

See discussions, stats, and author profiles for this publication at: <http://www.researchgate.net/publication/270454471>

Quantitative Function and Algorithm for Community Detection in Bipartite Networks

ARTICLE · JANUARY 2015

Source: arXiv

READS

39

4 AUTHORS, INCLUDING:



[Zhenping Li](#)

Beijing Wuzi University

18 PUBLICATIONS 212 CITATIONS

SEE PROFILE



[Shihua Zhang](#)

Chinese Academy of Sciences

54 PUBLICATIONS 1,108 CITATIONS

SEE PROFILE



[Xiang Zhang](#)

Fourth Military Medical University

754 PUBLICATIONS 21,606 CITATIONS

SEE PROFILE

Quantitative Function and Algorithm for Community Detection in Bipartite Networks

Zhenping Li¹, Rui-Sheng Wang², Shihua Zhang^{3*}, Xiang-Sun Zhang^{3*}

1. School of Information, Beijing Wuzi University, Beijing, China;

2. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA;

3. National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, CAS, Beijing, China;

* Corresponding author. E-mail: zsh@amss.ac.cn, zxs@amt.ac.cn

Abstract

Community detection in complex networks is a topic of high interest in many fields. Bipartite networks are a special type of complex networks in which nodes are decomposed into two disjoint sets, and only nodes between the two sets can be connected. Bipartite networks represent diverse interaction patterns in many real-world systems, such as predator-prey networks, plant-pollinator networks, and drug-target networks. While community detection in unipartite networks has been extensively studied in the past decade, identification of modules or communities in bipartite networks is still in its early stage. Several quantitative functions proposed for evaluating the quality of bipartite network divisions are based on null models and have distinct resolution limits. In this paper, we propose a new quantitative function for community detection in bipartite networks, and demonstrate that this quantitative function is superior to the widely used Barber's bipartite modularity and other functions. Based on the new quantitative function, the bipartite network community detection problem is formulated into an integer programming model. Bipartite networks can be partitioned into reasonable overlapping communities by maximizing the quantitative function. We further develop a heuristic and adapted label propagation algorithm (BiLPA) to optimize the quantitative function in large-scale bipartite networks. BiLPA does not require any prior knowledge about the number of communities in the networks. We apply BiLPA to both artificial networks and real-world networks and demonstrate that this method can successfully identify the community structures of bipartite networks.

Keywords: bipartite network; community; quantitative function; integer programming; label propagation algorithm

1 Introduction

Many real-world systems, such as the Internet, social relationships, food webs, and biological systems, can be represented as complex networks [1–4]. Community and its detection is an important and valuable topic of great interest in network science and many other disciplines from sociology, computer science to biology [4–7]. A community in a complex network is often defined as a set of nodes which are densely connected with each other, but sparsely connected with nodes outside of the set.

To evaluate the quality of a network partition or community structure, a proper quantitative function is needed. Modularity Q introduced by Newman and Girvan [8] is the most widely used function. Q measures the density of edges within communities as compared to a null model. In other words, a partition with high modularity should be the one that the density of edges in modules is significantly higher than random expectation. Modularity optimization has become a very popular method for community detection in the past decade. However, Q has been exposed to have serious resolution limit issues [9, 10]. It contains an intrinsic scale that depends on the total number of edges in the network which makes it fail to detect dense communities smaller than this scale [9]. To overcome the resolution limit, several

other quantitative functions have been proposed [11–14]. For example, Li et al. [11] proposed community density D based on the concept of average degree of a community and Ann et al. [13] defined a quantitative function based on the average link density which was improved recently by Li et al. [14].

Although most attention has focused on the community detection in unipartite networks (see [5] and references therein), many real-world relations, such as plant-pollinator, order-item, paper-author, and event-attendee, are more suitable to be represented as bipartite networks [15]. In bipartite networks, nodes are divided into two disjoint sets (i.e., a bipartite network is composed of two types of nodes), and only two nodes from different sets can be connected. In the past several years, community detection in bipartite networks has attracted great interests as well [15–23].

Several quantitative functions and algorithms for community detection in bipartite networks have been developed [15, 23–25]. Freeman [25] proposed a projection-based algorithm for bipartite community detection, where a bipartite network is projected to a unipartite network. In the projected network, two nodes are connected if they are adjacent to one or more common nodes in the bipartite network. Obviously, some information of the original bipartite network is lost in the projected network, even if the projection is weighted.

In Guimerà’s opinion, a bipartite network community is composed of nodes of the same type [15]. In order to measure the connectivity between nodes of different types, a bipartite modularity function considering both node types has to be used. Barber proposed a bipartite modularity function [24] based on the assumption that a community is a bipartite subgraph composed of nodes of both types. Murata and Ikeya [23] proposed another bipartite modularity that uses a pair of two values to represent bipartite modularity in both directions, which allows one-to-many correspondence of communities of different node types. This bipartite modularity can reveal results consistent with those obtained by Newman’s modularity when it is applied to unipartite networks. Suzuki and Wakita [26] modified the version of Murata’s bipartite modularity so as to reflect the multi-facet correspondence among communities. In addition, several studies have designed algorithms to maximize these modularity functions for community detection in bipartite networks [18–20, 23, 27].

So far there has been no agreement on a standard definition for community detection in bipartite networks. Since the above definitions of bipartite modularity are all based on the null model used in Newman-Girvan modularity, they all have the resolution limit issue [9, 23] and fail to detect communities smaller than a detectable scale that depends on the size of a network and the interconnectivity of its communities. When the scale of the network is sufficiently large, optimizing bipartite modularity favors network divisions with groups of small communities merged into larger communities, which may lead to ambiguities [19].

In this paper, we propose a novel quantitative function – bipartite partition density – based on the concept of average link density for evaluating a community partition in (both unweighted and weighted) bipartite networks. By maximizing the bipartite partition density, we can partition a bipartite network into communities. We formulate the community partition problem for a bipartite network into an integer programming model. In addition to the simple form, we show that the proposed criterion overcomes the resolution limit for community detection in bipartite networks based on theoretical analysis and numerical tests on artificial and real-world networks. Then we design a revised label propagation algorithm (BiLPA) for efficient community detection in large-scale bipartite networks. We demonstrate their effectiveness with applications onto artificial and real-world bipartite networks.

2 Partition Density of Bipartite Networks

A bipartite network $G = (U, V, E)$ is a graph with two disjoint types of nodes U and V , $U \cap V = \emptyset$, such that there is no edge connecting two nodes of the same type. Let $L(U, V)$ be the number of edges with one node in U and the other node in V , then $L(U, V) = |E|$ is the total number of edges in bipartite network G .

A complete bipartite network or a biclique $G = (U, V, E)$ is a bipartite network, with each node in U adjacent to each node in V . In other words, there are $L(U, V) = |U| \times |V|$ edges in the complete bipartite

network. Let $B(m, n)$ denote a biclique with m nodes in U and n nodes in V . $B(m, n)$ has the maximal number of edges among all bipartite networks with the same node sets U and V .

Definition 1. Bipartite graph density The density $D(G)$ of a bipartite network $G = (U, V, E)$ is defined as the ratio of the number of edges in G to the number of edges of a corresponding biclique $B(|U|, |V|)$, which can be calculated by $D(G) = \frac{L(U, V)}{|U| \times |V|}$. Obviously, for any bipartite network without parallel edges, the maximal graph density value is 1 when it is a biclique, and the minimal graph density value is 0 when it is an empty bipartite network.

Definition 2. Community partition density Given a bipartite network $G = (U, V, E)$, we assume that $G = \{G_1, \dots, G_K\}$ is a partition of G into K subgraphs, $G_c = (U_c, V_c, E_c)$, $c = 1, 2, \dots, K$, then $\cup_{c=1}^K U_c = U$, $\cup_{c=1}^K V_c = V$, and $E_c = \{e_{ij} \in E | u_i \in U_c, v_j \in V_c\}$. Let $L(U_c, V_c) = |E_c|$ be the number of edges with one node in U_c and the other node in V_c , then the density of subgraph G_c is defined as

$$D_c = \frac{L(U_c, V_c)}{|U_c| \times |V_c|}.$$

The partition density D of a bipartite network is defined as the weighted average of D_c , where the weight coefficient is the ratio of the number of edges in each community relative to the total number of edges in bipartite network G :

$$D = \sum_c \frac{L(U_c, V_c)}{L(U, V)} \cdot D_c = \frac{1}{L(U, V)} \sum_c \frac{L(U_c, V_c)^2}{|U_c| \times |V_c|}. \quad (1)$$

We can see that the maximum value of partition density D is 1 when each community is a biclique and there is no edge between the communities, and the minimum value of D is 0 when each community is an empty bipartite graph. If a bipartite network is partitioned into overlapping communities, we can also use formula (1) to calculate the partition density.

We can easily extend this definition for weighted bipartite networks. Let $G = (U, V, E, W)$ be a weighted bipartite network, where $W = (w_{ij})_{p \times q}$ is the weight matrix, and w_{ij} is the weight of edge e_{ij} taking values in $[0, 1]$ without loss of generality. Let $W(U, V)$ be the sum of all edges' weight between node sets U and V , then we can define the weighted partition density WD of weighted bipartite network $G = (U, V, E, W)$ as follows:

$$WD = \sum_c \frac{W(U_c, V_c)}{W(U, V)} \cdot D_c = \frac{1}{W(U, V)} \sum_c \frac{W(U_c, V_c)^2}{|U_c| \times |V_c|}.$$

By maximizing the partition density D and WD , we can find the optimal community partition of unweighted and weighted bipartite networks respectively.

Definition 3. Core degree Given a partition of bipartite network $G = \{G_1, \dots, G_K\}$, for any node $u_i \in U$ and any community $G_c = (U_c, V_c, E_c)$, the core degree of node u_i to community G_c is defined as $CD(u_i, G_c) = \frac{|N(u_i) \cap V_c|}{|V_c|}$; similarly, for any node $v_j \in V$ and the community G_c , the core degree of node v_j to community G_c is defined as $CD(v_j, G_c) = \frac{|N(v_j) \cap U_c|}{|U_c|}$, where $N(u_i)$ and $N(v_j)$ denote the set of nodes adjacent to u_i or v_j in bipartite network G .

If a node has a high core degree to its community, the node might be an important member of the community, so we call it a core node. If a node has a low core degree to its community, it might not be an important member of the community, thus we call it a peripheral node. In a biclique, every node has core degree 1, and is a core node. A sparse bipartite network has plenty of peripheral nodes.

3 Community Partition Density Improves Resolution Limits

Although Barber's bipartite modularity has been widely used for community detection in bipartite networks, it may fail to detect communities smaller than a scale even in case where communities are bicliques.

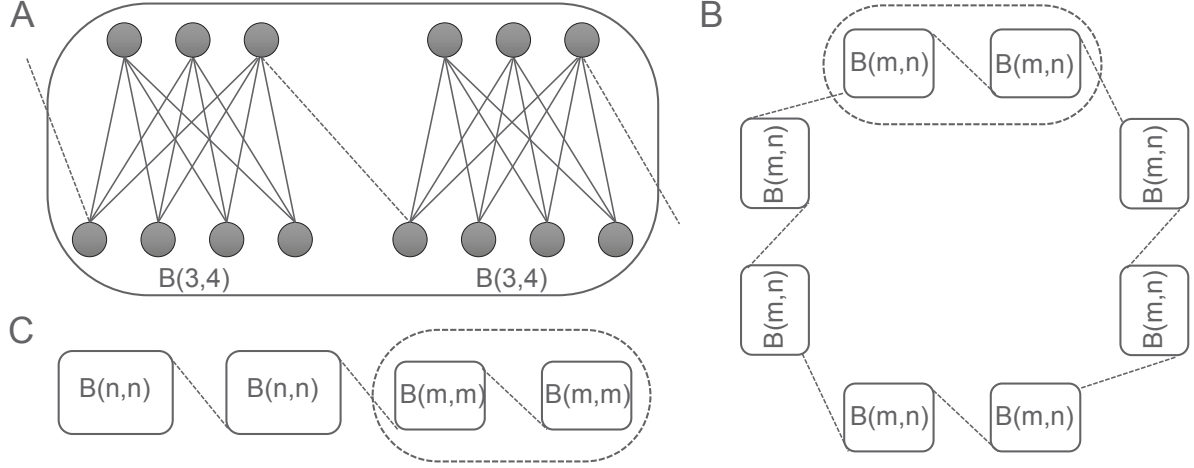


Figure 1. (A) The connection relationship of two bicliques in (B). (B) The ring of bicliques. Each square indicates a biclique of size $m + n$, and two adjacent bicliques are connected by one edge. (C) A bipartite network consists of four bicliques, two of which are $B(n, n)$ and the other two are $B(m, m)$.

The proposed partition density D can overcome such an issue to some extent. In the following, we will give some examples to demonstrate this point.

A ring of bicliques A ring of bicliques is a bipartite graph composed of s bicliques connected through single edges (see Figure 1A and 1B). Each biclique is a complete bipartite graph $B(m, n)$ with m nodes in U , n nodes in V ($m, n \geq 2$) and mn edges. The network has a total of $(m + n)s$ nodes and $(mn + 1)s$ edges.

The network has a clear community structure where each biclique corresponds to a single community, but the correct result cannot be obtained by optimizing Barber's bipartite modularity Q . If we partition the network into s communities, then each community corresponds to a single biclique. If we partition the network into $\frac{s}{2}$ communities, then each community consists of two adjacent bicliques. When $s > 2mn + 1$, we can easily see that Barber's bipartite modularity for these two partitions: $Q_s < Q_{\frac{s}{2}}$ (see Appendix A). Thus, by maximizing Barber's bipartite modularity, we cannot partition the network into reasonable communities.

Assuming that $s \geq 4$, and it can be exactly divided by an integer k , $k \geq 2$, now we calculate the partition density D_s , $D_{\frac{s}{2}}$, $D_{\frac{s}{3}}$ by dividing the network into s , $\frac{s}{2}$ and $\frac{s}{3}$ communities, respectively. We find that $D_s > D_{\frac{s}{2}}$ and $D_s > D_{\frac{s}{3}}$ with $mn \geq 2$, $s \geq 4$, which confirm that the partition density D_s has larger value.

Although the above analysis is conducted for the special partition that the k consecutive bicliques are considered as a single community, by similar argument we can deduce that such a result is actually valid for any kind of grouping bicliques (i.e. any combination of bicliques as communities). In fact, the partition density of any other combination of bicliques as communities is less than that of the special partition D_s mentioned above. Therefore, these results, along with the fact that the maximum value of D is obtained in complete bipartite networks lead to a conclusion that the maximum value of D corresponds to the correct partition with each single biclique as a community.

A bipartite network consisting of bicliques of different sizes Assume that there is a bipartite network consisting of four bicliques, two of which are $B(n, n)$, and the other two are $B(m, m)$, for $2 \leq m \leq n$ (see Figure 1C). This bipartite network has a natural community structure of four communities, where each biclique corresponds to one community. If we partition the bipartite network by optimizing Barber's bipartite modularity Q , we will obtain three communities, with two small bicliques merged into one community (Figure 1C).

Let Q_{separate} and D_{separate} be Barber's bipartite modularity and our partition density when the network is partitioned into four communities corresponding to the four bicliques in Figure 1C, and Q_{merge} and D_{merge} denote Barber's bipartite modularity and our partition density when two small bicliques merge into one community. We can see that $Q_{\text{separate}} < Q_{\text{merge}}$ with $n \geq m^2 + 1$, and the two small communities will merge into one community (Appendix A). In contrast, it is easy to see that $D_{\text{separate}} > D_{\text{merge}}$ with $m \geq 2$, which indicates the partition density D based optimization does not have the resolution problem mentioned above (Appendix A).

The above analysis is conducted for the special partition wherein two smaller bicliques are merged into a community with each other biclique as a separate community. With the fact that bicliques have maximum partition density, it is easy to see that any other partition has a lower partition density than the one with each biclique as a community. Therefore, the optimal value of partition density D corresponds to the correct partition. In contrast to bipartite modularity Q , optimizing partition density D can correctly detect communities of any size. In summary, optimization of partition density D can often achieve correct partitions while Barber's bipartite modularity not. Thus, we deduce that bipartite partition density D can act as a proper quantitative function for community detection in bipartite networks.

4 Integer Programming Model for Bipartite Community Partition

Let $U = \{u_1, u_2, \dots, u_p\}$, $V = \{v_1, v_2, \dots, v_q\}$ be two disjoint node sets of bipartite network $G = (U, V, E)$, and $A = (a_{ij})_{p \times q}$ be the adjacency matrix of G , where $a_{ij} = 1$ indicates there is an edge e_{ij} between nodes u_i and v_j , otherwise $a_{ij} = 0$. Note that $|E| = \sum_{i=1}^p \sum_{j=1}^q a_{ij}$ is the total number of edges in G . Let's define binary variables x_{ic} , y_{jc} to represent whether node u_i and node v_j belong to community G_c or not, respectively:

$$x_{ic} = \begin{cases} 1 & \text{if } u_i \in G_c \\ 0 & \text{otherwise,} \end{cases}$$

$$y_{jc} = \begin{cases} 1 & \text{if } v_j \in G_c \\ 0 & \text{otherwise.} \end{cases}$$

We can formulate the problem of partitioning bipartite network G into K communities into an integer programming model (Model-1):

$$\max D = \frac{1}{|E|} \sum_{c=1}^K \frac{(\sum_{i=1}^p \sum_{j=1}^q a_{ij} x_{ic} y_{jc})^2}{(\sum_{i=1}^p x_{ic})(\sum_{j=1}^q y_{jc})} \quad (2)$$

$$s.t. \begin{cases} \sum_{c=1}^K x_{ic} = 1 & i = 1, 2, \dots, p & (3) \\ \sum_{c=1}^K y_{jc} = 1 & j = 1, 2, \dots, q & (4) \\ x_{ic} \in \{0, 1\}; & i = 1, 2, \dots, p, c = 1, 2, \dots, K & (5) \\ y_{jc} \in \{0, 1\}; & j = 1, 2, \dots, q, c = 1, 2, \dots, K & (6) \end{cases}$$

where the objective function (2) is to maximize partition density D , constraints (3) and (4) indicate that each node must belong to one and only one community, and constraints (5) and (6) ensure that the variables are binary ones. We can easily solve the integer programming model by using the Lingo software for small-scale bipartite networks to see if the model can find communities properly. To partition a bipartite network into overlapping communities, we can simply relax the constraints (3) and (4) to allow a node to belong to more than one community by changing the constraints to $\sum_{c=1}^K x_{ic} \geq 1$ for $i = 1, 2, \dots, p$ and $\sum_{c=1}^K y_{jc} \geq 1$ for $j = 1, 2, \dots, q$. We represent such a model as Model-2.

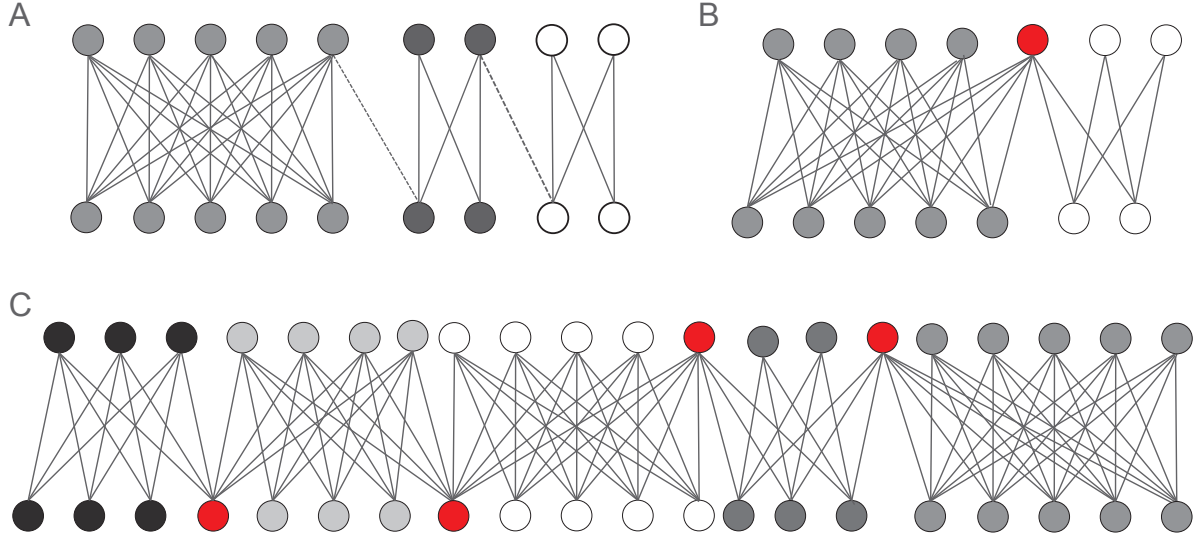


Figure 2. (A) Three communities of a bipartite network. The nodes in the same community are drawn in the same color, and the edges between communities are dotted. (B) Partitioning a bipartite network into two overlapping communities. The red node is overlapped by the two communities. (C) A chain of heterogeneous bicliques consisting of bicliques $B(3,4)$, $B(4,5)$, $B(5,5)$, $B(4,3)$ and $B(6,5)$. Two adjacent bicliques are overlapping with one red node.

For example, the bipartite network in Figure 2A has graph density 0.4321. We can easily partition it into two and three communities by solving Model-1 and obtain the partition density 0.6391 and 0.9429. If we partition it into more than 3 communities, the partition density will decrease and thus the maximum partition density is 0.9429 which seems reasonable. For the bipartite network in Figure 2B, we can partition it into two overlapping communities by solving Model-2 with the maximum partition density 1.

5 BiLPA for Bipartite Community Partition

To efficiently solve the community partition density, we further develop a revised label propagation algorithm (BiLPA) for community detection in large-scale bipartite networks, which does not need to pre-specify the number of communities K . Raghavan et al. [28] firstly proposed a label propagation algorithm (LPA) for community detection in large-scale unipartite networks. Barber and Clark [29] generalized it (LPAb) to optimize bipartite modularity. Liu and Murata [19, 20] further introduced an improved label propagation algorithm (LPAb+) for this task. However, both LPAb and LPAb+ are designed for maximizing Barber’s bipartite modularity. In this paper, we introduce a revised Label Propagation Algorithm (BiLPA) for community detection in large-scale bipartite networks by maximizing our partition density D .

Let $A = (a_{ij})_{p \times q}$ be the adjacency matrix of bipartite network G , and d_{u_i} and d_{v_j} be the degrees of node u_i and node v_j , respectively. The key idea of BiLPA is as follows. At first, each node $u_i \in U$ in the bipartite network is initially assigned a unique label, indicating its community membership (without causing confusion, we use r_i to indicate node u_i ’s label.). Then each node $v_j \in V$ acquires a label from its neighboring nodes in U , indicating its community membership (we use b_j to indicate the node v_j ’s label.). In the following steps, the labels of nodes in U and nodes in V are updated in turn, until a local maximum partition density is obtained. The updating rule can be described by the following formulae

respectively. For node v_j ,

$$b_j^{new} = \arg \max_g \frac{\sum_{i=1}^p a_{ij} \delta(g, r_i)}{\sum_{i=1}^p \delta(g, r_i)}. \quad (7)$$

For node u_i ,

$$r_i^{new} = \arg \max_g \frac{\sum_{j=1}^q a_{ij} \delta(g, b_j)}{\sum_{j=1}^q \delta(g, b_j)}. \quad (8)$$

where $g \in \{1, 2, \dots, K\}$ is a community label, and $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. Formulae (7) and (8) mean that a node should be assigned to a community to which the node has maximum core degree. The BiLPA pseudo-code is presented in **Algorithm BiLPA**.

Algorithm BiLPA:

- **Input**
Input the adjacency matrix A of bipartite network $G = (U, V, E)$, and calculate the number of nodes p ($p = |U|$) in U and the number of nodes q ($q = |V|$) in V , respectively. Set the maximum number of iterations $Iter_{\max}$ and the parameter value θ .
- **Output**
Output a community partition and its partition density.
- **Step 1.** Initialize the labels of nodes in U . For each node u_i , let $r_i = i$ and $t = 0$;
- **Step 2.** Calculate the number of nodes in U with the same labels. For each label k , calculate the number of nodes in U labeled with k using $R(k) = \sum_{i=1}^p \delta(k, r_i)$.
- **Step 3.** Update the labels of nodes in V . For each node v_j , update its label with

$$b_j^{new} = \arg \max_k \frac{\sum_{i=1}^p a_{ij} \delta(k, r_i)}{R(k)}.$$

- **Step 4.** Calculate the number of nodes in V with the same labels. For each label k , calculate the number of nodes in V labeled with k using $B(k) = \sum_{j=1}^q \delta(k, b_j)$.
- **Step 5.** Update the labels of nodes in U . For each node u_i , update its label with

$$r_i^{new} = \arg \max_k \frac{\sum_{j=1}^q a_{ij} \delta(k, b_j)}{B(k)}.$$

- **Step 6.** If $t < Iter_{\max}$, $t = t + 1$, go to Step 2.
- **Step 7.** Calculate the final label of each node and the corresponding partition density. For each node u_i in U (or similarly for node v_j in V), calculate its core degree $CD(u_i, k)$ to each community k and find its maximum core degree $\max_g CD(u_i, g)$. Calculate the core degree rate $RCD(u_i, k)$ with $RCD(u_i, k) = \frac{CD(u_i, k)}{\max_g CD(u_i, g)}$. For each community label k , determine the community membership of u_i by examining whether $RCD(u_i, k)$ is larger than θ or not, and assign the label to node u_i in U if so. Then calculate the final partition density.

We allow a node to belong to multiple communities in Step 7 which enables us to find overlapping nodes in a bipartite network.

6 Experiment Results

6.1 A chain of heterogeneous bicliques

We test our BiLPA algorithm on a type of exemplar bipartite networks, i.e. a chain network of K heterogeneous bicliques connected through single nodes (Figure 2C). Assuming that biclique $B_i = (U_i, V_i, E_i)$ has $s_i + t_i$ nodes and $|E_i| = s_i \times t_i$ edges, then the network has a total of $N = \sum_{i=1}^K (s_i + t_i) - K + 1$ nodes and $M = \sum_{i=1}^K |E_i|$ edges. The network has a clear bipartite community structure where each community corresponds to a single biclique, thus the optimal partition density is 1. Using the BiLPA algorithm described above, we can easily detect the optimal partition and identify the overlapping nodes.

6.2 Real-world Networks

In this subsection, we validate our algorithm on several real-world networks, including the women-event network, the Scotland corporate interlock network, the Malaria gene-substring network, and the protein complex-drug network. We applied BiLPA to these networks with $Iter_{\max} = 100$ and used θ to control the membership of each node. We can obtain a hard partition with $\theta = 1.0$ and overlapping communities with $\theta < 1.0$.

The women-event network The famous women-event bipartite network collected in 1930s described women’s attendance to social events in the town of Natchez and Mississippi [15, 25]. It consists of 18 women and 14 events with 89 edges between them. It has become a popular benchmark for discussing and exploring bipartite networks in the social science [15]. Guimerà [15] has analyzed the modules of both women and events by three methods, including an unweighted projection method, a weighted projection method and a bipartite modularity maximization method. The unweighted projection method did not capture the true modular structure of the network. The weighted projection method and the bipartite modularity approach capture the two-community structure of the network except that one woman was partitioned wrongly. We applied BiLPA to the women-event network and identified four overlapping communities with $\theta = 0.8$, consisting of {A3, A4, A5, A6, A9; B3, B4, B5, B7}, {A1, A2, A6, A7, A8, A9; B1, B2, B3, B5, B6, B8}, {A10, A11, A12, A13, A16; B8, B9, B10, B12, B13, B14}, and {A11, A14, A15, A17, A18; B9, B10, B11} respectively (Figure 3). The corresponding partition density is 0.601. If we use a smaller θ (e.g., $\theta = 0.7$), we can also partition the network into four communities. In such cases, more nodes will be determined as overlapping ones, and the partition density will be much larger. Compared with those obtained by Guimerà [15], the overlapping communities obtained by our method are more reasonable.

The Scotland corporate interlock network The Scotland corporate interlock network consisting of 244 nodes and 356 edges describes the corporate interlocks in Scotland in the beginning of the twentieth century (1904-1905). The 244 nodes are divided into two parts, where 136 nodes represent the board members who held multiple directorships, and 108 nodes denote the firms. The edges exist between each firm and its board members. The Scotland corporate interlock network is not connected. The largest component of the network contains 131 directors and 86 firms, which form 36 communities [35].

We applied our BiLPA algorithm to the largest component of the network. We determined 42 communities (Figure 4A) and obtained a partition density 0.524 with $\theta = 1.0$, which is similar to the real community structure. We can also determine 23 nodes belong to multiple communities with $\theta = 0.8$. However, by optimizing Barber’s modularity, the network can only be partitioned into 20 communities, and some small communities are merged into large communities. Obviously, their optimal solution is quite different from the real community structure of this network, indicating that optimizing our partition density with BiLPA algorithm provides higher quality community structure than other methods.

The Malaria gene-substring network The Malaria gene-substring network comes from the malaria parasite *P. falciparum*. The parasite can evade the human immune system via a protein camouflage encoded in *var* genes [45, 46]. In order to create novel camouflages, *var* genes frequently recombine, which amounts to the constrained splicing and shuffling of genetic substrings, giving rise to community structures naturally [46]. The network consists of two types of nodes corresponding to 297 genes and 806

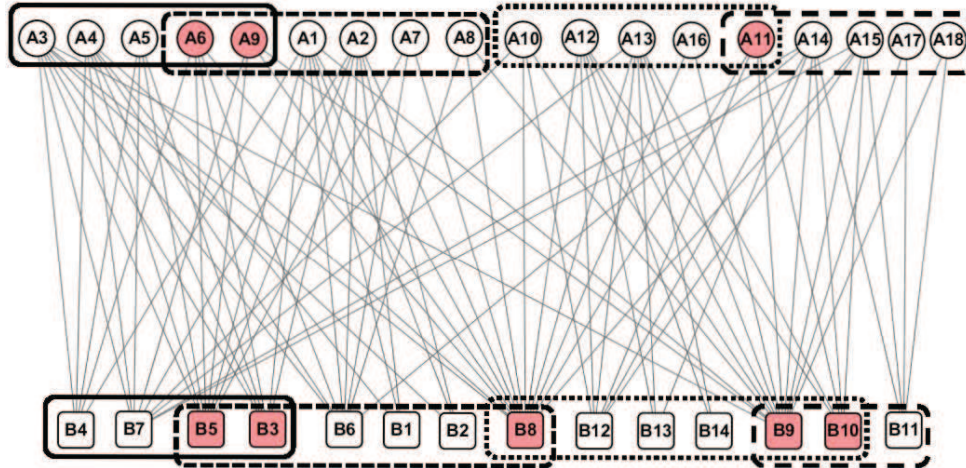


Figure 3. The four overlapping communities of women-event network detected by BiLPA. The red nodes A6, A9, A11, B3, B5, B8, B9, B10 are overlapping ones.

constituent substrings respectively. Each substring connects to a gene in which it is present. There are 2965 edges in the network and the link density of the bipartite network is 0.012.

In [46], the network is partitioned into three coarse communities. We applied our BiLPA algorithm to the network and identified 158 hard communities with a maximal partition density 0.41 using $\theta = 1.0$ (Figure 4B). From Figure 4B, we can see that this network is very sparse and thus partitioning it into three communities is unreasonable. Our method obtains very detailed communities. For example, BiLPA detects a biclique community consisting of four genes and 14 substring (the circled community in Figure 4B), while it was merged into a very large one in [46]. This result indicates that BiLPA can find better community structure and conquer the resolution limit showed in [46].

The protein complex-drug network The bipartite network of protein complexes and drugs investigated by Nacher and Schwartz [47] contains 1419 nodes (680 drugs and 739 complexes) and 3690 edges. This bipartite network has a small number of nodes connected to many edges and the majority of nodes connected with a few edges. There are both hubs of drugs and complexes. The most connected nodes are two drugs, flavopiridol (DB03496), which is used for the treatment of leukemia, and vorinostat (DB02546), which is used for the treatment of lymphoma, both interacting with over 90 complexes. A number of complexes interact with over 40 drugs, including several closely related complexes involving ER- α , BKCA- β and ESR1. The degree of top drug hubs decreases faster but a number of drugs interact with 20 to 40 complexes.

Nacher and Schwartz identified network communities in both the bipartite network itself and the projected networks [47]. The number of modules in the drug projection network is 23, while the number of modules in the protein complex projection network is 17. Using the bipartite network itself, they detected 48 modules of drugs and 42 modules of complexes. Since the network consists of 27 disconnected subgraphs, any two or more disconnected subgraphs should not belong to one community, so the number of communities (either the number of drug communities or the number of protein complex communities) in this bipartite network should not be less than 27. However, Nacher and Schwartz only detected 23 drug communities and 17 protein complex communities [47], both of which are less than 27. In the results of communities obtained from the original bipartite network, although the authors detected more drug communities (48) and protein complex communities (42), some communities consist of two or more disconnected subgraphs, and some other communities consist of two or more weakly connected subgraphs.

We applied our BiLPA to this network and partitioned it into 122 communities with partition density 0.57 (Figure 5A). We can see that each small disconnected subgraph is identified as one community. The giant connected subgraph is partitioned into about 90 communities. This network is very sparse and most

communities are very sparsely connected with others, but densely connected within themselves. For example, we demonstrated two distinct complex-drug communities in Figure 5B and Figure 5C. The circled community in Figure 5B consists of four complexes (CORUM_0441, CORUM_1054, CORUM_2124 and CORUM_5862) and 40 drugs. Most drugs are fully connected with these four complexes. Another circled dense community in Figure 5C consists of five complexes (CORUM_1223, CORUM_2242, CORUM_4158, CORUM_5189, CORUM_5526) and 16 drugs. The bipartite communities reveal that some drugs target to a same set of biological complexes. This is valuable for prediction of the detailed mechanisms of unknown ones from the well-known drugs' functions.

7 Conclusion and discussion

Community structure is one of the main characteristics of bipartite networks and is valuable for understanding the organization and functions of such networks. In this paper, we propose a new quantitative function — bipartite partition density — for evaluating the community partition of bipartite networks. We show that our quantitative function can conquer the serious resolution limit issues of modularity-based functions. We first formulate the community detection problem for bipartite networks into an integer nonlinear programming model. Furthermore, we design a fast heuristic algorithm BiLPA for solving the community detection problem and conduct validation experiments on some artificial and real-world networks. The computational results demonstrate that our model and algorithm can detect overlapping communities in bipartite networks. Moreover, the quantitative function proposed in this paper can be easily extended to the community detection in weighted bipartite networks.

Acknowledgments

This project was supported by the National Natural Science Foundation of China, No. 61379092, 61422309 and 11131009, the Foundation for Members of Youth Innovation Promotion Association, CAS. The Outstanding Young Scientist Program of CAS, the Scientific Research Foundation for ROCS, SEM, the Key Laboratory of Random Complex Structures and Data Science at CAS, the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (CIT&TCD20130327).

References

- [1] Albert R and Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74: 47-97.
- [2] Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45: 167-256.
- [3] Hu Y, Chen H, Zhang P, Li M, Di Z and Fan Y (2008) Comparative definition of community and corresponding identifying algorithm. *Phys Rev E* 78: 026121.
- [4] Zhang S, Jin G, Zhang XS and Chen L (2007) Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics* 7: 2856-2869.
- [5] Fortunato S (2010) Community detection in graph. *Physics Reports* 486: 75-174.
- [6] Newman MEJ (2012) Communities, modules and large-scale structure in networks. *Nature Physics* 8: 25-31.
- [7] Zhang S, Zhao J and Zhang XS. Common community structure in time-varying networks. *Physical Review E* 85, 056110.
- [8] Newman, MEJ and Girvan M (2004) Finding and evaluating community structure in networks, *Phys. Rev. E* 69: 026113.
- [9] Fortunato S and Barthélemy M (2007) Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* 104: 36-41.

- [10] Rosvall M and Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. USA* 104: 7327-331.
- [11] Li Z, Zhang S, Wang R, Zhang XS and Chen L (2008) Quantitative function for community detection. *Physical Review E* 77: 36109.
- [12] Zhang S, Ning XM, Ding C, Zhang XS (2010) Determining modular organization of protein interaction networks by maximizing modularity density. *BMC systems biology* 4(S2), S10.
- [13] Ahn YY, Bagrow JP and Lehmann S (2010) Link communities reveal multi-scale complexity in networks. *Nature* 466: 761-764.
- [14] Li Z, Zhang XS, Wang RS, Liu H and Zhang S (2013) Discovering link communities in complex networks by an integer programming model and a genetic algorithm, *PLoS ONE* 8(12): e83739.
- [15] Guimera R, Sale-Pardo M and Nunes Amaral L (2007) Module identification in bipartite and directed networks, *Physical Review E* 76: 036102.
- [16] Zhan W, Zhang Z, Guan J and Zhou S (2011) Evolutionary method for finding communities in bipartite networks. *Physical Review E* 83: 066120.
- [17] Costa A and Hansen P (2011) Comment on "Evolutionary method for finding communities in bipartite networks". *Physical Review E* 84(5): 058101.
- [18] Costa A and Hansen P (2014) A locally optimal hierarchical divisive heuristic for bipartite modularity maximization, *Optim Lett* 8: 903-917.
- [19] Liu X and Murata T (2009) Community detection in large-scale bipartite networks. In: *IEEE/WIC/ACM international conference on web intelligence and Intelligent Agent Technologies*, pp. 50-57.
- [20] Liu X and Murata T (2010) An efficient algorithm for optimizing bipartite modularity in bipartite networks. *J. Adv. Comput. Intell. Intell. Inform.* 14(4): 408-415.
- [21] Lee J and Lee J (2013) Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction. *PLoS ONE* 8(4): e60372.
- [22] Mao C (2012) A heuristic algorithm for bipartite community detection in social networks. *Journal of Software*, 7(1): 204-211.
- [23] Murata T and Ikeya T (2010) A new modularity for detecting one-to-many correspondence of communities in bipartite networks. *Advances in Complex Systems*, 13(1):19-31
- [24] Barber MJ (2007) Modularity and community detection in bipartite networks, *Phys. Rev. E* 76: 066102.
- [25] Freeman LC (2003) Finding social groups: a meta-analysis of the Southern Women data. In: Breiger, R., Carley, C., Pattison, P. (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. National Research Council, The National Academies Press, Washington, DC, pp. 39-97.
- [26] Suzuki S and Wakita K (2009) Extracing multi-facet community structure from bipartite networks, *Proc. of 2009 Int. Conf. on Computational Science and engineering (CSE 09)*, Vol.4, pp.312-319, August 2009.
- [27] Murata T (2009) Detecting communities from bipartite networks based on bipartite modularities, *Proc. of the 2009 International Conference on Computational Science and Engineering (CSE 09)*, Vol.4, pp. 50-57, August 2009.
- [28] Raghavan UN, Albert R and Kumara S. (2007) Near linear time algorithm for detect community structures in large scale networks, *Physical Review E* 76: 036106.
- [29] Barber MJ and Clark JW (2009) Detecting network communities by propagating labels under constraints, *Phys. Rev. E* 80: 026129.
- [30] Evans TS and Lambiotte R (2009) Line graphs, link partitions and overlapping communities. *Phys Rev E* 80: 016105.

- [31] Evans TS (2010) Clique graphs and overlapping communities. *J Stat Mech*: P12037.
- [32] Evans TS and Lambiotte R (2010) Line graphs of weighted networks for overlapping communities. *Eur Phys J B* 77: 265-272.
- [33] Zhang S, Liu HW, Ning XM and Zhang XS (2009) A hybrid graph-theoretic method for mining overlapping functional modules in large sparse protein interaction networks. *International Journal of Data Mining and Bioinformatics* 3: 68-84.
- [34] He DX, Liu D, Zhang W, Jin D and Yang B (2012) Discovering link communities in complex networks by exploiting link dynamics. *J Stat Mech*: P10015.
- [35] Chen BL, Chen L, Zhou SR and Xu XL (2013) Detecting community structure in bipartite networks based on matrix factorisation, *Int. J. Wireless and Mobile Computing*, 6(6): 599-607.
- [36] Leicht EA and Newman MEJ (2008) Community structure in directed networks, *Phys. Rev. Lett.* 100: 118703
- [37] Kim S and Shi T (2012) Scalable spectral algorithm for community detection in directed networks, *arXiv:1211.6807v1*.
- [38] Zhang S, Wang RS and Zhang XS (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A* 374: 483-490.
- [39] Tasgin M and Bingol H (2006) Community detection in complex networks using genetic algorithm. <http://arxiv.org/abs/0711.0491v1>.
- [40] Gregory S. (2010) Finding overlapping communities in networks by label propagation. *New Journal of Physics* 12: 103018.
- [41] Lehmann S, Schwartz M and Hansen LK (2008) Biclique communities, *Physical Review E* 78: 016108.
- [42] Du N, Wang B, Wu B and Wang Y (2008) Overlapping community detection in bipartite networks, *Proc. of the 2008 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT 08)*, Vol. 1, pp. 176-179, December 2008.
- [43] Miyauchi A and Sukegawa N (2013) Maximizing Barber's bipartite modularity is also hard. *arXiv:1310.4656v1*.
- [44] Souam F, Aïtelhadj A and Baba-Ali R (2014) Dual modularity optimization for detecting overlapping communities in bipartite network, *Knowl. Inf. Syst.* 40: 455-488.
- [45] Larremore DB, Clauset A and Buckee CO (2013) A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Computational Biology* 9(10): e1003268.
- [46] Larremore DB, Clauset A and Jacobs AZ (2014) Efficiently inferring community structure in bipartite networks. *Physical Review E* 90: 012805.
- [47] Nacher JC, Schwartz JM (2012) Modularity in protein complex and drug interactions reveals new polypharmacological properties. *PLoS ONE* 7(1): e30028.

Appendix A: Community Partition Density Improves Resolution Limits

A ring of bicliques The ring of bicliques has a distinct community structure where each biclique corresponds to a single community, but the correct result cannot be obtained by optimizing Barber's bipartite modularity Q . If we partition the network into s communities, and each community corresponds to a single biclique, then Barber's bipartite modularity Q is

$$\begin{aligned}
 Q_s &= \left(\frac{mn}{s(mn+1)} - \frac{mn+1}{s(mn+1)} \frac{mn+1}{s(mn+1)} \right) \cdot s \\
 &= \frac{mn}{mn+1} - \frac{1}{s}.
 \end{aligned}$$

If we partition the network into $\frac{s}{2}$ communities, and each community consists of two adjacent bicliques, then Barber's bipartite modularity Q is

$$\begin{aligned} Q_{\frac{s}{2}} &= \left(\frac{2mn+1}{s(mn+1)} - \frac{2mn+2}{s(mn+1)} \frac{2mn+2}{s(mn+1)} \right) \cdot \frac{s}{2} \\ &= \frac{2mn+1}{2(mn+1)} - \frac{2}{s}. \end{aligned}$$

When $s > 2(mn+1)$, we have

$$\begin{aligned} Q_s - Q_{\frac{s}{2}} &= \frac{mn}{mn+1} - \frac{1}{s} - \frac{2mn+1}{2(mn+1)} + \frac{2}{s} \\ &= \frac{1}{s} - \frac{1}{2(mn+1)} \\ &< 0. \end{aligned}$$

The partition density D_s of the natural partition can be easily and analytically calculated as follows:

$$D_s = \frac{1}{s(mn+1)} \left(\frac{(mn)^2}{mn} \cdot s \right) = \frac{mn}{mn+1}.$$

On the other hand, the partition density D_k of the partition in which the k consecutive bicliques are considered as a single community is

$$D_{\frac{s}{k}} = \frac{1}{s(mn+1)} \left(\frac{(kmn+k-1)^2}{(km)(kn)} \cdot \frac{s}{k} \right) = \frac{(kmn+k-1)^2}{k^3mn(mn+1)}.$$

Then,

$$\begin{aligned} D_s - D_{\frac{s}{k}} &= \frac{mn}{mn+1} - \frac{(kmn+k-1)^2}{k^3mn(mn+1)} \\ &= \frac{k^3(mn)^2 - (kmn+k-1)^2}{k^3mn(mn+1)} \\ &= \frac{(mn)(k - (1 + \frac{k-1}{kmn})^2)}{k(mn+1)}. \end{aligned}$$

Supposing $mn \geq 2$, $s \geq 4$, when $k = 2$,

$$\begin{aligned} D_s - D_{\frac{s}{2}} &= \frac{(mn)(2 - (1 + \frac{1}{2mn})^2)}{2(mn+1)} \\ &> 0. \end{aligned}$$

When $k \geq 3$,

$$\begin{aligned} D_s - D_{\frac{s}{3}} &= \frac{(mn)(k - (1 + \frac{k-1}{kmn})^2)}{k(mn+1)} \\ &> \frac{(mn)(k - (1 + \frac{1}{mn})^2)}{k(mn+1)} \\ &\geq \frac{(mn)(k - (1 + \frac{1}{2})^2)}{k(mn+1)} \\ &> 0. \end{aligned}$$

A bipartite network consisting of bicliques of different sizes Suppose there is a bipartite network consisting of four bicliques, two of which are $B(n, n)$, and the other two are $B(m, m)$, for $2 \leq m \leq n$. This bipartite network has a natural community structure of four communities, where each biclique corresponds to one community. If we partition the bipartite network by optimizing Barber's bipartite modularity Q , we will obtain three communities, where two small bicliques merge into one community. We can calculate their corresponding Q as follows:

$$\begin{aligned}
Q_{\text{separate}} &= \left(\frac{n^2}{2n^2 + 2m^2 + 3} - \frac{n^2 + 1}{2n^2 + 2m^2 + 3} \frac{n^2}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{n^2}{2n^2 + 2m^2 + 3} - \frac{n^2 + 1}{2n^2 + 2m^2 + 3} \frac{n^2}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{m^2}{2n^2 + 2m^2 + 3} - \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \frac{m^2}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{m^2}{2n^2 + 2m^2 + 3} - \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \frac{m^2}{2n^2 + 2m^2 + 3} \right). \\
\\
Q_{\text{merge}} &= \left(\frac{n^2}{2n^2 + 2m^2 + 3} - \frac{n^2 + 1}{2n^2 + 2m^2 + 3} \frac{n^2}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{n^2}{2n^2 + 2m^2 + 3} - \frac{n^2 + 1}{2n^2 + 2m^2 + 3} \frac{n^2}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{2m^2 + 1}{2n^2 + 2m^2 + 3} - \frac{2m^2 + 1}{2n^2 + 2m^2 + 3} \frac{2m^2 + 2}{2m^2 + 2} \right). \\
\\
Q_{\text{separate}} - Q_{\text{merge}} &= \left(\frac{m^2}{2n^2 + 2m^2 + 3} - \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \right) \\
&+ \left(\frac{m^2}{2n^2 + 2m^2 + 3} - \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \frac{m^2 + 1}{2n^2 + 2m^2 + 3} \right) \\
&- \left(\frac{2m^2 + 1}{2n^2 + 2m^2 + 3} - \frac{2m^2 + 1}{2n^2 + 2m^2 + 3} \frac{2m^2 + 2}{2m^2 + 2} \right) \\
&= \frac{-2n^2 + 2m^4 + m^2 - 2}{(2n^2 + 2m^2 + 3)^2}.
\end{aligned}$$

If $n \geq m^2 + 1$, then $Q_{\text{separate}} - Q_{\text{merge}} < 0$. This means that modularity Q -based optimization will merge two smaller bicliques $B(m, n)$ s into one community.

In the following, we prove that partition density D -based optimization does not have this resolution limit.

$$\begin{aligned}
D_{\text{separate}} &= \frac{1}{2n^2 + 2m^2 + 3} (n^2 + n^2 + m^2 + m^2) \\
&= \frac{2n^2 + 2m^2}{2n^2 + 2m^2 + 3}. \\
\\
D_{\text{merge}} &= \frac{1}{2n^2 + 2m^2 + 3} \left(n^2 + n^2 + \frac{(2m^2 + 1)^2}{4m^2} \right).
\end{aligned}$$

It is easy to verify that when $m \geq 2$,

$$D_{\text{separate}} - D_{\text{merge}} = \frac{(2m^2 - 1)^2 - 2}{4m^2(2n^2 + 2m^2 + 3)} > 0.$$

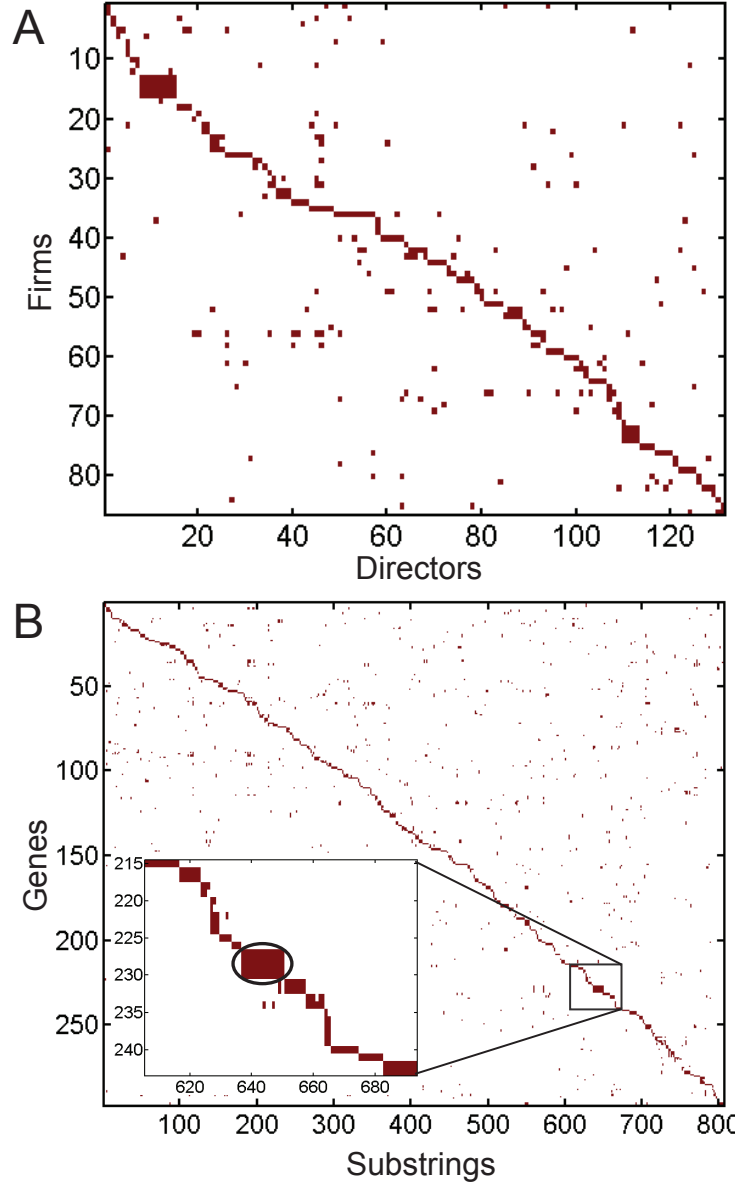


Figure 4. The communities of the Scotland corporate interlock network (A) and the Malaria gene-substrings network (B) detected by BiLPA.

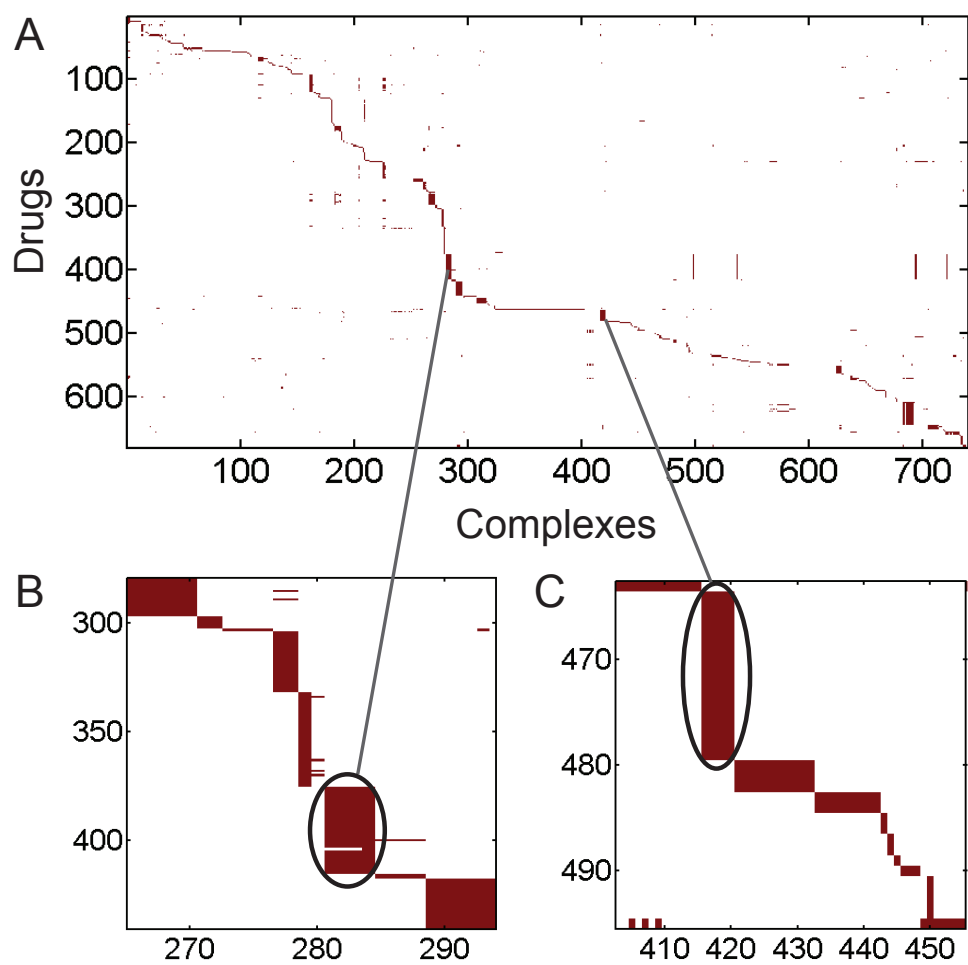


Figure 5. The communities of the protein complex-drug network detected by BiLPA (A) and two illustrative community examples in (B) and (C).