that NEMBP was able to exploit the structural information in node contents when $h_{out} < 12$ to better detect communities, and as a result outperformed SCI and DCSBM (Figure 3(a)). We then set $h_{out} = 8$ and varied z_{out} from 0 to 12. NEMBP was also able to fare well when the networks do not have any community structure when $z_{out} = 12$ (Figure 3(b)). In short, this new method can combine the information of communities and node contents to find better communities than the other methods.

4.2 A Case Study

We assessed whether the derived correlation $\mathrm{H} = (\eta_{rs})_{c \times c}$ between communities and content clusters (with their topics) can help to better interpret the communities. To this end, we used the LASTFM dataset (Cantador 2016) from an online music system *Last.fm*, which has 1,892 users connected in a social network of "friends". Each user is described by 11,946 attributes, including a list of most listened music artists and tag assignments. Since no ground-truth is known regarding user communities in the network, we set forth to look for 38 communities as did in (Wang et al. 2016).



(a) (b)

(c) (d)

Figure 4: *Three examples on community interpretation in the music domain. Shown are the word clouds of the dominant attributes of communities. Word sizes are proportional to the probability they belong to a topic.* ***(a)*** *"topic 3, hardcore punk music" is the main topic of the 33th community.* ***(b)*** *"topic 37, Lady Gaga" is the main topic of the 14th community.* ***(c)*** *"topic 17, downtempo" and* ***(d)*** *"topic 25, intelligent dance music" are types of "electronic music" and the two main topics of the 15th community.*

The 38 communities discovered form two groups. The first has 28 communities, each of which has one dominant topic, and the second has 10 communities, each of which has more than one topic. A close analysis of the results revealed that most of the communities are semantically meaningful and well supported by their topics. Due to limited space, we discuss here three examples of the results: two communities each with one dominant topic, and one community with two topics, shown in Figure 4.

The first example is a community that has one dominant topic. As shown in Figure 4(a), this is a group of fans of hardcore punk music. Therefore, "screamo" is a kind of hardcore punk, "post-hardcore" is evolved from hardcore punk, and "deathcore", "death metal" and "metalcore" are the same as hardcore punk in terms of styles and characteristics of loud noise. The nature of the hardcore punk is "alternative" and "experimental", and it is also labeled as "hardcore", "rock" and "hard rock".

The second example is a community whose dominant topic is closely related to Lady Gaga, a well-known female singer, so that the community may be a group of her fans (Figure 4(b)). Lady Gaga is "female vocalists", "diva", "sexy" and "beautiful". Her music is also known to have styles of "pop", "dance", "rock", "electronic" and "love".

The third example is a community with two dominant topics (Figures 4(c) and 4(d)). One topic is highly related to downtempo, a kind of electronic music (Figure 4(c)). This music style is slow and, while being similar to "chill-out", has a little more beats than "chill-out" and "ambient", and inclines to "instrumental". "Post-rock", another name for "experimental rock", is also part of the nature of downtempo. The other topic is on idm (i.e., "intelligent dance music"), a kind of electronic music that comes from the dance floor (Figure 4(d)). It contains hard edge dance and slow beat music (like "ambient"). In addition, "minimal" and "techno" are also similar to idm. Note that these two topics on downtempo and idm belong to electronic music of different branches. Thus, this community is a group of fans of downtempo and idm who like some electro music.

In summary, this case study not only validated that the new method can find semantically meaningful communities, but also showed that allowing more than one topic per community can help better interpret and understand network communities as illustrated by the third example.

4.3 Applications to Real Networks

We applied the new method NEMBP to 10 real networks with known communities (Table 1). We considered three types of existing methods for comparison. The first, including DCSBM (Karrer and Newman 2011) and BigCLAM

(Yang and Leskovec 2013), uses information of structures alone. The second, including LDA (Blei, Ng and Jordan 2003), employs only node attributes. The third, including Block-LDA (Balasubramanyan and Cohen 2011), PCL-DC (Yang et al. 2009), CESNA (Yang, McAuley and Leskovec 2013), DCM (Pool, Bonchi and Leeuwen 2014) and SCI (Wang et al. 2016), uses information of structures and contents. All these methods require the number of communities to be specified, which were set to the same number of communities of the ground truth. These algorithms ran with their default parameters.

Table 1: *Dataset used. n is the number of nodes, e the number of edges, m the number of attributes, c the number of communities.*

Datasets	*n*	*e*	*m*	*c*	Descriptions (Sen et al. 2008; Leskovec 2016)
Texas	187	328	1,703	5	The WebKB network consists of four subnetworks from four American universities, which are Texas, Cornell, Washington and Wisconsin, respectively
Cornell	195	304	1,703	5	
Washington	230	446	1,703	5	
Wisconsin	265	530	1,703	5	
Twitter	171	796	578	7	Largest subnetwork (id 629863) in Twitter data
Facebook	1,045	26,749	576	9	Largest subnetwork (id 107) in Facebook data
Citeseer	3,312	4,732	3,703	6	A Citeseer citation network
Cora	2,708	5,429	1,433	7	A Cora citation network
UAI2010	3,363	45,006	4,972	19	A Wikipedia articles network
Pubmed	19,729	44,338	500	3	Publications in PubMed on diabetes

Table 2: *Comparison on disjoint communities in terms of AC and NMI. The best results are in bold. N/A means out of memory.*

Metrics (%)	Methods	Datasets									
		Texas	Cornell	Washington	Wisconsin	Twitter	Facebook	Citeseer	Cora	UAI2010	Pubmed
AC	DCSBM	48.09	37.95	31.80	32.82	60.49	45.19	26.57	38.48	2.60	53.64
	LDA	56.28	44.62	**65.90**	**76.72**	37.04	31.59	31.34	37.19	34.07	46.30
	Block-LDA	54.10	46.15	39.17	49.62	35.80	37.66	24.35	25.52	16.04	49.01
	PCL-DC	38.80	30.26	29.95	30.15	56.79	51.04	24.85	34.08	28.82	63.55
	SCI	**62.30**	45.64	51.15	50.38	50.62	51.04	27.98	40.62	30.94	*N/A*
	NEMBP	53.55	**47.17**	42.85	63.35	**62.96**	**56.27**	**49.51**	**57.57**	**46.25**	**65.66**
NMI	DCSBM	16.65	9.69	9.87	3.14	57.48	43.38	4.13	17.07	31.21	12.28
	LDA	31.29	**21.09**	**38.48**	**46.56**	31.10	21.53	9.13	14.61	35.42	10.55
	Block-LDA	4.21	6.81	3.69	10.09	0	9.28	2.42	1.41	5.70	6.58
	PCL-DC	10.37	7.23	5.66	5.01	52.64	38.63	2.99	17.54	26.92	26.84
	SCI	17.84	11.44	12.37	17.03	43.00	30.01	4.87	19.26	24.80	*N/A*
	NEMBP	**35.12**	18.71	21.24	38.02	**59.73**	**47.52**	**24.27**	**44.08**	**47.21**	**28.30**

Table 3: *Comparison on overlapping community structures in terms of GNMI, F-score and Jaccard.*

Metrics (%)	Methods	Datasets									
		Texas	Cornell	Washington	Wisconsin	Twitter	Facebook	Citeseer	Cora	UAI2010	Pubmed
GNMI	BigCLAM	0.75	0.58	0.77	0.44	14.04	21.54	0	0	11.94	0.57
	CESNA	0.69	2e-14	0.32	2e-14	15.53	**27.02**	2e-14	2.64	7.59	0
	DCM	1.17	0	0.17	0.51	1.75	22.42	0	1e-14	2.80	2e-14
	NEMBP	**7.83**	**5.99**	**7.08**	**15.74**	**26.94**	26.74	**9.72**	**26.65**	**18.98**	**22.82**
F-score	BigCLAM	20.64	13.23	13.35	12.84	39.79	40.06	9.30	18.89	16.99	7.72
	CESNA	23.54	23.48	21.91	23.17	43.72	49.05	3.38	31.05	32.32	27.97
	DCM	11.15	14.38	12.45	10.45	10.57	39.21	2.50	3.43	9.65	0.38
	NEMBP	**38.56**	**41.73**	**38.92**	**50.00**	**49.07**	**51.25**	**46.41**	**56.55**	**43.43**	**64.45**
Jaccard	BigCLAM	12.18	7.18	7.25	7.01	26.13	28.94	5.01	10.89	9.87	4.04
	CESNA	13.57	13.47	12.40	13.14	29.63	**38.18**	1.73	19.10	21.26	16.26
	DCM	6.03	7.95	6.72	5.54	5.75	28.46	1.27	1.76	5.77	0.19
	NEMBP	**26.20**	**27.54**	**25.10**	**36.75**	**36.08**	37.91	**31.14**	**43.09**	**30.92**	**48.62**

Because the networks have known communities, we adopted accuracy (AC) (Liu et al. 2012) and normalized mutual information (NMI) (Danon et al. 2005) to compare all the methods against the ground truth. To accommodate overlapping communities, we also included the generalized NMI (GNMI) (Lancichinetti, Fortunato and Kertész 2009) for comparison. We adopted the metric of (Yang, McAuley and Leskovec 2013) for overlapping communities, i.e., we evaluated a set of detected communities C with the ground-truth communities C^* by $\frac{1}{2|C^*|}\sum_{C_i^* \in C^*} max_{C_j \in C}\delta(C_i^*, C_j) + \frac{1}{2|C|}\sum_{C_j \in C} max_{C_i^* \in C^*}\delta(C_i^*, C_j)$, where $\delta(C_i^*, C_j)$ is a similarity measure (F-score or Jaccard) between C_i^* and C_j.

Our method NEMBP outperformed all 5 existing methods on 7 of the 10 networks in terms of AC and NMI (Table 2). It was also the best on 9, 10 and 9 of the 10 networks in terms of GNMI, F-score and Jaccard, respectively (Table 3). It was among the top two except on Texas measured by AC.

5. Conclusion and Discussion

We proposed a generative model for attributed networks, and developed a novel and efficient learning method using a nested EM algorithm with belief propagation to train the model. The model describes communities and content clusters using separate hidden variables, and extracts and explores the latent correlation between the two to better identify communities. It is able to fully utilize network structural information even when information of node contents is erroneous. The learned correlation between communities and content clusters (as well as the topics) can be used to extract community semantics, sometimes more than one topic per community, so as to better understand and interpret the community. We evaluated the new method on artificial benchmarks and in a case study. The new method outperformed eight state-of-the-art community-finding methods on most of 10 large real complex networks.

The strong performance of the new method is not obvious, since it would be possible that combining structural and semantic information may possibly make a community-finding method ineffective. Indeed, LDA which uses information of node attributes alone outperforms several methods that use both structural and content information. The superior performance of the new methods is mainly due to three properties: 1) even when node contents does not match well with community structures, it is still able to utilize content information as much as possible to improve community detection; 2) when either the network or node contents contains no information of community structure, it can proceed with whatever information available; 3) it has no parameter to tune to balance the effects of network structures and node contents in training the model.

Similar to the existing methods, our method needs the number of communities to be given. In addition, the number of communities may be different from the number of content clusters (and topics). These issues are related to the problem of model selection, which may be addressed using cross-validation (Chen and Lei 2014) or hierarchical

Bayesian (Jin et al. 2016), which are topics of our on-going research.

Acknowledgments

Zhiyong Feng is the corresponding author. The work was supported by Natural Science Foundation of China (61502334, 61303110, 61373035, 61133011), the Talent Development Program of Wuhan, the municipal government of Wuhan, Hubei, China (2014070504020241), and an internal research grant of Jianghan University, Wuhan, China, as well as by United States National Institutes of Health (R01GM100364).

References

Balasubramanyan, R.; and Cohen, W. W. 2011. Block-LDA: Jointly modeling entity-annotated text and entity-entity links. *In Proceedings of 11th SIAM International Conference on Data Mining(SDM'11)*, 450-461.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3: 993-1022.

Blondel, V. D.; Guillaume, J.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008: P10008.

Cantador, I. 2016. The 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems. http://ir.ii.uam.es/hetrec2011/datasets.html

Chen, K.; and Lei, J. 2014. Network cross-validation for determining the number of communities in network data. arXiv:1411.1715

Danon, L.; Diazguilera, A.; Duch, J.; and Arenas, A; 2005. Comparing community structure identification. *J. Stat. Mech.* 2005: 09008.

Decelle, A.; Krzakala, F.; Moore, C.; and Zdeborova, L. 2011. Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.* 107: 065701.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39: 1-38.

Duch, J.; and Arenas, J. 2005. Community detection in complex networks using extremal optimization. *Phys. Rev. E* 72: 027104.

Fortunato, S.; and Hric, D. 2016. Community detection in networks: A user guide. arXiv:1608.00163

Fortunato, S. 2010. Community detection in graphs. *Phys. Rep.* 486: 75-174.

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.; Liu, D.; Jin, D.; and Zhang, W. 2015. A stochastic model for the detection of heterogeneous link communities in complex networks. *In Proceedings of the 29th AAAI Conference on Artificial Intelligence(AAAI'15)*, 130-136. Palo Alto, California, USA: AAAI Press.

Jin, D.; Wang, H.; Dang, J.; He, D.; and Zhang, W. 2016. Detect overlapping communities via modeling and ranking node popularities. *In Proceedings of the 30th AAAI Conference on Artificial Intelligence(AAAI'16)*, 172-178. Palo Alto, California, USA: AAAI Press.

Karrer, B.; and Newman M. E. J. 2011. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* 83: 016107.

Lancichinetti, A.; Fortunato, S.; and Kertész, J. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* 11: 033015.

Leskovec, J. 2016. Stanford Network Analysis Project. http://snap.stanford.edu

Liu, H.; Wu, Z.; Li, X.; Cai, D.; and Huang, T. S. 2012. Constrained nonnegative matrix factorization for image representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 34: 1299-1311.

Liu, L.; Xu, L.; Wang, Z.; and Chen, E. 2015. Community detection based on structure and content: A content propagation perspective. *In Proceedings of the 15th IEEE International Conference on Data Mining(ICDM'15),* 271-280. Piscataway, NJ, USA: IEEE Press.

Martin, T.; Ball, B.; and Newman, M. E. J. 2016. Structural inference for uncertain networks. *Phys. Rev. E* 93: 012306.

Mcauley, J.; and Leskovec, J. 2014. Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data* 8: 73-100.

Newman, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69: 066133.

Pei, Y.; Chakraborty, N.; and Sycara, K. 2015. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence(IJCAI'15)*, 2083-2089. San Francisco, California: Morgan Kaufmann Press.

Pool, S.; Bonchi, F.; and Leeuwen, M. 2014. Description-driven community detection. *ACM Trans. Intell. Syst. Technol.* 5: Article No. 28.

Rosvall, M.; and Bergstrom, C. 2008. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* 105: 1118-1123.

Ruan, Y.; Fuhry, D.; and Parthasarathy, S. 2013. Efficient community detection in large networks using content and links. *In Proceedings of the 22nd International World Wide Web Conference(WWW'13)*, 1089-1098. New York, NY: ACM Press.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Gallagher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI Magazine*. 29: 93-106.

Sun, Y.; Aggarwal, C. C.; and Relation, J. H. 2012. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *In Proceedings of the 37th International Conference on Very Large Data Bases(VLDB'12)*, 394-405. New York, NY: ACM Press.

Wang, X.; Jin, D.; Cao, X.; Yang, L.; and Zhang, W. 2016. Semantic community identification in large attribute networks. *In Proceedings of the 30th AAAI Conference on Artificial Intelligence(AAAI'16)*, 172-178. Palo Alto, California, USA: AAAI Press.

White, S. R.; and Smyth, P. 2005. A spectral clustering approach to finding communities in graphs. *In Proceedings of 5th SIAM International Conference on Data Mining(SDM'05)*, 76-84.

Xu, Z.; Ke, Y.; Wang, Y.; Cheng, H.; and Cheng, J. 2012. A model-based approach to attributed graph clustering. *In Proceedings of the 33rd ACM Conference on Management of Data(SIGMOD'12)*, 505-516. New York, NY: ACM Press.

Yang, T.; Jin, R.; Chi, Y.; and Zhu, S. 2009. Combining link and content for community detection: A discriminative approach. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD'09)*, 927-936. New York, NY: ACM Press.

Yang, J.; and Leskovec, J. 2013. Overlapping community detection at scale: A nonnegative matrix factorization approach. *In Proceedings of the 6th ACM International Conference on Web Search and Data Mining(WSDM'13)*, 587-596. New York, NY, USA: ACM Press.

Yang, J.; and Leskovec, J. 2014. Overlapping communities explain core-periphery organization of networks. *Proceedings of the IEEE* 102: 1892.

Yang, J.; McAuley, J.; and Leskovec, J. 2013. Community detection in networks with node attributes. *In Proceedings of the 13th IEEE International Conference on Data Mining(ICDM'13)*, 1151-1156. Piscataway, NJ, USA: IEEE Press.