Detect Overlapping Communities via Ranking Node Popularities

Di Jin[1], Hongcui Wang[1], Jianwu Dang[1,2], Dongxiao He[1], 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, hcwang, dangjianwu, hedongxiao}@tju.edu.cn, weixiong.zhang@wustl.edu

Abstract

Detection of overlapping communities has drawn much attention lately as they are essential properties of real complex networks. Despite its influence and popularity, the well studied and widely adopted stochastic model has not been made effective for finding overlapping communities. Here we extend the stochastic model method to detection of overlapping communities with the virtue of autonomous determination of the number of communities. Our approach hinges upon the idea of ranking node popularities within communities and using a Bayesian method to shrink communities to optimize an objective function based on the stochastic generative model. We evaluated the novel approach, showing its superior performance over five state-of-the-art methods, on large real networks and synthetic networks with ground-truths of overlapping communities.

1. Introduction

One of the challenging problems in study of complex networks, e.g., social networks, biological networks, and the world wide web, is the detection of community structures (Girvan and Newman 2002), a subject that has attracted a great deal of interest. A community within a network can be loosely defined as a set of nodes that are densely connected with respect to the rest of the network. So far, many different approaches have been proposed to uncover community structures in networks, as reviewed in (Fortunato 2010). Among the existing methods, the most popular are the ones that focus on partition of nodes, resulting in communities of disjoint sets of nodes and a node being belong to only one community (Girvan and Newman 2002). However, overlaps of communities, i.e., a node can be members of more than one community, are ubiquitous in reality (Palla et al. 2005). For example, an individual has a family and belongs to a group of co-workers, each of which has its own functions and forms its own circle of influence. Forcing a node into one community fail to accommodate multiple relationships and functions that a node may possess, resulting in erroneous representation of the network structure. Thus, it is imperative to develop methods that allow nodes to be members of multiple communities.

A few approaches have been proposed for overlapping community detection. One of them is based on the idea of clique percolation where a cluster is interpreted as the union of small, fully connected subgraphs that share common nodes (Palla et al. 2005; Shen et al. 2009). Another approach utilizes local expansion and optimization which is based on growing a natural community (Lancichinetti, Fortunato and Kertész 2009; Jin et al. 2011; Lancichinetti et al. 2011). Most of these methods rely on a local benefit function that characterizes the quality of a densely connected group of nodes. The third approach considers link community detection by partitioning links instead of nodes, where a node associated with different types of links may belong to different communities (Ahn, Bagrow and Lehmann 2010; Gopalan and Blei 2013; He et al. 2015). The fourth approach is based on dynamic label propagation (Raghavan, Albert and Kumara 2007), which has been extended to overlapping community detection (Gregory 2010; Xie, Szymanski and Liu 2011). In the process of label propagation, each node updates its community belonging coefficients by repeatedly averaging the coefficients passing from all its neighbors.

Thanks to a sound theoretical basis and good performance, the stochastic model constitutes a promising technique for identifying modular structures of networks and has been well studied recently. Furthermore, the model-based methods that maintain probabilistic community memberships have also been adopted to find overlapping community structures (Airoldi et al. 2008; Shen, Cheng and Guo 2011; Psorakis et al. 2011; Zhang and Yeung 2012; Yang and Leskovec 2013; Jin et al. 2015).

Despite their popularity, the model-based methods suffer from some common drawbacks. First, highly connected nodes in real networks are likely to play important roles and thus tend to appear in multiple communities (Yang and Leskovec 2014). The existing model-based methods do not model this phenomenon well since they assume that the sum of probabilities for a node belonging to different communities (or community membership of the node) is 1

(Yang and Leskovec 2014). This assumption is often not satisfied in reality. For example, a node may belong to the community 'Politics' with probability 0.9 and the community 'Economics' with probability 0.8, as the two communities are highly correlated. While it has been proposed to remove this assumption to better model overlapping communities (Zhang and Yeung 2012; Yang and Leskovec 2013), without the constraint the community memberships of two nodes may have different scales and the actual meaning of node community membership will be unclear.

Second, the model-based methods typically require a threshold on probabilistic memberships, which is difficult to determine in practice, in order to derive overlapping structures (Xie, Kelley and Szymanski 2013). It has been proposed to use some community metrics, e.g., modularity Q and AUC score (Zhang and Yeung 2012; Jin et al. 2015), or some empirical methods (Airoldi et al. 2008; Yang and Leskovec 2013) to determine the threshold. All these methods introduce additional factors that may not be necessarily consistent with the models to be built.

Third, the model-based methods require the number of communities c, which is often unknown in practice, to be specified. This problem has been addressed by multiple runs of the same method with different c's and choosing the result with the highest fitness based on some suitable quality metric, such as cross validation (Airoldi et al. 2008; Yang and Leskovec 2013), minimum description length (Shen, Cheng and Guo 2011) or consensus clustering (Jin et al. 2015). Unfortunately, this multi-run scheme makes the model-based methods inefficient on and unscalable to large networks. This problem has also been considered using a Bayesian approach in a single run (Psorakis et al. 2011). The method starts with a large initial value of c, and uses a prior to penalize the model for including many non-zero parameter values to balance the number of communities and the fitness of the model to the given data. However, the prior contains two hyperparameters that are nontrivial to determine. As a result, the problem remains unsolved.

Here we propose and develop a novel approach to extend the stochastic model method to find overlapping communities and autonomously determine the number of communities at the same time. The paper is organized as follows. In Section 2, we start with a description of the stochastic model to be learned from a given network. We then consider deriving overlapped communities using the model. We subsequently adopt a hierarchical Bayesian approach to determine the number of communities. We shift our attention to experimental evaluation of the new approach in Section 3. We extensively compare our method with five existing methods on large real networks and synthetic networks. We conclude with some discussions.

2. The Model and Method

Here, we first describe a stochastic model to be learned from a given network. Reciprocally given the model, the given network can be viewed as most likely to be generated from the model; in other words, the model is a generative model of the network. We then consider overlapping communities by inferring the importance of a node in all fixed number of communities. To remove the restriction of the number of communities c being provided, we introduce a hierarchical Bayesian approach to determine c in a single run. We complete this section with a complexity analysis.

2.1 The Stochastic Generative Model

A network G with n nodes can be represented by an adjacency matrix $A = (a_{ij})$ with $a_{ij} = 1$ if an edge exists between nodes i to j, or 0 otherwise. We model G by an ensemble of c *probabilistic* communities $\{G_1, G_2, \ldots, G_c\}$. This can be viewed as a generative model, from which the observed network G might have been generated.

The model is specified by two sets of parameters $D = (d_{ik})$ and $\Theta = (\theta_{kj})$, where d_{ik} denotes the expected degree of node i in the kth community G_k, and θ_{kj} the probability that G_k selects node j when generating an edge, which is also taken as the importance of node j in community G_k based on the rationale that the more important a node to a community, the more likely the node is included in the community. As θ_{kj} captures the propensity of node j belonging to community k, we have $\sum_{j=1}^{n} \theta_{kj} = 1$.

Since G_k describes an unseparable community, it is a random graph with little structure. Thus the expected number of links between nodes i to j in G_k is $d_{ik}\theta_{kj}$, meaning that G_k randomly selects node j by d_{ik} times with probability θ_{kj} when it generates links from i to j. In total, the expected number of links from node i to node j is:

$$\hat{a}_{ij} = \sum_{k=1}^{c} d_{ik}\theta_{kj} . \quad (1)$$

Using a Poisson distribution that corresponds to the KL-divergence, the log probability of generating a graph G with adjacency matrix $A = (a_{ij})_{n\times n}$ by the model in (1) is:

$$L = \log P(A \mid D, \Theta) = \sum_{ij} a_{ij} \log\left(\sum_k d_{ik}\theta_{kj}\right) - \sum_{ijk}\left(d_{ik}\theta_{kj}\right), \quad (2)$$

where the additive constants are ignored. This log likelihood describes the best fit between the expected network from the model and the observed network. The parameters can be learned by maximizing this log likelihood function.

Since direct maximizing (2) is nontrivial, we adopt an expectation-maximization algorithm. By applying Jensen's inequality to (2), we construct an auxiliary function as:

$$\bar{L}(d_{ik}, \theta_{kj}; q_{ij,k}) = \sum_{ijk}\left(a_{ij} q_{ij,k} \log \frac{d_{ik}\theta_{kj}}{q_{ij,k}} - d_{ik}\theta_{kj} \right) \le L(d_{ik}, \theta_{kj}) , \quad (3)$$

where the introduced probabilities $q_{ij,k}$ can be freely chosen, provided that they satisfy $\sum_k q_{ij,k} = 1$. Thus $\bar{L}$ is a lower bound of L and the equality holds when

$$q_{ij,k} = d_{ik}\theta_{kj} / \sum_r d_{ir}\theta_{rj} \quad (4)$$

To maximize L in the E-M algorithm, assume the current estimation of d_{ik} and θ_{kj} to be $\hat{d}_{ik}$ and $\hat{\theta}_{kj}$. We have $L(\hat{d}_{ik}, \hat{\theta}_{kj}) = \bar{L}(\hat{d}_{ik}, \hat{\theta}_{kj}; \hat{q}_{ij,k})$, where $\hat{q}_{ij,k}$ is derived from $\hat{d}_{ik}$ and $\hat{\theta}_{kj}$ using (4). We then maximize $\bar{L}$ with respect to d_{ik} and θ_{kj} under $\sum_j \theta_{kj} = 1$ with $\hat{q}_{ij,k}$ fixed. Introducing Lagrange multipliers γ_k to incorporate these constraints, we have the Lagrange form of $\bar{L}$

$$\tilde{L} = \bar{L} + \sum_k \gamma_k \left(\sum_j \theta_{kj} - 1 \right). \quad (5)$$

By taking partial derivative of $\tilde{L}$ in (5), we obtain

$$d_{ik} = \sum_j a_{ij} \hat{q}_{ij,k}; \;\; \theta_{kj} = \sum_i a_{ij} \hat{q}_{ij,k} / \sum_{is} a_{is} \hat{q}_{is,k}. \quad (6)$$

Therefore, we have $\bar{L}(d_{ik}, \theta_{kj}; \hat{q}_{ij,k}) \geq \bar{L}(\hat{d}_{ik}, \hat{\theta}_{kj}; \hat{q}_{ij,k})$. Next, we re-estimate the value of $q_{ij,k}$ using d_{ik} and θ_{kj}, which leads to $L(d_{ik}, \theta_{kj}) = \bar{L}(d_{ik}, \theta_{kj}; q_{ij,k}) \geq \bar{L}(d_{ik}, \theta_{kj}; \hat{q}_{ij,k}) \geq \bar{L}(\hat{d}_{ik}, \hat{\theta}_{kj}; \hat{q}_{ij,k}) = L(\hat{d}_{ik}, \hat{\theta}_{kj})$. By alternating between (4) and (6), the objective function L monotonically converges to a local maximum.

The derivation above applies to directed networks. We can generalize this model to undirected networks by introducing $\theta_{kj} = d_{jk} / \sum_s d_{sk}$, similar to the undirected configuration model (Jin et al. 2015). The derivation follows the case for directed and the results are the same as (4) and (6).

2.2 Inference of Overlapping Communities

We now consider inferring overlapping community structures for a given network by exploiting its model.

To reiterate, the stochastic model is specified by two sets of parameters, D and Θ. Consider first Θ. The kth row of Θ, θ_k, represents the memberships and, in essence, captures the relative importance of all nodes in the kth community. We thus sort the nodes in a decreasing order of θ_k and record the corresponding node order as I_k, where the most and least important nodes in community k appear the first and last in I_k. For clarity, we denote the jth node in this order as I_{kj}. We now consider D. Note that $\sum_i d_{ik} = s_k$ is the expected degree of all nodes in the kth community. We use s_k as a threshold to determine which nodes should be included to form a *natural* community. Specifically, we add to the community the nodes in I_k one by one starting from the first, until the sum of actual degrees of the chosen nodes exceeds s_k. Then, the members of the kth community are

$$O_k = \left\{ I_{kj} \,|\, 1 \leq j \leq pos, \; \sum_{i=1}^{pos-1} K_i < s_k \leq \sum_{i=1}^{pos} K_i \right\}, \quad (7)$$

where K_i is the actual degree of node i.

Note that a node may have large values of θ_{kj}'s in multiple communities simultaneously, thus rank highly in each of them, and consequently be included in more than one community, creating a potentially overlapping community structure $O = \{O_1, O_2, \ldots, O_c\}$. Furthermore, when setting $\theta_{kj} = d_{jk} / \sum_s d_{sk}$ the nodes with large degrees are more likely to reside in multiple communities. This is consistent with what was observed on most real networks previously (Yang and Leskovec 2014). Importantly, which nodes are included in a community is autonomously determined based on the model learned. This eminent feature sets apart the new method from the existing ones that often require a predefined threshold on community memberships.

To better deal with some bridge nodes, we may refine the resultant community structure O using a greedy optimization on a well-known local community function, *conductance* (Yang and Leskovec 2014). To be specific, for each community O_k, we first find a node i from the neighbors of O_k, which will bring the highest increase (Δ_{+i}) of its community quality when adding i to O_k; We then find the node j in O_k which will bring the highest increase (Δ_{-j}) of its community quality when removing j from O_k; We set $O_k = O_k \cup \{i\}$ if $\Delta_{+i} \geq \Delta_{-j}$, or $O_k = O_k - \{j\}$ otherwise. This stops when no node can increase the community quality of O_k when ether adding or removing it.

2.3 Statistical Model Selection

Our method for autonomous determination of the targeted number of communities c is essentially a hierarchical Bayesian approach to statistical model selection. We first initialize c to a large value $c_{max} > c$. We then place some proper prior over the model parameters, to evaluate each parameterized community, and to shrink the communities that contribute little to the generation of the network. Removing irrelevant communities gives rise to the desired c.

Without loss of generality, assume that the parameters D and Θ are independent. First we consider D and define an independent exponential distribution for each column of D with rate parameter λ_k. Then, the log prior over D is:

$$\log P(D \,|\, \lambda) = \sum_k \sum_i \left(\log \lambda_k - \lambda_k d_{ik} \right). \quad (8)$$

Since exponential prior corresponds to a l_1-regularization favoring a sparse representation, each λ_k controls the degree of suppressing the kth column of D to zero and hence can be regarded as the weight of group sparsity.

Furthermore, the rate parameters $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_c\}$ cannot be prespecified, but can be learned from the data. To find the best λ_k's, we impose a non-informative Jeffreys' hyper prior on each λ_k to control the degree of the column-sparsity of D. Then, the log prior over λ is:

$$\log P(\lambda) = -\sum_k \log \lambda_k. \quad (9)$$

We choose non-informative prior as it expresses the property by itself and requires no additional parameter.

Finally we consider Θ. As the constraint $\sum_j \theta_{kj} = 1$ corresponds to an improper (un-normalizable) prior, we have:

$$\log P(\Theta) = \sum_k \delta \left(\sum_j \theta_{kj} - 1 \right), \quad (10)$$

where $\delta(.)$ denotes the Dirac delta function.

Using Bayes rule, log posteriori of D, Θ and λ given A is:

$$\begin{aligned} \log P(D, \Theta, \lambda \,|\, A) &= \log P(A \,|\, D, \Theta) \\ &+ \log P(D \,|\, \lambda) + \log P(\lambda) + \log P(\Theta) - \log P(A) \end{aligned} \quad (11)$$

where the last term is a constant. The first term is the likelihood of observing the data A given the model (or model parameters D and Θ), and the remaining terms are the priors of D and Θ. This Bayesian approach is shown graphically in Fig. 1(a). Furthermore, in order to optimize (11) efficiently, we replace $P(\Theta)$ by $\theta_{kj} = \phi_{kj} / \sum_s \phi_{ks}$ and convert the above MAP (maximum a posteriori) problem to an equivalent NMF (nonnegative matrix factorization) formulization with KL-divergence (Tan and Févotte, 2013):

$$\begin{aligned} &\arg\min_{D,\Phi>0} \; O(D, \Phi, \lambda) \\ &= \sum_{ij} \left(\sum_k \frac{d_{ik} \phi_{kj}}{\sum_s \phi_{ks}} - a_{ij} \log \left(\sum_k \frac{d_{ik} \phi_{kj}}{\sum_s \phi_{ks}} \right) \right). \\ &+ \sum_k \left(\lambda_k \sum_i d_{ik} \right) - (n-1) \sum_k \log \lambda_k \end{aligned} \quad (12)$$

The first term in (12) is the loss function for the distance between the reconstructed and the original networks. The second term is the l_1-regularization over each column of D with weight λ_k. Because each l_1-norm $\sum_i d_{ik}$ denotes the expected degree of a community, this term corresponds to the suppression of each community with different degrees. The third term is used to adaptively adjust the weights of λ_k's. Through the adaptive l_1-norm selection, the columns of D will be segregated into two groups under optimization of (12). One group includes columns whose l_1-norms are significantly larger than 0, whereas the other contains columns whose l_1-norm is close to 0; the number of columns with large l_1-norms is the anticipated number of communities. An illustrative example on the well-known "karate club" network (Zachary 1977) is shown in Fig. 1(b).

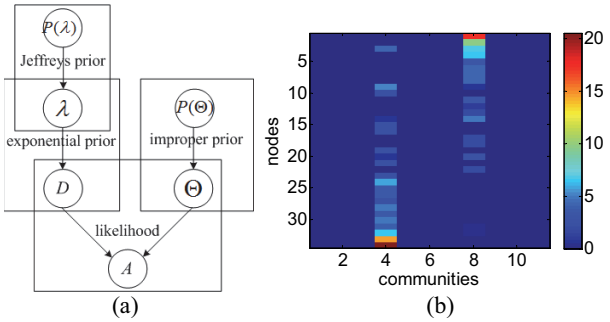


Figure 1: ***(a)*** *A graphical representation of our hierarchical Bayesian approach.* ***(b)*** *An illustratition of the result from this approach on Zachary's "karate club" network. Shown here is the learned matrix D. After the adaptive compression of the columns of D, only two columns (columns 4 and 8) remain nonzero, which is the targeted number of communities.*

Similar to the multiplicative update rules in the original NMF with KL-divergence (Lee and Seung 2000), we can construct an iterative procedure that reaches local minima and maintains nonnegativity of the parameters. First, we calculate gradients of $O(D, \Phi, \lambda)$ with respect to D, Φ and λ. Second, we update D and Φ by multiplying their current values with the ratio between the positive to the negative parts of the gradients. The update rules for D and Φ are:

$$d_{ik} = d_{ik} \frac{\sum_j \left(a_{ij} \theta_{kj} / \sum_r d_{ir} \theta_{rj} \right)}{1 + \lambda_k}$$
$$\phi_{kj} = \phi_{kj} \frac{\sum_i \left(a_{ij} d_{ik} / \sum_r d_{ir} \theta_{rj} \right)}{\sum_{is} \left(a_{is} d_{ik} \theta_{ks} / \sum_r d_{ir} \theta_{rs} \right)} \quad (13)$$

where $\theta_{kj} = \phi_{kj} / \sum_s \phi_{ks}$. We then update λ by setting its derivative equal to zero because it has analytic solution given D and Φ. Then, the updating rule for λ is:

$$\lambda_k = \frac{1}{\sum_i d_{ik} / (n-1)} . \quad (14)$$

Besides, if the kth column of D is suppressed to zero, λ_k will approach infinity. To avoid this, we add a small positive value ε (e.g., $\varepsilon = 10^{-3}$) to the numerator of (14) to have:

$$\lambda_k = \frac{1}{\varepsilon + \sum_i d_{ik} / (n-1)} . \quad (15)$$

We will perform some additional, detailed analysis of the parameter ε in the experiments.

The process of model selection is as follows. We set the initial number of communities to a large value (e.g., $c_{\max} = m/3$), optimize the objective function (12) by choosing a set of nonnegative initial values, and subsequently alternate between (13) and (15). The targeted number of communities c is the number of nonzero columns of D, i.e., we remove the irrelevant communities k whose expected degree $\sum_i d_{ik}$ is zero or very close to zero.

We can also extend this model selection to undirected networks by letting $\Phi = D$ in (12).

2.4 Complexity Analysis

The most time-consuming step of the new method is for updating D and Φ in (13). Since the adjacency matrix A is often sparse, the complexities to evaluate D and Φ once are $(4mc_{\max}+2nc_{\max}+2m+c_{\max})$ and $(4mc_{\max}+4nc_{\max}+2m)$, respectively, where n is the number of nodes, m the number of links, and $c_{\max}$ the initial number of communities. The total time complexity of the method is $O(Tmc_{\max})$, where T is the number of iterations to convergence. If an approximate initial value of $c_{\max}$ much smaller than $O(m)$ is available, the method will be nearly linear in network size.

3. Experimental Evaluation

To assess the performance of the proposed method, we evaluated it on real-world networks and synthetic benchmarks. We compared it with five state-of-the-art methods for overlapping community detection: i) CFinder (Palla et al. 2005), the most prominent algorithm using clique percolation theory; ii) Oslom (Lancichinetti et al. 2011), a local optimization method with an excellent performance especially on the LFR benchmarks; iii) SVI (Gopalan and Blei 2013), a new model-based method for detecting link communities; iv) BigClam (Yang and Leskovec 2013), a newly developed model-based method for finding overlapping communities using probabilistic memberships; and v) SLPA (Xie, Szymanski and Liu 2011), a representative algorithm based on dynamic label propagation.

Each of these methods has parameters to be adequately set. For CFinder, we set the clique size $k = 4$, which returns the best overall results (Palla et al. 2005). For Oslom, we used the default of 10 trial optimizations of the lowest hierarchical level, and selected the lowest hierarchical level as the resulting partition as suggested earlier (Lancichinetti et al. 2011). For SVI, following the guidelines in (Gopalan and Blei 2013), we assigned a link to a community if the approximate posterior probability of a link assignment to a community exceeded a threshold t, and took the best results from t=0.5 and t=0.9. Especially, for experiments on synthetic networks, we required at least three links of a node to be assigned to a community before assigning the node to that community. For BigClam, we used its default

values as in (Yang and Leskovec 2013). For SLPA, as suggested by (Xie, Szymanski and Liu 2011), we set the maximum number of iterations $T = 100$ and varied parameter r from 0.01 to 0.1 for synthetic networks and from 0.02 to 0.45 for real networks to determine the optimal value. Our method has an additional parameter ε which should be specified as a small positive value. We will introduce how to choose it in the following subsections.

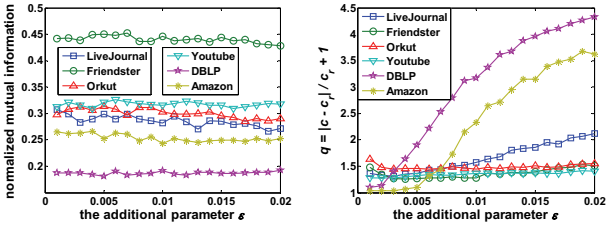
The accuracy of a community detection method was quantified by the level of correspondence between detected and ground-truth communities. The widely used normalized mutual information (NMI) index, which has been extended to overlapping community structures (Lancichinetti, Fortunato and Kertész 2009), was adopted as the accuracy measure in our study.

3.1 Real-World Networks

A practical issue in network structure analysis is the lack of the ground-truth of the underlying structures of a network. This issue is exacerbated on networks of overlapping structures since nodes in multiple communities often render ambiguous explanations. Fortunately, six real networks and their overlapping communities have been published by the Stanford Network Analysis Project (Leskovec 2015). These include four online social networks (LiveJournal, Friendster, Orkut and Youtube), one collaboration network (DBLP) and one information network (Amazon), where the communities, including overlapping ones, in each of these networks are explicitly labeled (see Table 1 for details).

The networks that were analyzed in our study were very large and beyond all currently available community detection methods. We thus adopted a sampling scheme to extract a large portion of a network with a manageable size. Following what was suggested earlier (Yang and Leskovec 2013), we randomly chose a node u in a given graph G belonging to at least two communities, and then took the subnetwork to be the induced subgraph of G consisting of all the nodes that share at least one known community membership with u. To obtain a credible subnetwork with well-defined overlapping community structures, we disregarded the subnetworks whose values of *overlapping modularity* (Shen et al. 2009) under the ground-truth were less than a threshold of $t = 0.1$, which can be considered as having no well-defined community structure. The result for each of the 6 datasets we tested was averaged on 500 randomly generated networks with overlapping communities.

Our method has only one parameter ε to be set, which should be a small positive value. For each network, we ran the method by varying ε from 1.0e-3 to 20e-3 with an increment of 1.0e-3, recorded the obtained number of communities c and the corresponding accuracy in the NMI index, and used $q = |c - c_r| / c_r + 1$ to represent the quality of the obtained c, where c_r is the actual number of communities. So, $q = 1$ corresponds to a perfect match between the actual number of communities and that obtained by the method. The experimental results showed that the NMI accuracy of the new method is not so sensitive to parameter ε on all six real networks (Fig. 2(a)) and the quality q of the obtained number of communities is also not too sensitive to ε on four real networks expect DBLP and Amazon (Fig. 2(b)). In sum, ε = 1.0e-3 typically corresponds to good performance on all of the six real networks for both the clustering accuracy and the number of communities. In the subsequent experiments ε was set to 1.0e-3.



(a) (b)

Figure 2: *Analysis of potential impact of parameter ε on six real-world networks.* ***(a)*** *The community accuracy represented by the NMI index and* ***(b)*** *the quality of the obtained number of communities calculated using $q = |c - c_r| / c_r + 1$ with the change of parameter ε. The larger the NMI index (NMI ≤ 1), the better the result. The smaller the q value (q ≥ 1), the better the result.*

Table 1: *Comparison of NMIs of different methods on six large Stanford networks with ground-truth of overlapping communities. Here, n is the number of nodes, m the number of links and c_r the number of communities. M denotes one million and k one thousand. The larger the NMI, the more closely the detected overlapping structure matches the ground truth. The best NMIs of these networks are in bold. We set ε = 1.0e-3 for our method.*

Datasets/ NMIs (%)	*n*	*m*	c_r	Methods					
				CFinder	Oslom	SVI	BigClam	SLPA	Ours
LiveJournal	4.0M	34.9M	310k	14.73	22.05	12.22	18.45	21.07	**30.75**
Friendster	120M	2,600M	1.5M	25.26	29.07	17.12	23.30	28.96	**44.19**
Orkut	3.1M	120M	8.5M	14.93	22.92	16.03	18.76	25.71	**29.75**
Youtube	1.1M	3.0M	30k	9.34	13.83	12.98	12.34	18.31	**31.21**
DBLP	0.43M	1.3M	2.5k	13.73	12.16	10.29	14.96	12.02	**18.78**
Amazon	0.34M	0.93M	49k	15.54	17.32	13.65	18.49	19.83	**26.47**

We compared the new method with five state-of-the-art methods. Quantified in NMI, our method outperformed all the other methods on all six networks (see Table 1). To highlight, this method is 8.70%, 15.12%, 4.04%, 12.90%, 3.82% and 6.64% more accurate than the second best on real networks of *LiveJournal*, *Friendster*, *Orkut*, *Youtube*, *DBLP* and *Amazon*, respectively. While the factors for such a superb performance remained to be further investigated, it may be partialy attributed to the way the relative node importance within a community was measured, especially for those nodes with large degrees, to the way which nodes to be included in a community, and to a hierarchical Bayes approach to determine the number of communities.

3.2 Synthetic Networks

A class of well-known synthetic benchmarks with overlapping community structures has been proposed by (Lancichinetti and Fortunato 2009), which is referred to as

LFR hereafter. The degree and the community size of a LFR graph follow power law distributions, a statistical property that most real-world networks seem to share.

The parameters of the LFR benchmarks were set following (Lancichinetti and Fortunato 2009). The network size n was set to 1000, the minimum community size c_{min} was either 10 or 20, the mixing parameter μ (each vertex shares a fraction μ of its edges with vertices in other communities) was set to either 0.1 or 0.3, the fraction of overlapping vertices (o_n/n) varied from 0 to 1 with an increment of 0.1. The remaining parameters were kept fixed: the average degree $d = 20$, the maximum degree $d_{max} = 2.5\times d$, the maximum community size $c_{max} = 5\times c_{min}$, the number of communities that each overlapping vertex belongs to $o_m = 2$, and the exponents of the power-law distributions of vertex degree $\tau_1 = -2$ and community size $\tau_2 = -1$.

For these four sets of benchmarks above, the parameter ε of the new method was set to 16e-3, 14e-3, 9e-3 and 9e-3, respectively. Again the results were not very sensitive to the variation of ε in a large range, while here we used the best parameter for comparison against other methods.

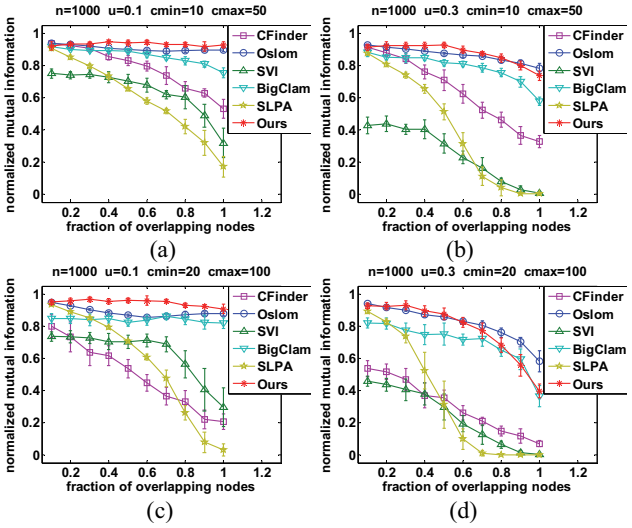


Figure 3: *NMIs of six methods compared as a function of the fraction of overlapping nodes. Error bars show the standard deviations estimated on 20 graphs. Shown are results on networks of* ***(a)*** *small mixing parameter & small communities (μ = 0.1, c_{min} = 10, c_{max} = 50),* ***(b)*** *big mixing parameter & small communities (μ = 0.3, c_{min} = 10, c_{max} = 50),* ***(c)*** *small mixing parameter & big communities (μ = 0.1, c_{min} = 20, c_{max} = 100) and* ***(d)*** *big mixing parameter & big communities (μ = 0.3, c_{min} = 20, c_{max} = 100).*

Fig. 3 shows the results that compare our method with CFinder, Oslom, SVI, BigClam and SLPA on the four sets of LFR benchmarks. As shown, our method and Oslom outperformed the other four methods in all four cases with our method being even slightly better than Oslom overall. The third most consistently performing method is BigClam. The performance of all the other methods deteriorated with the fraction of overlapping nodes increased.

In comparison with other methods, the performance of our method is relatively stable with the change of the average sizes of communities, the fraction of overlapping vertices, and the ratio of the external degree of each node. This provided another piece of supporting evidence to the key intuition behind our method, i.e., a node (particularly that of a high degree) can be important in multi-communities.

We further examined the performance of our model selection method on the four sets of LFR benchmarks. With the same parameters of ε as used before, we compared our method with the spectral method (Krzakala et al. 2013), which has been considered as one of the best methods for determining the number of communities. As shown in Fig. 4, on the first three sets of benchmarks, when the fraction of overlapping communities (o_n/n) is small, the two methods are comparable, accurately finding the actual number of communities; but when o_n/n becomes larger, our method outperforms the spectral method. On the fourth set of benchmarks, however, our method dose not perform well when o_n/n is small (in the range of 0.1 to 0.5); but when o_n/n increases (in the range of 0.6 to 1.0), the new method outperforms again the spectral method. More importantly, for each of these four sets of benchmarks, our method can follow precisely the trend of the number of communities as the fraction of overlapping communities increases (using the same ε-value), whereas the spectral method does not.

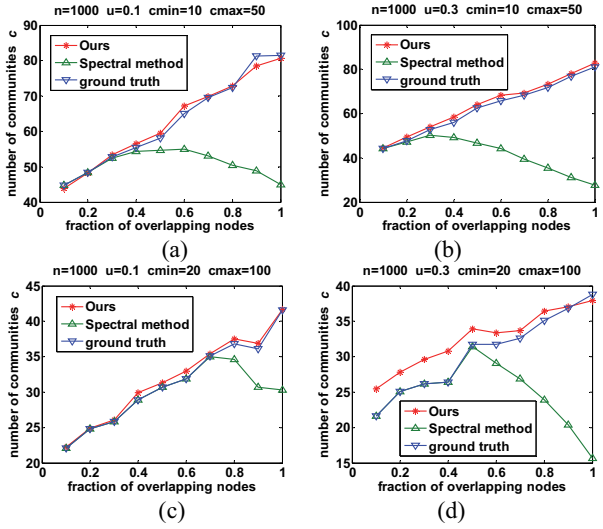


Figure 4: *Comparison of the new method and the spectral method for finding the number of communities of four sets of benchmarks, which are specified in the legend of Figure 3.*

4. Conclusion and Discussion

We proposed a novel method for overlapping community detection. It was built upon a stochastic generative model, learned from a given network, which was used to explore the relative importance of a node in a community to determine whether the node should be included in the community. The method also adopted a hierarchical Bayesian

scheme to derive the number of communities *using* shrinkage priors to adaptively shrink or remove irrelevant communities. The new method was evaluated on both real and synthetic networks with ground-truths, and compared against five state-of-the-art overlapping community detection methods. The results showed the superior performance of the new method over the competing ones.

The superb performance of the new method may be attributed to 1) the stochastic model used to quantify the relative importance of nodes in every community, 2) the node importance was adequately contrasted with its expectation to derive a natural community, and 3) a hierarchical Bayesian scheme for autonomously determining the number of communities. Nevertheless, our method can be improved. An additional parameter ε was introduced to the model selection process to avoid oversuppression of some of the candidate communities. While this parameter is easily determined in experiments, it may still affect finding the right number of communities. We plan to improve the Bayesian model selection scheme in the future.

Acknowledgments

The work was supported by National Basic Research Program of China (2013CB329301, 2012CB316301), Natural Science Foundation of China (61502334, 61303110, 61303109, 61133011, 61373053, 61503281), 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) and United States National Science Foundation (DBI-0743797). Correspondence should be addressed to H.W. (hcwang@tju.edu.cn)

References

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

Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; and Xing, E. P. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9: 1981-2014.

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.

Gopalan, P. K.; and Blei, D. M. 2013. Efficient discovery of overlapping communities in massive networks. *Proc. Natl. Acad. Sci. USA* 110: 14534-14539.

Gregory, S. 2010. Finding overlapping communities in networks by label propagation. *New J. Phys.* 12: 103018.

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*, 130-136. Palo Alto, California, USA: AAAI Press.

Jin, D.; Chen, Z.; He, D.; and Zhang, W. 2015. Modeling with node degree preservation can accurately find communities. *In Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 160-167. Palo Alto, California, USA: AAAI Press.

Jin, D.; Yang, B.; Baquero, C.; Liu, D.; He, D.; and Liu, J. 2011. A Markov random walk under constraint for discovering overlapping communities in complex networks. *J. Stat. Mech.* 2011: P05031.

Krzakala, F.; Moore, C.; Mossel, E.; Neeman, J.; Sly, A.; Zdeborova, L.; and Zhang, P. 2013. Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* 110: 20935-20940.

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

Lancichinetti, A.; and Fortunato, S. 2009. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E* 80: 016118.

Lancichinetti, A.; Radicchi, F.; Ramasco, J. J.; and Fortunato, S. 2011. Finding statistically significant communities in networks. *PLoS ONE* 6: e18961.

Lee, D. D.; and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. *In Proceedings of the 13th Annual Conference on Neural Information Processing Systems*, 556-562. Cambridge, Massachusetts, USA: MIT Press.

Leskovec, J. 2015. Stanford Network Analysis Project. http://snap.stanford.edu

Palla, G.; Derényi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435: 814-818.

Psorakis, I.; Roberts, S.; Ebden, M.; and Sheldon, B. 2011. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review* E 83: 066114.

Raghavan, U. N.; Albert, R.; and Kumara, S. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* 76: 036106.

Shen, H.; Cheng, X.; Cai, K.; and Hu, M. 2009. Detect overlapping and hierarchical community structure in networks. *Physica A* 388: 1706.

Shen, H.; Cheng, X.; and Guo, J. 2011. Exploring the structural regularities in networks. *Phys. Rev. E* 84: 056111.

Tan, V. Y. F.; and Févotte, C. 2013. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35: 1592-1605.

Xie, J.; Kelley, S.; and Szymanski, B. K. 2013. Overlapping community detection in networks: the state of the art and comparative study. *ACM Comput. Surv.* 45: Article No. 43.

Xie, J.; Szymanski, B. K.; and Liu, X. 2011. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *In Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW'11)*, 344-349. Piscataway, NJ, USA: IEEE 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*, 587-596. New York, NY, USA: ACM Press.

Yang, J.; and Leskovec, J. 2014. Structure and overlaps of ground-truth communities in networks. *ACM Transactions on Intelligent Systems and Technology* 5: Article No. 26.

Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. *J Anthropol Res* 33: 452-473.

Zhang, Y.; and Yeung, D. 2012. Overlapping community detection via bounded nonnegative matrix tri-factorization. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 606-614. New York, NY: ACM Press.