A uniform framework of projection and community detection for one-mode network in bipartite networks*

Guolin Wu(吴果林)[1,2,3], Changgui Gu(顾长贵)[1,†], Lu Qiu(邱路)[4], and Huijie Yang(杨会杰)[1]

[1] *Business School, University of Shanghai for Science and Technology, Shanghai* 200093, *China*
[2] *Faculty of Science, Guilin University of Aerospace Technology, Guilin* 541004, *China*
[3] *Guangxi Aviation Logistics Research Center, Guilin University of Aerospace Technology, Guilin* 541004, *China*
[4] *School of Finance and Business, Shanghai Normal University, Shanghai* 200234, *China*

(Received 9 April 2017; revised manuscript received 26 August 2017; published online 20 October 2017)

Projection is a widely used method in bipartite networks. However, each projection has a specific application scenario and differs in the forms of mapping for bipartite networks. In this paper, inspired by the network-based information exchange dynamics, we propose a uniform framework of projection. Subsequently, an information exchange rate projection based on the nature of community structures of a network (named IERCP) is designed to detect community structures of bipartite networks. Results from the synthetic and real-world networks show that the IERCP algorithm has higher performance compared with the other projection methods. It suggests that the IERCP may extract more information hidden in bipartite networks and minimize information loss.

Keywords: bipartite networks, community, projection, information exchange

PACS: 89.75.Fb, 89.75.Hc **DOI:** 10.1088/1674-1056/26/12/128901

1. Introduction

Many real-world complex systems, ranging from socio-economic to biological and medical fields, can be considered as bipartite networks, such as bus transport networks,[1,2] plants-pollinators networks,[3,4] phage-bacteria infection networks,[5,6] scientific collaboration networks,[7,8] recommender systems networks,[9–11] and so on. These networks describe relationships of interactions between two disjoint sets of nodes.[12] At present, the study of bipartite networks is mainly focused on nestedness and community structures.[13] Community, although standard definition is not given, is generally considered as a set of nodes with dense connections internally and sparser connections between sets. Identifying community structures of a bipartite network is helpful for understanding how its network topology and function interact with each other.[14] For example, if we know the community structures of users' preferences in the user-web networks,[15,16] the enterprises can target their advertisements to users selectively.

In the past decade, detecting the community structures of bipartite networks has become a popular topic. Many methods have been proposed for community detection. In general, these methods are roughly divided into two main categories. One is called the projection method, which projects the bipartite network into two one-mode networks and then uses the existing methods to detect communities for the one-mode networks. Another type of methods is to identify communities by directly analyzing structures of bipartite networks (named the combined methods for short).[16,18–27] In particular, the projection methods include unweighted projection,[28,29] simple weighted projection (SWP),[18,30–32] hyperbolic weighted projection (HWP),[8] weighted projection based on resource allocation (WRAP),[33,34] clustering triangular projection (CTP),[35] and so on. The simplest method is the unweighted projection, which does not take into account the frequencies of jointly connecting nodes. Unfortunately, the unweighted projection network will become denser and denser when the number of nodes increases in the opposing set.[18] Due to this dense network, the unweighted projection algorithm fails to identify community structures. A way to obtain the original network topology is to use the weighted methods. The SWP measures the weights of edges of one-mode network by calculating the frequencies of jointly connecting nodes. This SWP can effectively extract the structure information of bipartite networks especially when the frequencies of jointly connecting nodes are interpreted as the number of opportunities for goods or flow of information or interaction.[36] In some cases, if we want to reveal unseen relationships be-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11505114 and 10975099), the Program for Professor of Special Appointment (Orientational Scholar) at Shanghai Institutions of Higher Learning (Grant Nos. QD02015016 and DUSST02), the Shanghai Project for Construction of Discipline Peaks, the Natural Science Foundation of Guangxi Zhuang Guangxi Zhuang Autonomous Region (Grant No. 2016GXNSFDA380031), and the Fundamental Ability Enhancement Project for Young and Middle-aged University Teachers in Guangxi Zhuang Autonomous Region (Grant No. 2017KY0859).

†Corresponding author. E-mail: gu_changgui@163.com

© 2017 Chinese Physical Society and IOP Publishing Ltd

tween nodes in the same type in bipartite networks, the SWP is obviously inappropriate. For example, in the scientific collaboration network, two scientists whose names appear on a paper together with many other coauthors are expected to know one another less than two who are the sole two authors of a paper.[8] In order to measure the unseen relationships, Newman proposed a HWP model. It normalizes the weights of edges which are multiplied by a factor $1/(n-1)$, where n represents the number of jointly connecting nodes. Recently, a new projection method was proposed,[35] which projected a bipartite network into two weighted one-mode networks by bipartite clustering triangular. The CTP algorithm for detecting one-mode communities in bipartite networks mainly consists of two steps. The first step acquires two weighted one-mode networks. In the second step, the algorithm extracts all maximal complete sub-graphs from two one-mode networks and merges the maximal sub-graphs by the weighted clustering threshold. A merit for the CTP algorithm is to detect overlapping one-mode communities. However, the performance of this algorithm is greatly affected by the weighted clustering threshold whose definition is subjective. A blemish in these methods mentioned above is that information contained by edges whose "target" nodes are of degree equal to 1 in the original network will be lost in the projection. In order to overcome the shortcoming, Zhou *et al.*[33] proposed a projection method (WRAP), which assigned the weights of edges based on resource allocation between nodes. The WRAP generates a network with directional weights. Numerical simulations indicate that the WRAP is preferred to the collaborative filtering algorithm, which is a widely used method in recommender systems. However, articles where communities are detected by the WRAP have not been reported. A review for more projection methods can be found in the Ref. [37].

In many cases, people may be more interested in the one-mode networks for bipartite networks. For example, in affiliation networks of a social field, we may be more interested in the relationships between people than the relationships between people and events which they affiliate to or participate in. An effective method to deal with the problem is the projection model. In fact, there are also many advantages for the projection model. At first, the combined methods do not directly detect the community structures of projected one-mode networks. To identify the communities of one type of nodes, the combined methods must analyze the community structures of the original bipartite networks firstly, and then make a projection onto the nodes in the same type. The identifying process is just opposite to the projection model. Secondly, a proper projection method can extract more valuable information from bipartite networks and reduce the information to be lost. Guimera[18] found that the performances for the weighted projection and bipartite modularity method are very similar. In particular, when each community in the bipartite network contains the same number of nodes, the two kinds of methods are almost equivalent.[32] Thirdly, there are many excellent community detection algorithms for one-mode networks, such as the GN algorithm,[38] Infomap algorithm,[39] overlapping community detection algorithm,[40–44] and core-vertex and maximal sub-graphs algorithms.[45,46] We select these algorithms to detect the community structures of the projected one-mode networks in which we are interested, and do not need to design additional algorithms for these networks.

Based on these analyses mentioned above, we select the projection model for detecting one-mode communities in bipartite networks. However, these projection methods mentioned above do not set proper edge weights for the projected one-mode networks. They set either identical edge weights or impertinent weight definitions, and do not take into account the characteristics of the community. Although there are some uniform frameworks for community detection, for example, Long *et al.*[47] generalized traditional graph partitioning methods and Kim *et al.*[48] proposed a generalized form of modularity in directed networks, there does not exist a uniform framework for projection in bipartite networks at present. Thus, it is necessary to set up a uniform framework for projection especially when different types of nodes need to choose different projection patterns in bipartite networks. We hold a point of view that network analysis is usually considered as an approach to study the information exchanges.[49] Each relationship between nodes can be considered as a particular type of information exchange in the network. In this paper, we regard the edges as the channels of information exchanges and propose a uniform framework of projection in bipartite networks. Then, an IERCP projection is deduced by taking into account the property of cohesion of community structures. At last, we apply the IERCP to identify the community structures of the bipartite network, and find that the IERCP can extract more hidden information from bipartite networks when the community structures are detected. As shown in the synthetic networks, the performance of IERCP is better than other projection methods, such as SWP, HWP, CTP, and WRAP. For the real-world networks, communities identified by our method are proven to be more reasonable than the other four methods.

The following sections of this paper are organized as follows. In Section 2, we describe information exchanges between the same type of nodes, give a uniform framework of projection, and propose an IERCP algorithm for community detection in the bipartite network. Subsequently, the algorithm is applied to capturing communities in both synthetic and real-

world networks, and the detailed results and analyses of the experiments are displayed in Section 3. At last, conclusions and discussions are put forward.

2. Method

2.1. Information exchanges in bipartite networks

The bipartite network is composed of two disjoint sets of nodes and edges across the nodes in the two sets. For instance, the scientific collaboration network is a bipartite network, whose nodes are composed of authors and papers. In the network, each paper connects with all of its authors and edges represent the relationships of writing between authors and papers. In many cases, our purpose of collecting two-mode data is not to analyze the pattern of relationships between two disjoint sets, but to analyze the pattern of relationships between one-mode nodes and to examine the availability and the exchanges of resources between these nodes. For example, in the scientific collaboration network, we may be interested in analyzing the relationships among authors and exploring the community structures of authors. It is an obvious fact that the relationships between authors are identified only by the intersections of another-mode nodes which they connect with. So an extensively used method for extracting the information from the bipartite network is the projection.

Inspired by the network-based resource-allocation dynamics, Zhou *et al.*[33] proposed a new projection method, which could extract better the hidden information of networks than other projection methods, such as SWP and HWP. It generates the one-mode network with directional weights. This projection is validated that it is a very effective way to provide personalized recommendations. Because the relationships that community reflects between nodes are undirected, this projection is not suitable for community detection. In this section, we regard the edges as the channels of information exchanges, and obtain a uniform framework of projection of the bipartite network by the network-based information exchange dynamics. For the convenience of the following descriptions, we briefly introduce some basic concepts and notations in bipartite networks.

Let $G = (U, V, E)$ denote a bipartite network, where U and V are two disjoint sets of nodes and E is a set of edges that connect nodes across the two disjoint sets. $u_1, u_2, \ldots, u_m$ and $v_1, v_2, \ldots, v_n$ denote the nodes in U and V, respectively. Normally, U is called the bottom set, and V is called the top set. $A = (a_{\mathrm{st}})_{m\times n}$ denotes an adjacency matrix for the bipartite network G and $d(i)$ denotes the degree of the node i. Figure 1 shows an example for a bipartite network.

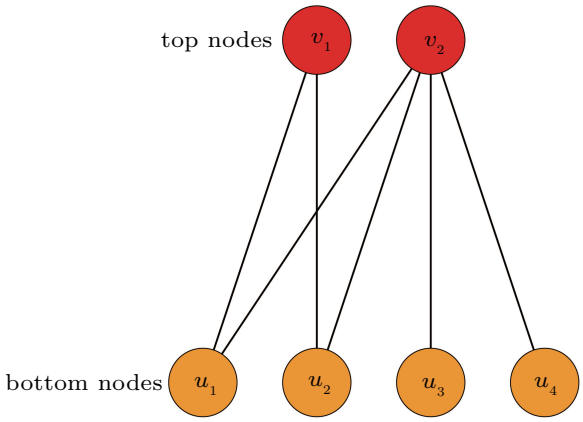


Fig. 1. (color online) A bipartite network. Orange represents bottom nodes. Red represents top nodes.

Let $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ represent the information located on $u_1, u_2, \ldots, u_m$ and $v_1, v_2, \ldots, v_n$, respectively. $f_{v_l}(x_k)$ represents the information that the bottom node u_k flows to the top node v_l, which is named as the information exchange function. $f_{u_k}(y_l)$ represents the information that the top node v_l flows to the bottom node u_k. Without loss of generality, we use bottom nodes as an example to discuss the information exchanges between the same type of nodes. At first, when all the information in the bottom nodes flows to the top nodes, the information in the top node v_l is

$$\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k), \quad l = 1, 2, \ldots, n. \tag{1}$$

Secondly, all the information in the top nodes flows back to the bottom nodes, and the final information located on the bottom node u_i is

$$EI_U(i) = \sum_{l=1}^{n} a_{il} f_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big], \quad i = 1, 2, \ldots, m. \tag{2}$$

Apparently, $EI_U(i)$ represents the amount of information that all the bottom nodes flow to the node u_i through the top nodes. Taking into account the fact that the bipartite network we discuss is an undirected graph, the total information shared by nodes u_i and u_j is

$$\sum_{l=1}^{n} a_{il} a_{jl} \Big(f_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] + f_{u_j}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big]\Big). \tag{3}$$

So the average information exchanged between nodes u_i and u_j is

$$EI_U(i, j) = \sum_{l=1}^{n} a_{il} a_{jl} \frac{f_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] + f_{u_j}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big]}{2}. \tag{4}$$

It is not difficult to find that $EI_U(i,j)$ reflects the extent of relationship between the bottom nodes u_i and u_j.

2.2. Uniform framework of projection

In many cases, we are not interested in investigating the amount of information exchanged between nodes, but rather in investigating the rate of information exchanged between nodes. Because the rate of information exchanged reflects the intimacies between nodes more clearly. Intuitively, a simple and effective way to obtain the information exchange rate is to solve the derivative of the information exchange function. In order to seek the derivative of information exchange function, we come back to Eq. (2) again. In the bottom nodes, $EI_U(i)$ represents the quantity of information that all the bottom nodes flow to the node u_i across the top nodes. Solving the derivative of $EI_U(i)$, we acquire

$$\frac{\partial(EI_U(i))}{\partial x_k} = \sum_{l=1}^{n} a_{il} f_{u_i}{}' \Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] a_{kl} f_{v_l}{}'(x_k), \quad k = 1,2,\ldots,m. \tag{5}$$

The term $\partial(EI_U(i))/\partial x_k$ represents the IER that node u_k flows to node u_i. Since the exchanged information between nodes u_i and u_j comes not only from two nodes, but also from other nodes. So, before we investigate the IER between two nodes, we must calculate the total IER which all the bottom nodes flow to the nodes u_i or u_j firstly. Integrating the Eq. (5) for all the k, we obtain the total IER in which all the bottom nodes flow to nodes u_i,

$$\begin{aligned} IER_U(i) &= \sum_{k=1}^{m}\sum_{l=1}^{n} a_{il} f'_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] a_{kl} f'_{v_l}(x_k) \\ &= \sum_{l=1}^{n} a_{il} f'_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] \sum_{k=1}^{m} a_{kl} f'_{v_l}(x_k). \end{aligned} \tag{6}$$

Analogously, the total IER in which all the bottom nodes flow to nodes u_j is

$$IER_U(j) = \sum_{l=1}^{n} a_{jl} f'_{u_j}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] \sum_{k=1}^{m} a_{kl} f'_{v_l}(x_k). \tag{7}$$

Therefore, the average IER between nodes u_i and u_j is

$$IER_U(i,j) = \sum_{l=1}^{n} a_{il} a_{jl} \sum_{k=1}^{m} a_{kl} f'_{v_l}(x_k)\Big\{f'_{u_i}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big] + f'_{u_j}\Big[\sum_{k=1}^{m} a_{kl} f_{v_l}(x_k)\Big]\Big\}/2. \tag{8}$$

Equation (8) is a uniform framework of projection (also named as IER projection), which is a general form for the bipartite networks projection.

2.3. IERCP algorithm

Generally, a network is said to have community structures if the nodes can be easily grouped into sets of nodes so that each set of nodes is densely connected internally. The nodes in the same community play similar roles and share common properties within the network. So the community structures reflect the property of cohesion between nodes. Identifying the community structures of the network is essentially a process that extracts the intimacies hidden between nodes. In bipartite networks, the edges represent the relationships between the different type of nodes. The intimacies of nodes in one mode are identified only by the intersections of nodes in another mode. The IER discussed above reflects the intimacies between nodes. In order to identify the community structures of a bipartite network more accurately, we combine the manner of information flows of the bipartite network to give the expressions of information exchange functions.

We still take the bottom nodes as an example. Because the community structures reflect the property of cohesion between nodes, we hold a point of view that more information exchange opportunities lead to more closer ties. When we measure the intimacies between the bottom nodes, we should consider the relative sizes of the top nodes connected to them. For example, in a scientific collaboration network if two authors are involved in the creation of an article which has two or three authors, the probability that the two authors know each other and communicate with each other is reasonably high. This article should be defined as a high weight. On the contrary if two authors are involved in the creation of a paper which tens of authors co-write, the paper should be given a low weight. A simple way of dealing with this trouble is to weight events inversely by their sizes.[8,36] Here we need to do a little modification. Back to Eq. (8), in order to eliminate the offset caused by summation, we select $f_{v_l}(x_i) = x_i/d^2(v_l)$ as the information exchange function at the first step, which represents the information that the bottom node u_i flows to the top node v_l. At the second step, we select $f_{u_i}(y_l) = y_l/d(u_i)$ as the information exchange function from the top node v_l to the bottom node u_i. Substituting $f_{v_l}(x_i) = x_i/d^2(v_l)$ and $f_{u_i}(y_l) = y_l/d(u_i)$ into Eq. (8), we obtain the IER between nodes u_i and u_j

$$\begin{aligned} IERCP_U(i,j) &= \sum_{l=1}^{n} a_{il} a_{jl} \sum_{k=1}^{m} a_{kl} \frac{1}{d^2(v_l)} \frac{1/d(u_i) + 1/d(u_j)}{2} \\ &= \sum_{l=1}^{n} a_{il} a_{jl} \frac{1}{d(v_l)} \frac{1/d(u_i) + 1/d(u_j)}{2}. \end{aligned} \tag{9}$$

We call Eq. (9) the IERCP.

Interestingly, the WRAP is

$$WRAP_U(i,j) = \frac{1}{d(u_j)} \sum_{l=1}^{n} a_{il} a_{jl} \frac{1}{d(v_l)}. \tag{10}$$

It is not difficult to find that the WRAP is directional. This projection is certified to extract more hidden information from bipartite networks.[33] But it does not reflect the characteristics of the community, and is not suitable for detecting one-mode communities in bipartite networks (to see the following experimental results). The IERCP is non-directional, which not only merges the advantages of the WRAP, but also reflects the characteristics of the community that the relationships between nodes are undirected.

In addition, if we select $f_{v_l}(x_i) = x_i/d(v_l)$ and $f_{u_i}(y_l) = y_l$ as the information exchange functions at the first and the second step respectively, we obtain the IER between the bottom nodes u_i and u_j,

$$SWP_U(i,j) = \sum_{l=1}^{n} a_{il} a_{jl}. \tag{11}$$

This is an SWP of the bipartite network on the bottom set.

Analogously, if we select $f_{v_l}(x_i) = x_i/d(v_l)$ and $f_{u_i}(y_l) = y_l/d(v_l) - 1$ as the information exchange functions at the first step and the second step respectively, we obtain the IER between the bottom nodes u_i and u_j,

$$HWP_U(i,j) = \sum_{l=1}^{n} a_{il} a_{jl} \frac{1}{d(v_l) - 1}. \tag{12}$$

This is an HWP of the bipartite network on the bottom set.

Figure 2 shows four projections of the bipartite network from Fig. 1 on the bottom set. Intuitively, the relationship between nodes u_1 and u_2 is the most intimate, followed by the relationship between nodes u_3 and u_4, and finally the relationship between other nodes. The SWP and HWP do not reflect the facts. Except the weight between nodes u_1 and u_2, the weights of all edges are equal in Figs. 2(a) and 2(b). The IERCP shows that the weight between nodes u_3 and u_4 is the second largest among all weights of edges in Fig. 2(c). As for the WRAP, we obtain a directed network that complicates the relationships between nodes. This example indicates that the IERCP reflects the intimacies between nodes more accurately. Therefore, it is very appropriate to choose the IERCP to identify the community structures in the one-mode network.

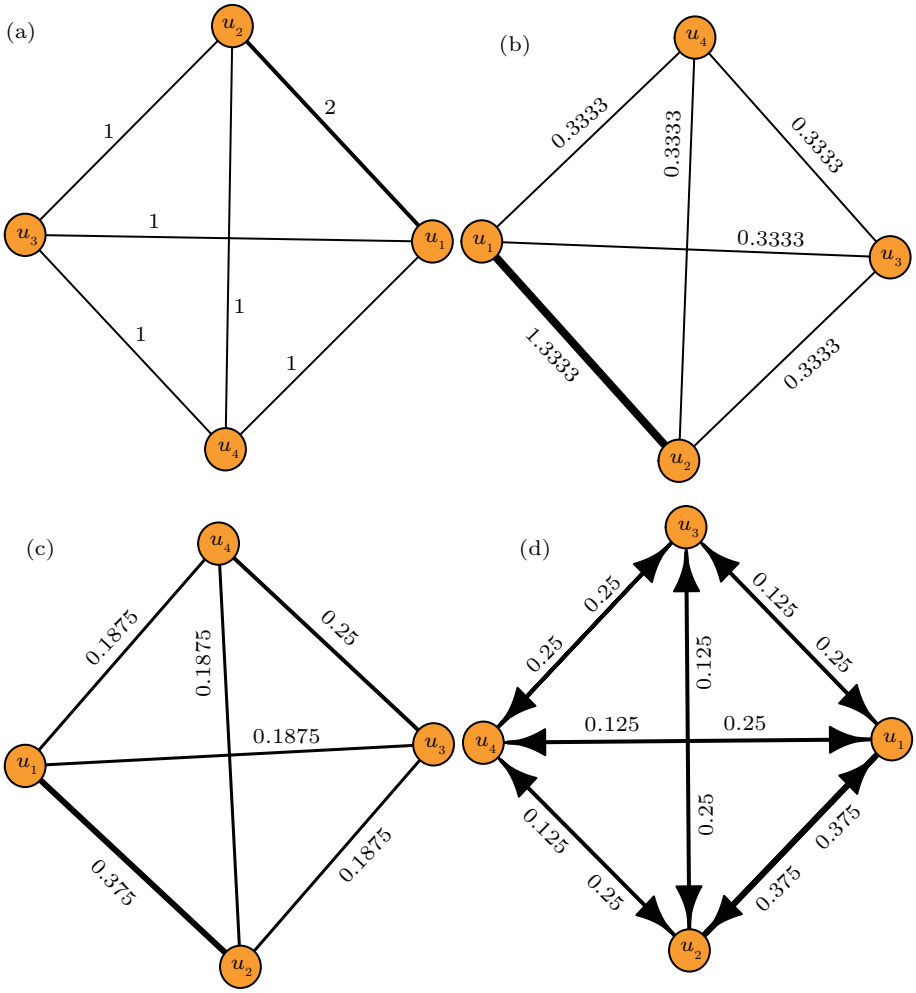


Fig. 2. (color online) Four projection networks on its bottom nodes of bipartite network from Fig. 1. (a) SWP network. (b) HWP network. (c) IERCP network. (d) WRAP network. The numbers in the figure represent the weights of the edges.

In view of the above analysis, we propose the IERCP algorithm for detecting one-mode communities in the bipartite networks. The steps for this method are as follows. At the first step, we use the IERCP to project the bipartite network into bottom nodes, and obtain a one-mode network. At the second step, we utilize the standard community detection algorithm of the one-mode network to detect the community structures. Because the Infomap algorithm is the best-performed algorithm for clustering one-mode networks, especially for very large networks.[39] We choose the Infomap algorithm to detect the community structures of one-mode network.

3. Experimental results

To evaluate the performance of the proposed algorithm, synthetic and real-world networks are applied to the IERCP algorithm.

3.1. Synthetic networks

Unlike the one-mode networks, standard benchmark graphs do not exist for testing the community structures of bipartite networks. An available model of bipartite networks with built-in communities was proposed by Guimera.[18] Guimera partitioned the bipartite network into k groups $C_1, C_2, \ldots, C_k$, where each group C_i was composed of the subgroups C_i^B and C_i^{T} $(i = 1, 2, \ldots, k)$ which only included the bottom nodes and top nodes, respectively. The cluster parameter μ represents a ratio of the sum of edges within each group to the total number of edges in the bipartite network. Obviously, for $\mu = 1$, there is no edge between groups, and the bipartite network has perfect community structures. Conversely, for $\mu = 0$, the bipartite network becomes a random network without community structures. Guimera placed the same number of nodes in each group, which is inconsistent with the real-world networks. We relax this restriction and allow the number of nodes within each group to be different. In addition, we define the probability p for drawing an edge across the two subgroups C_i^B and C_i^{T} $(i = 1, 2, \ldots, k)$. Figure 3 shows a synthetic bipartite network, which contains 64 bottom nodes and 64 top nodes with 4 groups, cluster parameter $\mu = 0.9$ and probability $p = 0.5$.

In our experiments, we choose synthetic networks with 128 nodes and 4 groups to test the proposed algorithm. Each synthetic network contains 64 bottom nodes and 64 top nodes, with the probability $p = 0.5$ for drawing an edge in the same group. The cluster parameter μ is tuned at different values in the range $[0, 1]$. Each synthetic network constructed with given cluster parameter μ is analyzed by using five projection methods, i.e. SWP, HWP, CTP, WRAP, and IERCP. Because the merging of the maximal complete sub-graphs depends on the weighted clustering threshold in the CTP algorithm and the choice of threshold is subjective, we use the GN algorithm[38] instead of the second step of the CTP algorithm.[35] Then, the discovered partitions are compared with the planted partitions by using the normalized mutual information (NMI)[50] and are calculated for the modularity defined in Newman's paper.[38] Figures 4 and 5 show the average NMI and modularity results of 1000 times, respectively.

In theory, the average NMIs are equal to 1 when the cluster parameter is $\mu = 1$. Conversely, the average NMIs are equal to 0 when the cluster parameter is $\mu = 0$. As is shown in Fig. 4, the average NMIs by the SWP, HWP, CTP, and IERCP algorithms are perfectly synchronized with the cluster parameters. The WRAP method does not reflect the variations between the average NMIs and the cluster parameters. So the WRAP method is not suitable for capturing the one-mode communities in bipartite networks. The average NMIs of the SWP, HWP, and IERCP algorithms are greater than those of the CTP method when the cluster parameter is $\mu > 0.6$. This means that the performances of the SWP, HWP, and IERCP algorithms are superior to that of the CTP method when the one-mode networks in bipartite networks have clear community structures. In addition, the IERCP method performs better than the SWP and HWP algorithms for all values of the cluster parameter μ. But there are no differences between the SWP and HWP algorithms. One reason for these phenomena is that the weights between the two nodes in the SWP method represent the number of the nodes jointly connected with each other. The HWP method normalizes the number of the nodes jointly connected with each other in the second step. However, the IERCP method normalizes the number of the nodes jointly connected with each other in each step. The normalized weights are considered as an indicator, and are consistent with the nature of the community which reflects the intimacy of the relationship between nodes.

In Fig. 5, the average modularities of the three projection methods rigorously increase when the cluster parameter $\mu > 0.1$. However, the average modularities of the WRAP algorithm are between 0 and 0.05 for all values of the cluster parameter μ, which are obviously inconsistent with the community structures of the bipartite networks. A reasonable explanation for this observation is that the WRAP algorithm overly partitions the one-mode network and generates more groups than the network actually owns. Besides that, we also find that the average modularities of the IERCP algorithm are larger than those of the SWP, HWP, and CTP methods for all values of $\mu \geq 0.6$. This indicates that the inner connections of communities detected by the IERCP algorithm are denser and community structures detected by the IERCP algorithm are more reasonable.

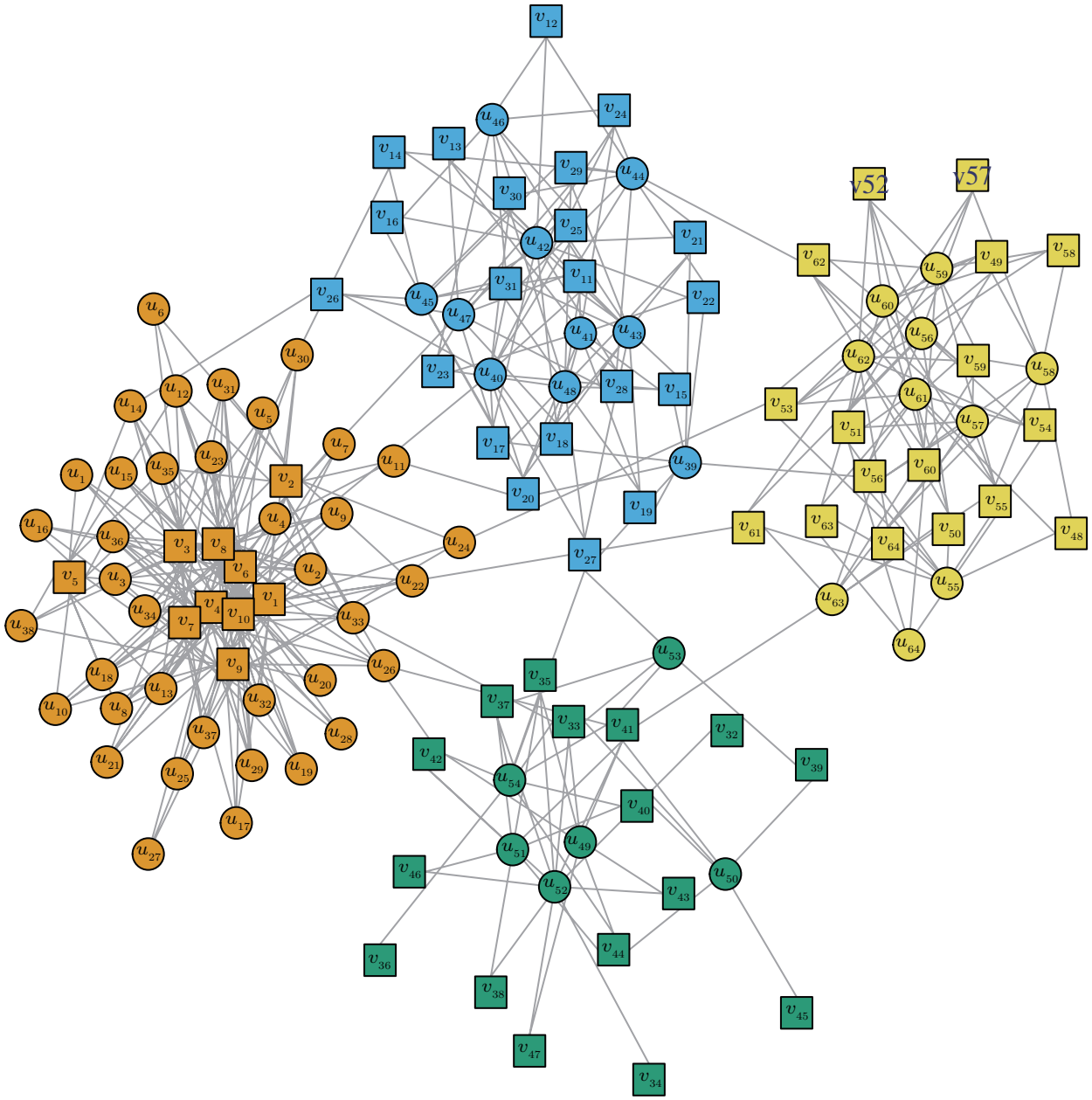


Fig. 3. (color online) Synthetic bipartite network with four groups. Circles represent the bottom nodes and squares represent the top nodes. Different colors represent different groups.

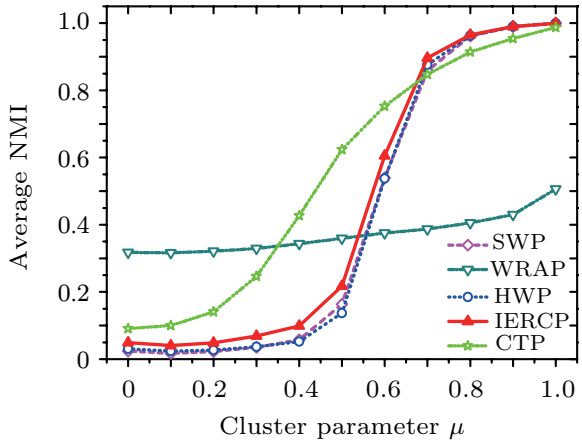


Fig. 4. (color online) The NMI performances for four algorithms on synthetic networks.

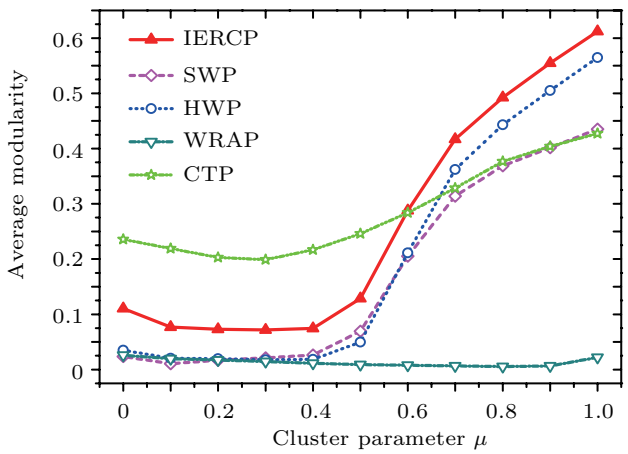


Fig. 5. (color online) The modularity performances for four algorithms on synthetic networks.

3.2. Real-world networks

3.2.1. Constraint-based recommender systems network

The dataset of the constraint-based recommender systems network came from the domain of premium cigars and was collected by an online cigar store in Switzerland from October 2005 to May 2009. It was compiled by Zanker *et al*. as a dataset of a constraint-based recommendation study case.[10] The dataset contains 535 distinct sessions where users disclosed their preferences in an online dialogue and implicitly rated at least one item by an addition to basket action or by accessing the item's details page at least twice. This dataset is compiled into a bipartite network. The bipartite network contains 676 nodes (535 users and 141 items) and 1422 edges. Each edge represents a user's preference.

We are interested in whether users are grouped, which will help us recommend items for new users. The bipartite network is projected into the user nodes and a weighted one-mode network is acquired for users. Clustering the user one-mode network using the IERCP algorithm results in 31 groups with modularity 0.4768. Table 1 shows the results that five projection methods identify communities of user one-mode network. The WRAP algorithm partitions the user one-mode network into 342 groups with modularity 0.0017. However, the network contains only 535 nodes in total. Hence the WRAP algorithm overly partitions the user one-mode network. The CTP algorithm partitions the user one-mode network into only 8 groups. The IERCP algorithm captures the most communities with the highest modularity.

In order to examin which community structures captured by the algorithms above are more reasonable, we use the weighted conductance proposed by Leskovec *et al*.[52] to evaluate the quality of clusters. The weighted conductance is a ratio between the sum of external strength and the total strength of the cluster. It is defined as follows:

$$C_{w,c} = \frac{w_c^{\text{ext}}}{w_c}. \tag{13}$$

In Eq. (13), the external strength is $w_c^{\text{ext}} = \sum_{i\in c, j\notin c} W_{ij}$ and the total strength is $w_c = \sum_{i\in c,j} W_{ij}$, where c is a subgraph or cluster and W_{ij} is an element of the weight matrix of one-mode network. As the ratio is insensitive to the size of the clusters, people use it to compare the quality of clusters of different sizes.[53] For the user one-mode network, the weighted conductance of each cluster captured by the four projection algorithms is calculated respectively. For these clusters containing the same number of nodes in each projection algorithm, we calculate their average weighted conductances. Furthermore, we exclude these clusters that their weighted conductances are equal to zero. Because the community structures detected by the WRAP algorithm has been demonstrated to be unreasonable, the weighted conductances of its clusters are not calculated. All the calculated results are shown in Fig. 6. As is shown in Fig. 6(a), the HWP and SWP algorithms capture a cluster of more than 120 nodes, respectively. The clusters detected by the CTP algorithms are always comprised of many nodes. By and large, the sizes of clusters detected by the IERCP algorithm are relatively reasonable, and its weighted conductances are lower than those of other algorithms. Therefore, it indicates that the result identified by the IERCP algorithm may be more reasonable.

Table 1. Comparisons of algorithm performances on the user and author one-mode networks.

Algorithm	User one-mode network		Author one-mode network	
	Communities	Modularity	Communities	Modularity
SWP	19	0.3192	1792	0.7971
HWP	25	0.4128	1900	0.7972
CTP	8	0.3738	803	0.8736
IERCP	31	0.4768	2151	0.8552
WRAP	342	0.0017	16614	0.0003

3.2.2. Scientific collaboration network

The scientific collaboration network contains authorship links between authors and publications in the arXiv condensed matter section from 1995 to 1999.[51] Initially this network contains 16726 authors and 22015 publications. The number of edges is 58595, and each edge represents an authorship which connects a paper and an author. Taking into account that some nodes do not provide important information for the overall network topology, we remove paper nodes with degree $k \leq 1$ firstly. After a new scientific collaboration network is obtained, author nodes with degree $k < 1$ are removed again. Finally, we obtain a scientific collaboration network which contains 34101 nodes (16264 authors and 17837 papers) and 54417 edges.

We project the scientific collaboration network to the author nodes and obtain an author one-mode network. Clustering the author one-mode network by using the IERCP algorithm results in 2151 groups of authors with modularity 0.8552. The CTP algorithm obtains 803 groups with modularity 0.8736. The WRAP algorithm partitions the author one-mode network into 16614 groups with modularity 0.0003. Based on the view of modularity maximization, it seems that the partition for the author one-mode network by the CTP algorithm is optimal. Analogous to the user one-mode network, we also draw the weighted conductance curves of clusters about the author one-mode network for the four projection algorithms. As shown in Fig. 6(b), the weighted conductances of the CTP algorithm are

indeed lower than those of the other three algorithms. However, it is easily observed that most of the clusters captured by the CTP algorithm contain more than 100 nodes. This result is obviously inconsistent with the fact that community size meets the power-law distribution.[54] Except for the CTP algorithm, the weighted conductance of the IERCP algorithm is the smallest among three algorithms.

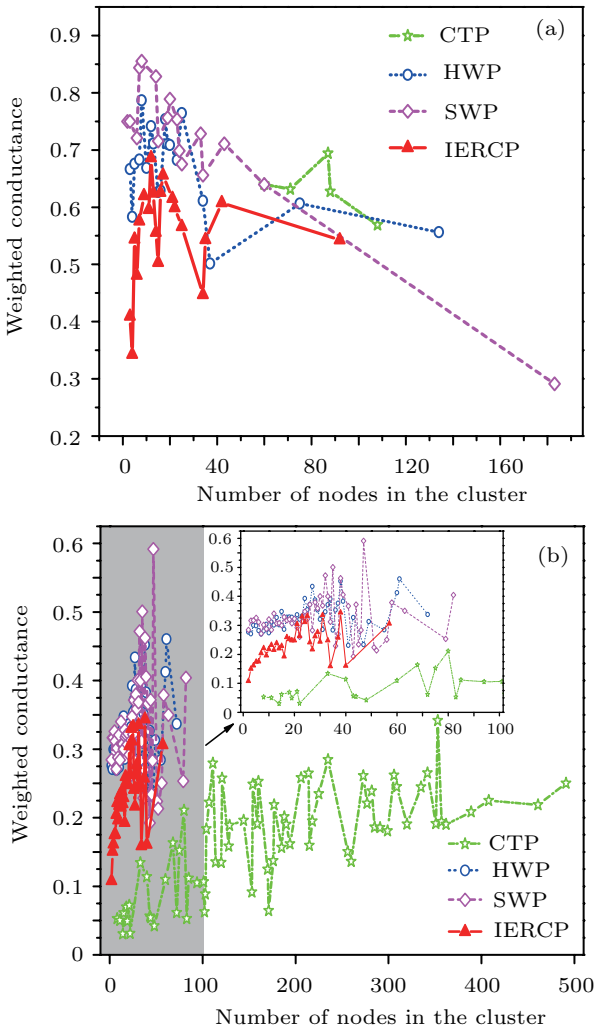


Fig. 6. (color online) The weighted conductances of clusters: (a) the user one-mode network, (b) the author one-mode network.

In order to further explore whether clusters captured by our method are reasonable, we plot the degree distribution of the author one-mode network and the distribution of community size identified by our method. In Fig. 7, nodes with few degrees account for a large proportion and there are few nodes with degree > 60 in the author one-mode network. The distribution of community size is analogous to that of node degree. Most of the communities contain a few nodes. There are 380 communities with only two nodes, accounting for 0.1763. The largest community contains 57 nodes, accounting for 0.0005. Therefore we make a conclusion that there is a well-defined hierarchical structure in this network and the communities of the network are well captured by the IERCP algorithm.

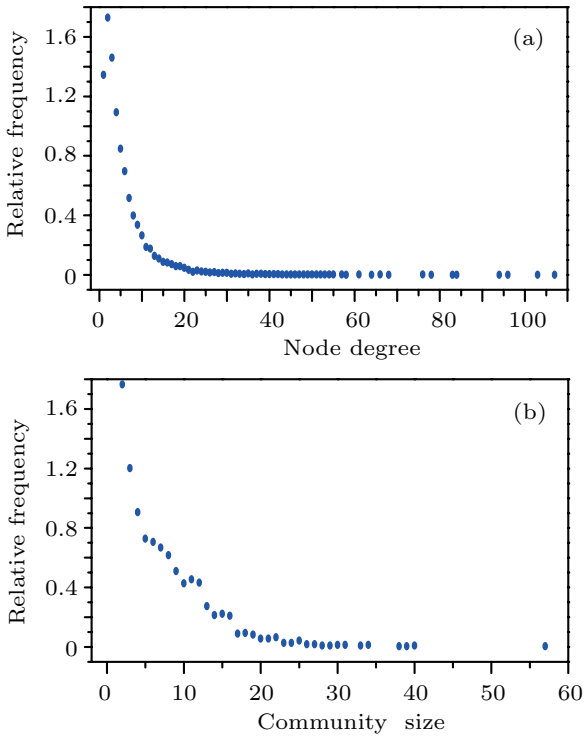


Fig. 7. (color online) The author one-mode network: (a) node degree distribution, relative to network size, (b) community size distribution, relative to total number of community.

4. Conclusions and discussions

In the present paper, we have reviewed briefly all kinds of projections of bipartite networks. A uniform framework of projection is proposed, in which different selection of the information exchange function can extract different information from a bipartite network such as community structure and nestedness. In addition, the model provides a new view to understand the projection of bipartite networks and facilitates theoretical analysis of projection of bipartite networks.

As a typical application, we derive the IERCP projection for detecting community structures of a one-mode network in bipartite networks. Its advantage is multi-fold. First, extensive experiments show that the IERCP algorithm can capture the community structures of one-mode networks from bipartite networks. The performance is significantly higher compared with the other methods. Second, it can extract more information hidden in bipartite networks. Third, it is particularly suited to the situation that people are only interested in community structures of one-mode network in bipartite networks. Because additional community detection algorithms do not need to be designed and we may select different community detection algorithms for different bipartite networks, even for different types of nodes in the same bipartite network. For example, in some affiliation networks, a set representing people may entail overlapping communities, while another set entails hierarchical communities. We can choose different community algorithms for different sets.

The IERCP can be extended straightforwardly to identify community structures of bipartite networks (to see Appendix A), however the size of a network should be large enough to guarantee an acceptable performance. For example, if we select events as the preferred nodes in the Davis southern women network, the algorithm does not identify community structures of the network. This will be our focus and direction for our next research.

Appendix A: An extension of the IERCP algorithm

In fact, the IERCP algorithm can also be used for detecting community structures of bipartite networks. We just need to do a simple extension for this algorithm. This method is particularly suitable for the situation that one of the sets in bipartite networks is preferred for detecting community structures. For example, we pay more attention to the clusters of authors when we detect community structures of the scientific collaboration network. Taking the bottom nodes of bipartite networks as an example, we design an algorithm that identifies the community structures of bipartite networks. This algorithm consists of two steps. At first, we select the bottom nodes as the preferred nodes and capture the clusters of the bottom nodes by the IERCP algorithm. At the second step, each node in the top set is allocated to one of these clusters by the conductance Eq. (13). In this way we obtain the community structures of bipartite networks.

We apply this algorithm to three real-world networks and evaluate its performances. The three real-world networks are the Davis southern women network,[36] the diseases-genes network,[55] and the protein complex-drug network,[56] respectively. The performance criterions we adopt are Barber modularity[17] and density-based modularity.[57] Similar to Newman modularity, the bigger the modularity value is, the better the algorithm performs. For the Davis southern women network, our algorithm partitions it into 2 groups with priority to process the women nodes. The result of the algorithm is shown in Fig. A1.

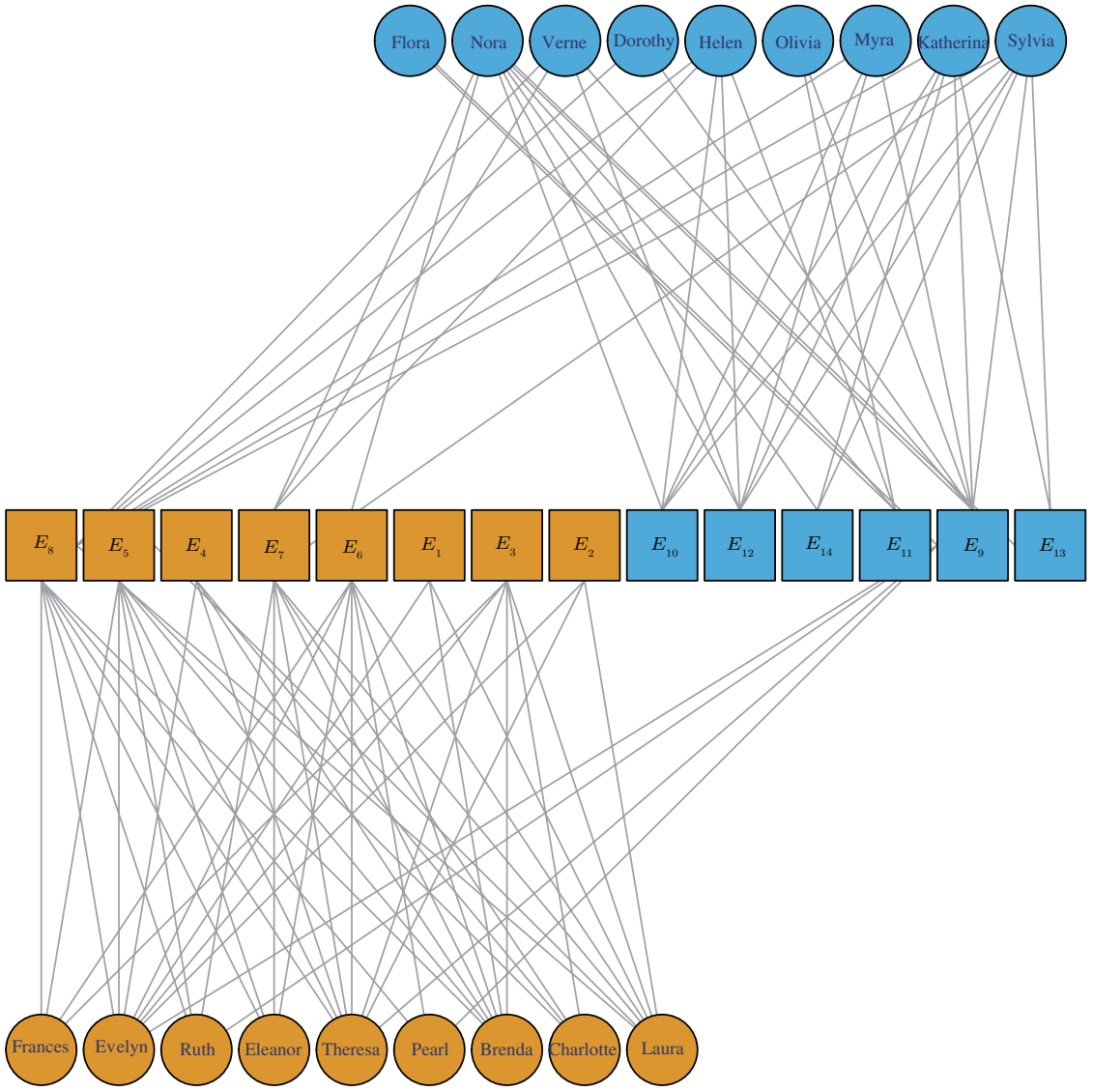


Fig. A1. (color online) Community partitioning scheme of the Davis southern women. Circles represent women and squares represent events. Different colors represent different clusters.

In the Ref. [58], the Davis southern women network is also partitioned into 2 groups. But three women nodes are overlapping nodes. The asymmetric intimacy algorithm[27] partitions the Davis southern women network into 2 groups with priority to process the women nodes. If we give priority to the nodes of events, this network is partitioned into 1 group, which coincides with the result captured by the asymmetric intimacy algorithm. As for the diseases-genes network and the protein complex-drug network, their results captured by our algorithm are shown in Table A1. The results of the asymmetric intimacy algorithm are directly quoted from the Ref. [27]. As is shown in Table A1, the performances of our algorithm are almost the same as those of the asymmetric intimacy algorithm in the two networks.

Table A1. Comparison of algorithm performances on the diseases-genes and protein complex-drug networks.

Network	Order of priority	Asymetric intimacy algorithm			Our algorithm		
		N[a]	Q_B^b	Q_D^c	N	Q_B	Q_D
Diseases-genes	Disease gene nodes	3	0.388	1.068	3	0.387	1.032
	Genetic disease nodes	2	0.379	0.534	2	0.384	0.497
Protein complex-drug	Complex nodes	67	0.828	32.589	72	0.841	39.787
	Drug nodes	63	0.819	31.268	71	0.838	37.312

[a] The number of identified clusters. [b] Barber modularity. [c] Density-based modularity.

References

[1] Chen Y Z, Fu C H, Chang H, Li N and He D R 2008 *Chin. Phys. B* **17** 3580
[2] Feng S M, Hu B Y, Nie C, Shen X H and Ci Y S 2016 *Chin. Phys. B* **25** 030504
[3] Jones P L and Agrawal A A 2017 *Ann. Rev. Entomol.* **62** 53
[4] Stout J C and Tiedeken E J 2016 *Funct. Ecol.* **31** 38
[5] Flores C O, Valverde S and Weitz J S 2013 *ISME J.* **7** 520
[6] Weitz J S, Poisot T, Meyer J R, Flores C O, Valverde S, Sullivan M B and Hochberg M E 2013 *Trends Microbiol.* **21** 82
[7] Newman M E J 2001 *Phys. Rev. E* **64** 01631
[8] Newman M E J 2001 *Phys. Rev. E* **64** 01632
[9] Zanker M and Jessenitschnig M 2009 *User Model. User-adapt. Interact.* **19** 133
[10] Zanker M, Jessenitschnig M and Schmid W 2010 *Constraints* **15** 574
[11] Chen G, Qiu T and Shen X Q 2015 *Chin. Phys. B* **24** 078901
[12] Guillaume J L and Latapy M 2006 *Phys. A: Stat. Mech. Its Appl.* **371** 795
[13] Flores C O, Poisot T, Valverde S and Weitz J S 2016 *Methods Ecol. Evol.* **7** 127
[14] Bashan A, Bartsch R P, Kantelhardt J W, Havlin S and Ivanov P C 2012 *Nat. Commun.* **3** 702
[15] Reddy P K, Kitsuregawa M 2001 *Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on* **1** 301
[16] Eustace J, Wang X Y, Li J Q 2014 *Knowledge-Based Systems* **70** 118
[17] Barber M J 2007 *Phys. Rev. E* **76** 066102
[18] Guimerà R, Sales-Pardo M and Amaral L A N 2007 *Phys. Rev. E* **76** 036102
[19] Barber M J and Clark J W 2009 *Phys. Rev. E* **80** 026129
[20] Liu X and Murata T 2010 *J. Adv. Comput. Intell. Intell. Informatics.* **14** 408
[21] Peixoto T P 2013 *Phys. Rev. Lett.* **14** 148701
[22] Larremore D B, Clauset A and Jacobs A Z 2014 *Phys. Rev. E* **90** 012805
[23] Kheirkhahzadeh M, Lancichinetti A and Rosvall M 2016 *Phys. Rev. E* **93** 032309
[24] Liebig J and Rao A 2016 *EPL* **113** 28003
[25] Li Z, Wang R S, Zhang S and Zhang X S 2016 *Inf. Sci. (Ny).* **367-368** 874
[26] Aitkin M, Vu D and Francis B 2014 *Soc. Networks.* **38** 74
[27] Wang X Y, Qin X M 2016 *Phys. A Stat. Mech. its Appl.* **462** 569
[28] Barabàsi A L, Jeong H, Néda Z, Ravasz E, Schubert A and Vicsek T 2002 *Phys. A Stat. Mech. its Appl.* **311** 590
[29] Zhou T, Wang B H, Jin Y D, He D R, Zhang P P, He Y, Su B B, Chen K, Zhang Z Z and Liu J G 2007 *Int. J. Mod. Phys. C.* **18** 297
[30] Fan Y, Li M, Zhang P, Wu J and Di Z 2007 *Phys. A Stat. Mech. its Appl.* **378** 583
[31] Alzahrani T and Horadam K J 2016 *Complex Systems and Networks* 1st edn. (Berlin: Spring Link) pp.25-50
[32] Melamed D 2014 *PLoS ONE* **9** e97823
[33] Zhou T, Ren J, Medo M and Zhang Y C 2007 *Phys. Rev. E* **76** 046115
[34] Wang Y L, Zhou T, Shi J J, Wang J and He D R 2009 *Phys. A Stat. Mech. its Appl.* **388** 2949
[35] Cui Y z, Wang X Y 2016 *Phys. A Stat. Mech. its Appl.* **457** 307
[36] Borgatti S P and Halgin D S 2011 *The Sage Handbook of Social Network Analysis* 1st edn (Thousand Oaks: SAGE) pp.417-433
[37] Qiao J, Meng Y Y, Chen H, Huang H Q and Li G Y 2016 *Phys. A Stat. Mech. its Appl.* **457** 270
[38] Newman M E J and Girvan M 2004 *Phys. Rev. E* **69** 026113
[39] Rosvall M and Bergstrom C T 2008 *Proc. Natl. Acad. Sci.* **105** 1118
[40] Palla G, Derényi I, Farkas I, Vicsek T 2005 *Nature* **435** 03607
[41] Li J Q, Wang X Y, Eustace J 2013 *Phys. A Stat. Mech. its Appl.* **392** 6125
[42] Cui Y Z, Wang X Y,Li J Q 2014 *Phys. A Stat. Mech. its Appl.* **405** 85
[43] Li J Q, Wang X Y, Cui Y Z 2014 *Phys. A Stat. Mech. its Appl.* **415** 398
[44] Eustace J, Wang X Y, Cui Y Z 2015 *Phys. A Stat. Mech. its Appl.* **421** 510
[45] Wang X Y, Li J Q 2013 *Phys. A Stat. Mech. its Appl.* **392** 2555
[46] Cui Y Z, Wang X Y, Eustace J 2014 *Phys. A Stat. Mech. its Appl.* **416** 198
[47] Long B, Wu X Y, Yu P S, Zhang Z F 2007 *Seventh IEEE International Conference on Data Mining (ICDM 2007), October 28-31,2007, Omaha,NE,USA*, p. 232
[48] Kim Y D, Son S W, Jeong H 2010 *Phys. Rev. E* **81** 016103
[49] Haythornthwaite C 1996 *Lisr* **18** 323
[50] Lancichinetti A and Fortunato S 2014 *Phys. Rev. E* **89** 049902
[51] Newman M E J 2001 *Proc. Natl. Acad. Sci.* **98** 404
[52] Leskovec J, Lang K J, Dasgupta A, Mahoney M W 2009 *Internet Math.* **6** 29
[53] Fortunato S, Hric D, 2016 *physics reports* **659** 1
[54] Scott J 2017 *Social network analysis*, 4th edn. (Thousand Oaks: SAGE) pp. 113–137
[55] Goh K, Cusick M E, Valle D, Childs B, Vidal M, Barabási A L 2007 *Proc. Natl. Acad. Sci.* **104** 8685
[56] Nacher J C, Schwartz J M 2012 *PLoS ONE* **7** e30028
[57] Xu Y C, Chen L, Li B, Liu W 2015 *Inf. Sci. (Ny).* **317** 278
[58] Cui Y Z, Wang X Y 2014 *Phys. A Stat. Mech. its Appl.* **407** 7