# A Stochastic Model for Detecting Heterogeneous Link Communities in Complex Networks

**Dongxiao He[1], Dayou Liu[2*], Di Jin[1], Weixiong Zhang[3,4]**

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China, [2]College of Computer Science and Technology, Jilin University, Changchun 130012, China, [3]Department of Computer Science and Engineering, Washington University, St. Louis, MO 63130, USA, [4]Institute for Systems Biology, Jianghan University, Wuhan, Hubei 430056, China

hedongxiao@tju.edu.cn, liudy@jlu.edu.cn, jindi@tju.edu.cn, weixiong.zhang@wustl.edu

**Abstract**

Discovery of communities in networks is a fundamental data analysis problem. Most of the existing approaches have focused on discovering communities of nodes, while recent studies have shown great advantages and utilities of the knowledge of communities of links. Stochastic models provides a promising class of techniques for the identification of modular structures, but most stochastic models mainly focus on the detection of node communities rather than link communities. We propose a stochastic model, which not only describes the structure of link communities, but also considers the heterogeneous distribution of community sizes, a property which is often ignored by other models. We then learn the model parameters using a method of maximum likelihood based on an expectation-maximization algorithm. To deal with large complex real networks, we extend the method by a strategy of iterative bipartition. The extended method is not only efficient, but is also able to determine the number of communities for a given network. We test our approach on both synthetic benchmarks and real-world networks including an application to a large biological network, and also compare it with two existing methods. The results demonstrate the superior performance of our approach over the competing methods for detecting link communities.

## 1. Introduction

Many complex systems in the real world exist in the form of networks, such as social networks, biological networks, and the Internet, which are collectively referred to as complex networks. One of the main problems in the study of complex networks is the detection of community structures (Girvan and Newman 2002), which has drawn a great deal of interest. Although no common definition has been agreed upon, a community within a network is usually considered as a group of nodes that are densely connected with respect to the rest of the network. In the past few years, many different approaches have been proposed to uncover network community structures, as reviewed in (Fortunato 2010; Xie, Kelley, and Szymanski 2013).

Although previous research on community detection has been mainly focusing on finding communities of nodes, several recent works considered finding communities of links (Ahn, Bagrow, and Lehmann 2010; Evans and Lambiotte 2009; Evans and Lambiotte 2010; Kim and Jeong 2011; Pan et al. 2011; He et al. 2012). In many real networks, link communities are often more informative and intuitive than node communities, because links usually have unique identities, while nodes typically have multiple roles. In a social network, for instance, most individuals belong to multiple communities such as families, friends, and co-workers, while the link between two individuals often exists for a dominant reason which may represent family ties, friendship, or professional relationships. Furthermore, multiple links connecting to a node may belong to distinct link communities, so that the node can be assigned to multiple communities of links. Accordingly, overlapping communities of nodes, another attractive topic in community detection (Palla et al. 2005), could be detected as a natural byproduct of link communities.

Thanks to the good performance and sound theoretical basis, stochastic models constitute a promising technique for identifying modular properties of networks, and thus have been actively researched (Newman 2012). However, most of them focused on the detection of node communities (Wang et al. 2011; Psorakis et al. 2011; Zhang and Yeung 2012; Ren et al. 2009; Shen, Cheng, and Guo 2011; Karrer and Newman 2011; Zhang, Wang, and Ahn 2013), with only one exception for detecting link communities (Ball, Karrer, and Newman 2011). However, like most stochastic models, Ball's model requires the number of communities to be given and is computationally inefficient, so that its applicability is limited, particularly for large-scale networks. More importantly, it fails to describe heterogeneous distributions of community sizes of real networks, providing distorted resulting link communities.



* Corresponding author

Therefore, much needs to be done for stochastic models for link community detection.

In this work, we introduce a stochastic model for link communities, namely LM (link model), which considers the heterogeneity of link community sizes when describing community structures. We learn the parameters of this model using a maximum likelihood method based on an expectation-maximization algorithm. We then extend the above method, by introducing a scheme of iterative bipartition, to link model with iterative bipartition (LMBP). Our new LMBP method can not only autonomously determine the number of communities but is also efficient. Therefore it is applicable to large networks. Our experimental results show that LMBP outperforms two related competing methods on both synthetic and real-world networks for detecting link communities.

## 2. Related Work

A number of approaches to the detection of link communities in networks have been proposed. (Ahn, Bagrow, and Lehmann 2010) used a hierarchical clustering with similarity between links to build a dendrogram to describe hierarchical link structures. In order to obtain the most relevant communities, they introduced a link density function to determine the best level to cut the tree to determine the number of communities. (Evans and Lambiotte 2009; Evans and Lambiotte 2010) transformed a given network into a line graph based on several types of random walk, and detected link communities by applying some existing algorithms for node partitioning to the line graph. (Kim and Jeong 2011) extended the map equation method (Rosvall and Bergstrom 2008), which was originally developed for node communities, to link community detection by assigning the communities to links instead of nodes, modifying the encoding rule for the random walk to represent this change in the community structure, and proposing the corresponding map equation for the link community. (Pan et al. 2011) proposed a local-based method for finding natural link communities through expanding a selected seed to optimize a local function. (He et al. 2012) presented a stochastic process based on a link-node-link random walk to unfold the community structure of links, and then utilized the local mixing properties of the Markov chain to extract emerged link communities.

Moreover, stochastic models provided a promising technique for identifying communities from networks, which has been actively researched (Newman 2012). Several model-based methods have been proposed; they were based on a blockmodel or its variations and employed different inference algorithms, e.g., expectation-maximization and nonnegative matrix factorization, to derive the number of communities. Nevertheless, most of these methods focused on the detection of node communities (Wang et al. 2011; Psorakis et al. 2011; Zhang, and Yeung 2012; Ren et al. 2009; Shen, Cheng, and Guo 2011; Karrer and Newman 2011; Zhang, Wang, and Ahn 2013). One exception for detection of link communities that we are aware of is the algorithm designed by (Ball, Karrer, and Newman 2011). While Ball's model seemed to have a high similarity with the one that we proposed here, there are several key differences. Compared with Ball's model, the most salient feature of ours is its high flexibility. In our model, the size of a link community is modeled by a set of parameters $\omega_z$. This enables it to better describe the heterogeneous community sizes, such as that following a power law distribution which often appears in the real world. This feature, lacking in Ball's model, allows us to better characterize community structures of links of real-world networks. Moreover, our extended LMBP method does not need the number of communities a *priori*, which in contrast is required by the Ball's method BModel. LMBP is also more efficient than BModel. Therefore, compared with BModel, our LMBP method is more suitable for large networks, a necessity for real-world applications.

## 3. The Methods

We first introduce a model for the description of link communities, and then present a method based on maximum-likelihood estimation to learn the model parameters. For clarity, we present an example to illustrate the method. We then extend the basic method to make it more suitable for large real networks.

### 3.1 Stochastic Model

We define a stochastic model of link communities to characterize networks with a given number $n$ of vertices and $m$ undirected edges divided among a given number $c$ of communities. Taking the notion of *soft* membership of links, the model is parameterized by two sets of parameters, $\omega_z$'s and $\theta_{iz}$'s. Here, $\omega_z$ denotes twice the expected number of links in community $z$, which is defined as the sum of all expected counts of $z$-links (links in community $z$) that a node connects to, and $\theta_{iz}$ denotes the probability that community $z$ selects node $i$ when generating edges, which is defined as the expected proportion of $z$-links node $i$ connects to in this community. Thus, we have $\sum_z \omega_z = 2m$ and $\sum_i \theta_{iz} = 1$.

Based on the model above, an edge $<i, j>$ can be generated as follows. A link community $z$ is chosen with size $\omega_z$, and within community $z$, nodes $i$ and $j$ are selected with probabilities $\theta_{iz}$ and $\theta_{jz}$, respectively, to form an edge. Consequently, the expected number of links between nodes $i$ and $j$ in community $z$ is

$$\hat{A}_{ij}^{z} = \omega_z \theta_{iz} \theta_{jz} . \quad (1)$$

Summing over communities $z$, the expected number of link between $i$ and $j$ can be written as

$$\hat{A}_{ij} = \sum_z \hat{A}_{ij}^{z} = \sum_z \omega_z \theta_{iz} \theta_{jz} . \quad (2)$$

Under this model, link communities will appear with the generating of networks. Intuitively, two nodes $i$ and $j$ which have large values of $\omega_z$, $\theta_{iz}$ and $\theta_{jz}$ for some value of $z$ have a high probability of being connected by a link within community $z$. Thus, groups of such nodes will tend

to be connected by relatively dense webs of $z$-links, and these sets of edges correctly form the link communities we expect to see.

Formally, assume that the community assignments are represented by a set of variables $R_{ij}^z$'s, where $R_{ij}^z$ denotes the fraction that a link $<i, j>$ belongs to community $z$. Then we have

$$R_{ij}^z = \frac{\hat{A}_{ij}^z}{\hat{A}_{ij}} = \frac{\omega_z \theta_{iz} \theta_{jz}}{\sum_s \omega_s \theta_{is} \theta_{js}}, \quad (3)$$

As the soft membership of communities cannot be used directly, we simply assign each link $<i, j>$ to community $r$ satisfying $r = arg\max_z \{R_{ij}^z, z=1,2,\ldots,c\}$, and then derive a hard partition of links.

## 3.2 Parameter Learning

Since the (stochastic) model parameters for a given network are unknown, we need to learn the parameters in order to infer the link communities in the network. This can be done by maximizing a likelihood function that the network was presumably generated from the model. Since the number of links between two nodes is given in the expectation of a Poisson distribution (Karrer and Newman 2011; Ball, Karrer, and Newman 2011), the probability of generating a graph $G$ with adjacency matrix $(A_{ij})_{n\times n}$ by the model specified in (2) is

$$P(G \mid \omega, \theta) = \prod_{i,j} \frac{\left(\sum_z \omega_z \theta_{iz} \theta_{jz}\right)^{A_{ij}}}{A_{ij}!} \exp\left(-\sum_z \omega_z \theta_{iz} \theta_{jz}\right). \quad (4)$$

The best fit between the given network $G$ and its expected network in (2) can be achieved by maximizing the likelihood function in (4).

Likelihood maximization does not typically work directly with the likelihood itself, but with its logarithm. Taking the log of (4), rearranging, and dropping additive and multiplicative constants, we derive the log-likelihood

$$L = \sum_{ij} A_{ij} \log\left(\sum_z \omega_z \theta_{iz} \theta_{jz}\right) - \sum_{ijz} \omega_z \theta_{iz} \theta_{jz}. \quad (5)$$

Direct maximization of this expression by differentiating leads to a set of nonlinear implicit equations for $\omega_z$ and $\theta_{iz}$ that seem to be difficult to solve. Here we adopt an expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). We apply Jensen's inequality to (5), and obtain

$$L \geq \bar{L} = \sum_{ijz} \left( A_{ij} q_{ij,z} \log \frac{\omega_z \theta_{iz} \theta_{jz}}{q_{ij,z}} - \omega_z \theta_{iz} \theta_{jz} \right), \quad (6)$$

where the probabilities $q_{ij,z}$ can be freely chosen, provided they satisfy $\sum_z q_{ij,z} = 1$. Especially, the exact equality can always be achieved by making a particular choice $q_{ij,z} = \omega_z \theta_{iz} \theta_{jz} / \sum_s \omega_s \theta_{is} \theta_{js}$. Thus, it follows that the double maximization of the new function $\bar{L}$ with respect to both the model parameters ($\omega_z$ and $\theta_{iz}$) and the probabilities ($q_{ij,z}$) is equivalent to maximizing the original log-likelihood $L$ with respect to the model parameters alone. Given the optimal model parameters $\omega_z$ and $\theta_{iz}$, the optimal values of probabilities $q_{ij,z}$ are given by

$$q_{ij,z} = \frac{\omega_z \theta_{iz} \theta_{jz}}{\sum_s \omega_s \theta_{is} \theta_{js}}, \quad (7)$$

since these are the values that give the inequality in (6) an exact equality. Meanwhile, given the optimal probabilities $q_{ij,z}$, the optimal values of model parameters $\omega_z$ and $\theta_{iz}$ can be found by maximizing $\bar{L}$ with the constraints $\sum_z \omega_z = 2m$ and $\sum_i \theta_{iz} = 1$. Introducing Lagrange multipliers $\rho$ and $\gamma_z$ to incorporate these constraints, the Lagrange form of $\bar{L}$ is

$$\tilde{L} = \bar{L} + \rho\left(\sum_z \omega_z - 2m\right) + \sum_z \gamma_z \left(\sum_i \theta_{iz} - 1\right). \quad (8)$$

By differentiating (8), the optimal values of $\omega_z$ and $\theta_{iz}$ are given as

$$\omega_z = \sum_{ij} A_{ij} q_{ij,z}; \quad \theta_{iz} = \frac{\sum_j A_{ij} q_{ij,z}}{\sum_{kj} A_{kj} q_{kj,z}}. \quad (9)$$

Maximizing the log-likelihood $L$ is now equivalent to simultaneously solving (7) and (9), which can be done iteratively by choosing a random set of initial values and alternating back and forth between the two equations. The EM algorithm implemented here is guaranteed to converge under the above conditions.

Notice that the $q_{ij,z}$ are only defined for node pairs $i$, $j$ that are actually connected in the network (so that $A_{ij} = 1$), and hence there are only as many of them as there are observed edges. Thus the time to evaluate (7) once is $O(mc)$, where $m$ is the number of edges and $c$ the number of communities. Similarly, the time for calculating (9) once is $O(mc)$ as we only need to consider the observed edges. Therefore, the time complexity of our method is $O(Tmc)$, where $T$ is the number of iterations to convergence.

## 3.3 An Illustrative Example

We now illustrate the idea of our method using a simple example shown in Figure 1 and Table 1.

The given network $G$ is in Figure 1(a). Under our model, given the parameters $\omega_z$'s and $\theta_{iz}$'s such as that in Table 1, we can form the expected graphs of all the link communities in $G$ according to (1), which are shown in Figure 1(b) and (c). Further, we can form the expected graph of the whole network $G$ according to (2), which is an ensemble of the expected graphs of all its communities, shown in Figure 1(d). However, the model parameters are unknown, they must first be learned in order to find the communities in the network. To this end, we consider network $G$ and its expected graph by optimizing (4), and then get the best $\omega_z$'s and $\theta_{iz}$'s, as shown as Table 1. Thereafter, we infer the community structure of links according to (3), which perfectly matches the ground-truth given in Figure 1(a).

**Table 1**: *The learned model parameters $\omega_z$'s and $\theta_{iz}$'s*

| | $\omega_z$ | $\theta_{iz}$ | | | |
|---|---|---|---|---|---|
| | | $i$=1 | $i$=2 | $i$=3 | $i$=4 |
| $z$=1 | 19.99992 | 0.2 | 0.2 | 0.2 | 0.2 |
| $z$=2 | 20.00008 | 5.20E-07 | 5.36E-07 | 5.19E-07 | 5.04E-07 |
| | | $\theta_{iz}$ | | | |
| | | $i$=5 | $i$=6 | $i$=7 | $i$=8 | $i$=9 |
| $z$=1 | 0.199999 | 1.39E-09 | 1.36E-09 | 1.43E-09 | 8.37E-10 |
| $z$=2 | 0.200001 | 0.199999 | 0.199999 | 0.199999 | 0.199999 |

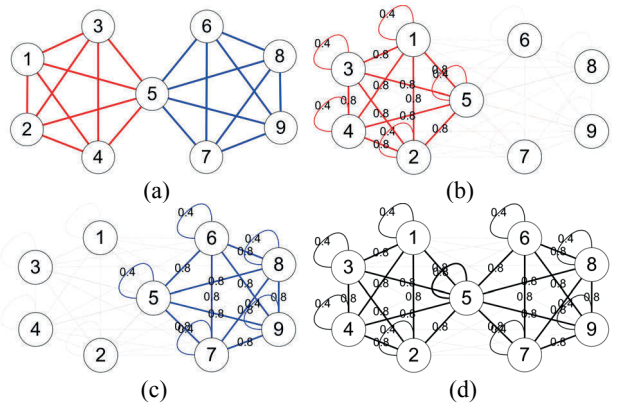

**Figure 1:** *An illustration of our method for identifying the community structure of links.* ***(a)*** *The given network G with two link communities (in red and blue).* ***(b)*** *and* ***(c)*** *The expected graph of the red and blue link community, respectively.* ***(d)*** *The expected graph of G, which is an ensemble of the expected graphs of its red and blue communities. The width of a link corresponds to its expected values, and values smaller than 1.0e-3 are omitted.*

### 3.4 A Practical Extension

The method discussed above can be improved. A drawback of our basic method is that it offers no criterion for determining the value of parameter $c$, i.e., the number of communities in a network. This is also a common problem suffered by all exiting methods based on stochastic models. A statistical model selection method can be, in principle, applied to stochastic models to find the number of communities (Brunet et al. 2004; Tan and Févotte 2012). Such a model selection method is too computationally demanding to be desirable to any but some small networks (Ball, Karrer, and Newman 2011). Even if the number $c$ of communities is given, because large networks often have large values of $c$, the convergence rate of the core learning algorithms, such as expectation maximization and nonnegative matrix factorization, will become very slow. This is also a crucial limitation for the existing model-based methods when dealing with large real networks. It remains an open problem whether an accurate and efficient model selection method can be developed for large real networks (Karrer and Newman 2011).

To mitigate the above problems, we extend our original method LM (link model) to "link model with iterative bipartition", or LMBP for short. In LMBP, we first divide a network into two link modules using LM with community number $c = 2$, and then recursively subdivide the two parts. In dividing a subnetwork, we isolate it from the rest of the network and perform a 'nested' LMBP on it, resulting in a partition of the subnetwork with two smaller link communities. For each partition, we decide whether to accept a bipartition based on the quality of the resulting link partition. We summarize the algorithm LMBP using the following recursive algorithm:

**Algorithm** $P$ = LMBP($G$) // $G$ is a graph, $P$ is a link partition of $G$

1. $P = \{E(G)\}$; // $E(G)$ denotes the edge set of $G$
2. Divide $G$ into two link modules $N_1$ and $N_2$ by LM; // $E(N_1) \cap E(N_2) = \Phi$, $E(N_1) \cup E(N_2) = E(N)$
3. If the link partition quality cannot be improved by this bipartition, return $P$; // the quality function is to be introduced later
4. $P_1$ = LMBP($N_1$);
5. $P_2$ = LMBP($N_2$);
6. Return $P = P_1 \cup P_2$.

We now consider the termination condition for the repetitive process of subdividing the links of network $G$, so as to obtain a superior link community structure. Several measures for community structures exist, most of which were developed for node communities (Fortunato 2010; Newman and Girvan 2004; Lancichinetti et al. 2011). Fortunately, partition density $D$ (Ahn, Bagrow, and Lehmann 2010) was specially designed for link communities. Here we adopt it as our quality metric, i.e., we iteratively bipartition each (sub)network until the density $D$ cannot be further improved to determine the acceptance of the bipartition.

For a network with $m$ links and $n$ nodes, $P = \{P_1, P_2, \ldots, P_c\}$ is a partition of the links into $c$ communities. The number of links in community $z$, $P_z$, is $m_z = |P_z|$. The number of induced nodes, the nodes that those links connect to, is $n_z = |\cup_{eij \in Pz,} \{i, j\}|$. The link density $D_z$ of $P_z$ is

$$D_z = \frac{m_z - (n_z - 1)}{n_z(n_z - 1)/2 - (n_z - 1)}. \quad (10)$$

This is $m_z$ normalized by the minimum and maximum numbers of links among $n_z$ connected nodes. Thus, $D_z = 1$ when $P_z$ is a clique, or $D_z = 0$ when $P_z$ is a tree. In particular, we assume that $D_z = 0$ if $n_z = 2$ without loss of generality. In essence, $D_z$ measures how 'clique-ish' versus 'tree-ish' that $P_z$ is. Then, the partition density, $D$, is the average of $D_z$, weighted by the fraction of links that are present:

$$D = \frac{2}{m}\sum_z m_z \frac{(m_z - n_z + 1)}{(n_z - 2)(n_z - 1)}. \quad (11)$$

## 4. Experiments

In order to evaluate the performance of our method LMBP, we tested it on synthetic networks and widely used real-world networks. The synthetic networks allow us to test LMBP's ability to detect known communities, while the real networks allow us to assess its performance in practice. As an application, we applied LMBP to a large biological network.

In our analysis, we compared LMBP with two well-known and closely related methods. The first (denoted as BModel) is a model-based method for link communities proposed by (Ball, Karrer, and Newman 2011), and the second (denoted as LC) is the notable method of link communities proposed by (Ahn, Bagrow, and Lehmann 2010). To the best of our knowledge, LMBP and BModel are the only two methods based on stochastic models for link communities, and LMBP and LC are the only two

hierarchical methods using partition density $D$ (Ahn, Bagrow, and Lehmann 2010) as a quality metric to detect link communities. Note that BModel needs the number $c$ of communities as a given parameter, thus we used the community number obtained by our LMBP as its input. Besides, the LMBP and BModel converge to local minima, thus we ran each of them 20 times and reported the best results.

## 4.1 Synthetic Networks

Several benchmarks of synthetic network have been proposed for node communities (Girvan and Newman 2002; Lancichinetti, Fortunato, and Radicchi 2008; Lancichinetti and Fortunato 2009). In contrast, only one benchmark, to our knowledge, has been designed for testing algorithms for link community detection (Ball, Karrer, and Newman 2011), which we used in this evaluation. We also employed two accuracy measures introduced in (Ball, Karrer, and Newman 2011), namely "Fraction of Vertices Classified Correctly (FVCC)" and "Jaccard index", to compare the planted community structures of a network and the ones delivered by the algorithms compared. Notice that LC does not appear here, because it often finds very small communities, and fails to detect the communities defined in this benchmark.

Following (Ball, Karrer, and Newman 2011), the parameter setting for this benchmark is given as follows. The networks have $n$ = 10000 nodes each, divided into two overlapping (link) communities. We placed $x$ nodes in the first community only, i.e., these nodes have connections exclusively within the community, $y$ nodes in the second community only, and the remaining $z=n-x-y$ nodes in both communities, with equal numbers of connections to nodes in these two communities on average. We set the expected degree of all nodes to a fixed value $\langle k \rangle$. We also varied the parameters $x$, $y$, $z$, and $\langle k \rangle$ to generate networks with stark community structures or no structure at all, so as to vary the difficulty of the network instances posed to the algorithms.

We performed three sets of tests. In the first set of experiments, we fixed the size of the overlap between the communities at $z$ = 500, divided the remaining nodes evenly (i.e., $x = y$ = 4750), and varied the value of $\langle k \rangle$ from 1 to 15 with an increment of 1. For the second set of tests, we again set the overlap at $z$ = 500 but fixed $\langle k \rangle$ = 10 and varied the ratio between $x$ and $y$. Finally, for the third set of tests, we set $\langle k \rangle$ = 10, constrained $x$ and $y$ to be equal, and varied the amount of overlap $z$.

As BModel requires the number of communities to be given, we set the number of communities for BModel to 2, the actual number of communities. For fairness, the first bipartition result from LMBP was used for comparison with BModel. In Figure 2, we show the fraction of corrected classified nodes by the two algorithms for each of the three sets of experiments. To be considered correctly classified, a node's membership in both communities must be reported correctly by an algorithm. As shown in Figure 2, LMBP outperforms BModel in terms of FVCC accuracy in all the three tests. This may be mainly due to our parameter $\omega_z$, which controls the size of each link community, and thus makes our model more flexible to describe link communities compared with Ball's model.

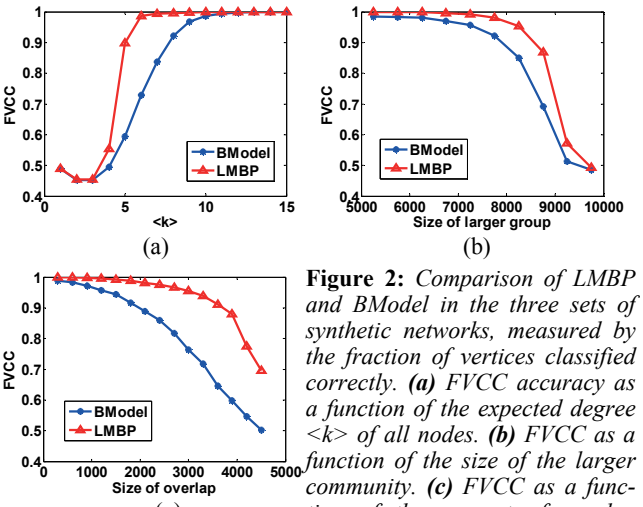


**Figure 2:** *Comparison of LMBP and BModel in the three sets of synthetic networks, measured by the fraction of vertices classified correctly.* ***(a)*** *FVCC accuracy as a function of the expected degree $\langle k \rangle$ of all nodes.* ***(b)*** *FVCC as a function of the size of the larger community.* ***(c)*** *FVCC as a function of the amount of overlap between the two communities.*

Furthermore, we adopted the Jaccard index to compare the two algorithms' ability for identifying overlapping (link) communities using the same sets of network instances. Let $S$ be the set of truly overlapping nodes and $V$ be the set of predicted overlapping nodes, the Jaccard index is $J = |S \cap V| / |S \cup V|$. This index is a standard measure of similarity between sets that rewards accurate identification of the overlap while penalizes both false positives and false negatives. Figure 3 shows the result comparing the two algorithms. As shown, LMBP is also superior to BModel in all the three sets of experiments. This result is similar to the results in Figure 2, and they both confirm the validity of LMBP.

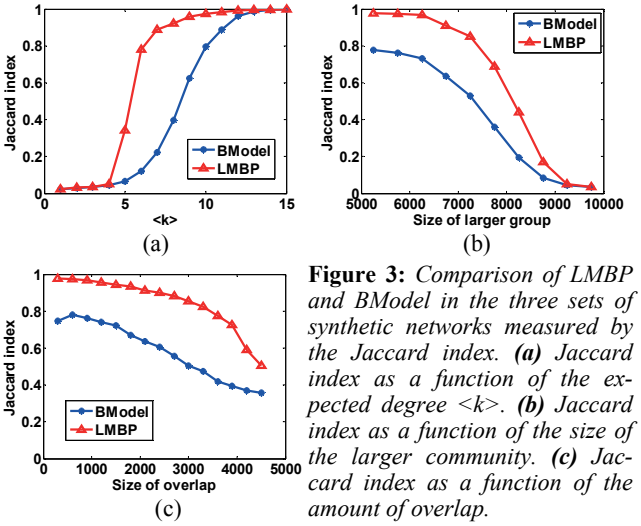


**Figure 3:** *Comparison of LMBP and BModel in the three sets of synthetic networks measured by the Jaccard index.* ***(a)*** *Jaccard index as a function of the expected degree $\langle k \rangle$.* ***(b)*** *Jaccard index as a function of the size of the larger community.* ***(c)*** *Jaccard index as a function of the amount of overlap.*

## 4.2 Real Networks

As real-world networks may have some unique topological properties not present in synthetic ones, we considered some widely used real networks to compare these algo-

rithms. All the networks we used here were obtained from Newman's website (Newman, 2013), except that 'protein-protein interaction' and 'word association' that were from (Palla et al. 2005). We adopted the partition density $D$ (Ahn, Bagrow, and Lehmann 2010), discussed earlier, as the quality metric for comparison.

Table 2 shows the result comparing our method LMBP with BModel and LC on these real-world networks. As shown, LMBP has the best performance on 13 of the 16 networks in terms of partition density $D$, BModel performs the best on two networks, and LC performs the best on one network. This result shows the superior performance of our method over the others methods on real-world networks.

**Table 2:** *Comparison of three methods for detecting link communities on real networks. Here, the greater a D-value, the better. In the table, '−' denotes run time >48 hours or the program ran out of memory.*

| Datasets | $n$ | $m$ | $c$ | partition density $D$ | | |
|---|---|---|---|---|---|---|
| | | | (Ours) | LMBP | BModel | LC |
| Zachary's karate club | 34 | 78 | 19 | **0.5405** | 0.4496 | 0.2847 |
| Dolphin social network | 62 | 160 | 29 | **0.3308** | 0.3199 | 0.3155 |
| High school friendship | 69 | 220 | 41 | **0.4932** | 0.4576 | 0.3600 |
| Les Miserables | 77 | 254 | 27 | **0.6772** | 0.5518 | 0.5765 |
| Political books | 105 | 441 | 90 | **0.5151** | 0.4958 | 0.2866 |
| Word adjacencies | 112 | 425 | 82 | **0.2863** | 0.2701 | 0.0632 |
| American college football | 115 | 613 | 98 | 0.5432 | **0.5508** | 0.5500 |
| Jazz musicians collaborations | 198 | 2,742 | 181 | **0.6234** | 0.6033 | 0.4155 |
| C. Elegans neural | 297 | 2,148 | 308 | **0.4067** | 0.3553 | 0.0823 |
| E. coli metabolic | 453 | 2,025 | 412 | 0.5626 | **0.5983** | 0.3333 |
| E-mail network URV | 1,133 | 5,451 | 910 | **0.3846** | 0.3186 | 0.1018 |
| Political blogs | 1,490 | 16,717 | 921 | **0.2690** | 0.1971 | 0.1204 |
| Network science collaborations | 1,589 | 2,742 | 518 | **0.8207** | 0.7517 | 0.6937 |
| Power grid | 4,941 | 6,594 | 58 | 0.0344 | -3.1e-4 | **0.1370** |
| Protein-protein interaction | 2,640 | 6,600 | 917 | **0.2740** | 0.2217 | 0.1705 |
| Word association | 5,017 | 29,148 | 5,335 | **0.2687** | − | 0.0767 |

### 4.3 Application

The large real network we considered as an application is the protein-protein interaction (PPI) network of budding yeast *Saccharomyces cerevisiae* (Palla et al. 2005; Xenarios et al. 2000). It contains 2,640 nodes (proteins) and 6,600 links (physical interactions between pairs of proteins).

We used the Gene Ontology (GO) terms (Ashburner et al. 2000), the most elaborate gene function annotations, as domain metadata for quality assessment. The GO terms include information on functions and cellular locations of a gene and biological pathways that a gene may be involved in. The biological significance of a community of genes (nodes) can be measured by the GO terms enriched in the genes in the community. Enrichment of GO terms can be evaluated by a hyper-geometric test (Altman 1991), providing a GO term a $p$-value to quantify the significance of the term. To quantify the biological significance of a community structure, we used as quality metric the average number of significantly enriched GO terms with $p$-values less than a given threshold for all communities. The larger this average number of significant GO terms, the more biologically significant the community structure is.

As shown in Figure 4, our method LMBP identified PPI community structures with many more significant GO terms than the LC method and with slightly more significant GO terms than the BModel method under all 10 different $p$-value thresholds tested. It serves as an additional example of the consistent superior performance of our method over the competing methods compared.

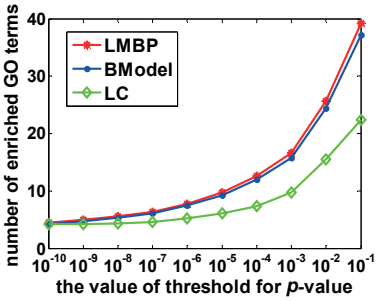


**Figure 4:** *Comparison of LMBP with BModel and LC on budding yeast PPI network.*

## 5. Conclusions

We proposed a stochastic model, namely Link-Model or LM, to not only describe the structure of link communities but also deal with heterogeneous sizes of community structures. In our method, we learned the model parameters by a combination of likelihood optimization and expectation-maximization. We extended the basic method by an iterative bipartition to autonomously determine the number of communities. The new method, named as LMBP, is more suitable for large, real networks. We tested LMBP and compared it with two existing competing methods on synthetic benchmark problems and real-world networks including a large biological network. Experimental results demonstrated the superior performance of our method over the competing methods for the detection of link communities in large networks.

There are other quality metrics for link communities (e.g., the extended map equation (Kim and Jeong 2011)) which may be also suitable for our iterative bipartition procedure. We will include in our software an option for choosing different quality metrics to make our method more applicable to various problems. Besides, we will use our method to analyze multimedia and social networks, and try to unfold significant community structures in real life.

## Acknowledgments

The work was supported in part by National Basic Research Program (973 Program) of China (2013CB329301), Natural Science Foundation of China (61303110, 61133011, 61373035 and 31300999), National High Technology Research and Development Program (863 Program) of China (2013AA013204), the municipal government of Wuhan, Hubei, China (2014070504020241 and the Talent Development Program), and an internal research grant of Jianghan University, Wuhan, China, as well as by United States National Institutes of Health (R01GM100364).

## References


Ahn, Y. Y.; Bagrow, J. P.; and Lehmann, S. 2010. Link Communities Reveal Multiscale Complexity in Networks. *Nature* 466(7307): 761-764.

Altman, D. G. eds. 1991. *Practical Statistics for Medical Research*. London, UK: Chapman& Hall/CRC.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* 25(1): 25-29.

Ball, B.; Karrer, B.; and Newman, M. E. J. 2011. Efficient and Principled Method for Detecting Communities in Networks. *Physical Review* E 84(3): 036103.

Brunet, J. P.; Tamayo, P.; Golun, T. R.; and Mesirov, J. P. 2004. Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *Proceedings of National Academy of Sciences of the United States of America* 101(12): 4164-4169.

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): 1-38.

Evans, T. S.; and Lambiotte, R. 2009. Line Graphs, Link Partitions, and Overlapping Communities. *Physical Review* E 80(1): 016105.

Evans, T. S.; and Lambiotte, R. 2010. Line Graphs of Weighted Networks for Overlapping Communitie. *The European Physical Journal* B 77(2): 265-272.

Fortunato, S. 2010. Community Detection in Graphs. *Physics Reports* 486(3-5): 75-174.

Girvan, M.; and Newman M. E. J. 2002. Community Structure in Social and Biological Networks. *Proceedings of National Academy of Sciences of the United States of America* 99(12): 7821-7826.

He, D.; Liu, D.; Zhang, W.; Jin, D.; and Yang, B. 2012. Discovering Link Communities in Complex Networks by Exploiting Link Dynamics. *Journal Statistical Mechanics: Theory and Experiment* 2012: P10015.

Karrer, B.; and Newman, M. E. J. 2011. Stochastic Blockmodels and Community Structure in Networks. *Physical Review* E 83(1): 016107.

Kim, Y.; and Jeong, H. 2011. Map Equation for Link Communities. *Physical Review* E 84(2): 026110.

Lancichinetti, A.; Fortunato, S.; and Radicchi, F. 2008. Benchmark Graphs for Testing Community Detection Algorithms. *Physical Review* E 78(4): 046110.

Lancichinetti, A.; and Fortunato, S. 2009. Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities. *Physical Review* E 80(1): 016118.

Lancichinetti, A.; Radicchi, F.; Ramasco, J. J.; and Fortunato, S. 2011. Finding Statistically Significant Communities in Networks. *PLoS ONE* 6(4): e18961.

Newman M. E. J.; and Girvan M. 2004. Finding and Evaluating Community Structure in Networks. *Physical Review* E 69(2): 026113.

Newman, M. E. J. 2012. Communities, Modules and Large-scale Structure in Networks. *Nature Physics* 8: 25-31.

Newman, M. E. J. 2013. Real-world Network Data in Newman's Homepage. http://www-personal.umich.edu/~mejn/netdata/

Palla, G.; Derenyi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the Overlapping Community Structures of Complex Networks in Nature and Society. *Nature* 435(9): 814-818.

Pan, L.; Wang, C.; Xie, J.; and Liu, M. 2011. Detecting Link Communities Based on Local Approach. In *Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence,* 884-886. Piscataway, NJ, USA: IEEE Press.

Psorakis, I.; Roberts, S.; Ebden, M.; and Sheldon, B. 2011. Overlapping Community Detection Using Bayesian Non-negative Matrix Factorization. *Physical Review* E 83(6): 066114.

Ren, W.; Yan, G.; Liao, X.; and Xiao, L. 2009. Simple Probabilistic Algorithm for Detecting Community Structure. *Physical Review* E 79(3): 036111.

Rosvall, M.; and Bergstrom, C. T. 2008. Maps of Random Walks on Complex Networks Reveal Community Structure. *Proceedings of National Academy of Sciences of the United States of America* 105(4): 1118-1123.

Shen, H.; Cheng, X.; and Guo, J. 2011. Exploring the Structural Regularities in Networks. *Physical Review* E 84(5): 056111.

Tan, V. Y. F.; and Févotte, C. 2013. Automatic Relevance Determination in Nonnegative Matrix Factorization with the $\beta$-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(7): 1592-1605.

Wang, F.; Li, T.; Wang, X.; Zhu, S.; and Ding, C. H. Q. 2011. Community Discovery Using Nonnegative Matrix Factorization. *Data Mining and Knowledge Discovery* 22(3): 493-521.

Xenarios, I.; Rice, D. W.; Salwinski L.; Baron M. K.; Marcotte E. M.; Eisenberg D. 2000. DIP: the Database of Interacting Proteins. *Nucleic Acids Research* 28(1): 289-291.

Xie, J.; Kelley S.; and Szymanski B. K. 2013. Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study. *ACM Computing Survey* 45(4): Article No. 43.

Zhang, Z.; Wang, Y.; and Ahn, Y. Y. 2013. Overlapping Community Detection in Complex Networks Using Symmetric Binary Matrix Factorization. *Physical Review* E 87(6): 062803.

Zhang, Y.; and Yeung, D.-Y. 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, USA: ACM Press.