Hindawi
Scientific Programming
Volume 2019, Article ID 4310561, 12 pages
https://doi.org/10.1155/2019/4310561

Research Article

New Community Estimation Method in Bipartite Networks Based on Quality of Filtering Coefficient

Li Xiong,[1] Guo-Zheng Wang,[1] and Hu-Chen Liu[2]

[1] *School of Management, Shanghai University, Shanghai 200444, China*
[2] *College of Economics and Management, China Jiliang University, Hangzhou, Zhejiang 310018, China*

Correspondence should be addressed to Hu-Chen Liu; huchenliu@foxmail.com

Received 20 January 2019; Revised 13 April 2019; Accepted 24 April 2019; Published 21 May 2019

Academic Editor: Emiliano Tramontana

Community detection is an important task in network analysis, in which we aim to find a network partitioning that groups together vertices with similar community-level connectivity patterns. Bipartite networks are a common type of network in which there are two types of vertices, and only vertices of different types can be connected. While there are a range of powerful and flexible methods for dividing a bipartite network into a specified number of communities, it is an open question how to determine exactly how many communities one should use, and estimating the numbers of pure-type communities in a bipartite network has not been completed. In our paper, we propose a method named as "biCNEQ" (bipartite network communities number estimation based on quality of filtering coefficient), which ensures that communities are all pure type, for estimating the number of communities in a bipartite network. This paper makes the following contributions: (1) we show how a unipartite weighted network, which we call similarity network, can be projected from a bipartite network using a measure of correlation; (2) we reveal the relation between the similarity correlation and community's edges in the vertices of a unipartite network; (3) we design a measure of the filtering quality named QFC (quality of filtering coefficient) to filter the similarity network and construct a binary network, which we call approximation network; and (4) the number of communities in each type of unipartite networks is estimated using Riolo's method with the approximation network as input. Finally, the proposed biCNEQ is demonstrated by both synthetic bipartite networks and a real-world network, and the results show that it can determine the correct number of communities and perform better than two classical one-mode projection methods.

1. Introduction

The bipartite network is a network whose vertices can be divided into two types *a* and *b*, where every edge connects a vertex of type-*a* to one of type-*b*, and there are no edges connecting vertices of the same type. There are many examples of bipartite networks, such as those described in [1–3]. Regarding unipartite networks, a common task is to find groups or communities of vertices that connect to the rest of the network in similar ways. Finding this underlying group structure is of significant, which can, for example, divide a heterogeneous network into homogeneous subgraphs for subsequent analysis or modeling [4].

Beginning from Newman's [5] study, community detection has attracted considerable attention from researchers [6], aiming to identify good ways to divide up a network into communities. A range of powerful and flexible methods for dividing a bipartite network into a specified number of communities have been proposed in recent years [4, 7, 8]. However, most of them have one key shortcoming; that is, they require us to know the number of communities of a network in advance. In the real world, however, we usually do not know this number a priori, and thus, we need to estimate it from the data. Recently, several methods have been proposed for making such estimates for unipartite networks [9–12] and bipartite networks [13–16]. Barber [13] in his work introduced bipartite modularity, a variant of the modularity proposed by Newman and Girvan [17]. A dual-projection approach proposed by Han et al. [14] aims to maximize the

Newman's one-mode modularity. The authors of [15, 16] maximized Barber's bipartite modularity for bipartite community detection. However, maximizing both modularities noted above proved to be a NP-hard problem [6, 18]. The bipartite network communities generated in the previous studies are of mixed type, and so far, there is no exploration inferring to the numbers of pure-type communities in a bipartite network.

In our paper, we propose a method named "biCNEQ" (bipartite network communities number estimating based on quality of filtering coefficient), which ensures that communities are all pure type. The main innovations and contributions of this study can be illuminated as follows: (1) a percolation idea-based (PIB) method, proposed by Lambiotte and Ausloos [19], is used to project a bipartite network to unipartite correlation networks and reveal the emergence of social communities and music genres by filtering correlation matrices and (2) a first principles method given by Riolo et al. [11] is used for inferring the number of communities in a unipartite network. The quality of filtering coefficient (QFC) is designed to select a threshold to filter the correlation matrix in constructing a binary unipartite network. This method can roughly match the structural features of the correlation and degree of the vertices of the original ones, which cannot be done using PIB. Finally, we use Riolo et al.'s [11] method to estimate the number of communities in each type of unipartite networks. In addition, the proposed biCNEQ is demonstrated by both synthetic bipartite networks and a real-world network, and the results show that our method performs better than two classical one-mode projection methods.

2. Methods

Tests were performed on both synthetic bipartite networks and a real-world bipartite network with a known community structure.

2.1. Synthetic Networks. We construct a synthetic network based on a degree-corrected bipartite stochastic block model (biSBM) formulated by Larremore et al. [4]. Given a bipartite network G with $N \times N$ adjacency matrix A (where $N = N_a + N_b$ and N_a are the vertices of type-a), we divide the N_a vertices of type a into K_a groups and the N_b type-b vertices into K_b groups and express the matrix of group interrelationships as a $K \times K$ matrix, where $K = K_a + K_b$. Let vertex i of type t_i belongs to group g_i and T_r be the type of group r, imposing the constraint $t_i = T_{g_i}$, which indicates that vertex types and group types must match and ensures that groups will be pure type. Let the number of edges between vertices i and j follow a Poisson distribution with mean $\theta_i \theta_j \omega_{g_i g_j}$ and choose the normalization $\sum_i \theta_i \delta_{g_i,r} = 1$, where θ_i controls the expected degree of vertex i, ω_{rs} is a $K \times K$ symmetric matrix of parameters to control the number of edges between groups r and s, and δ is the Kronecker delta. The probability of observing a network G with adjacency matrix A can be written as

$$\begin{aligned} P(G \mid g, \theta, \omega, T) &= \prod_{\substack{i<j \\ t_i \neq t_j}} \frac{\left(\theta_i \theta_j \omega_{g_i g_j}\right)^{A_{ij}}}{A_{ij}!} \exp\left(-\theta_i \theta_j \omega_{g_i g_j}\right) \\ &= \frac{\prod_i \theta_i^{k_i}}{\prod_{\substack{i<j \\ t_i \neq t_j}} A_{ij}} \times \prod_{\substack{rs \\ T_r \neq T_s}} \omega_{rs}^{m_{rs}/2} \exp\left(-\frac{1}{2}\omega_{rs}\right), \end{aligned} \tag{1}$$

where k_i is the observed degree of vertex i and $m_{rs} = \sum_{ij} A_{ij} \delta_{g_i,r} \delta_{g_j,s}$ is the number of edges between groups r and s. After taking partial derivatives with respect to ω_{rs} on the logarithm of equation (1), we can get the maximum likelihood parameter as follows:

$$\widehat{\omega}_{rs} = m_{rs}. \tag{2}$$

The maximum likelihood $\widehat{\theta}_i$ can be found via the constrained maximization of the logarithm of equation (1) subject to $\sum_i \theta_i \delta_{g_i,r} = 1$ using Lagrange multipliers, i.e.,

$$\widehat{\theta}_i = \frac{k_i}{\kappa_{g_i}}, \tag{3}$$

where $\kappa_r = \sum_s m_{rs}$ is the sum of the degrees in group r.

Empirically observed networks are often noisy with missing or spurious edges. Therefore, we examine the ability of biCNEQ to analyze a range of synthetic networks generated by a mixed model, which is a combination of planted structure ω^{planted} and a random network model ω^{random}. The later model is used to create various levels of uniformly random noise. We consider two forms, as in [4], an easy and a difficult case, to illustrate the biCNEQ's performance under different conditions.

We specify g and ω^{planted} and create mixed networks using g. Then,

$$\omega = \lambda \omega^{\text{planted}} + (1 - \lambda) \omega^{\text{random}}, \tag{4}$$

where the mixed parameter λ takes values between 0 (all noise) and 1 (all planted structure) and $\omega_{rs}^{\text{planted}} = m_{rs}$ according to equation (2). We let $\omega_{rs}^{\text{random}} = \kappa_r \kappa_s / m$, where m is the total number of edges in the network.

2.1.1. An Easy Case. In the easy case, we define the mixed matrix to have an easily identifiable community structure which consists of four equally sized, unambiguous, and nonoverlapping components with each made up of one type-a and one type-b community. Let $N = 60$ for each type and divide these vertices evenly across the four components, where $m_{1,5} = m_{2,6} = m_{3,7} = m_{4,8} = 150$. The symmetric entry has the same value. We create networks using ω^{random}. Finally, we use the code, downloaded from http://www.danlarremore.com/bipartiteSBM/makeEasyCaseNetworks.m, to generate mixed synthetic networks of an easy case for testing with the specification above and with its degree distribution unchanged.

2.1.2. A Difficult Case. In the difficult case, the mixed matrix we define is given a less easily identifiable community structure by creating partially overlapping communities, $k_a \neq k_b$, and has a broad degree distribution. We set different sizes for the communities with 70 type-a vertices, divided evenly into 2 communities {35, 35}, and 30 type-b vertices, divided into 3 communities {10, 15, 5}. Then, we let $m_{1,3} = m_{2,4} = 250$ and $m_{1,5} = m_{2,5} = 150$; θ_i can be obtained using equation (3). The symmetric entry has the same value. Finally, we use the code, downloaded from http://www.danlarremore.com/bipartiteSBM/makeDifficultCaseNetworks.m, to generate the mixed synthetic network of a difficult case for testing with the specification above and degree distribution unchanged.

2.2. Empirical Networks. The Southern women network collected by Davis et al. [20] contains the observed attendance at 14 social events by 18 Southern women. This network was commonly used as a benchmark for bipartite network community detection algorithms [4, 13, 21, 22], much like the Zachary "karate club" that was used for benchmarking unipartite community detection algorithms.

2.3. Projection Procedure

2.3.1. Projection. Given a bipartite network with an $N_a \times N_b$ bipartite adjacency matrix B, where N_a and N_b are the number of type-a and type-b vertices, respectively. $B_{il} = 1$, if there is an edge between type-a vertex a_i and type-b vertex b_ℓ; otherwise, $B_{il} = 0$.

A common way to represent and study bipartite networks consists of projecting them onto links of one kind of vertex [23]. The standard projection method simplifies the system to a unipartite network. For instance, from a bipartite network of scientists and papers, one can extract a network of scientists only, who are related by coauthorship. However, such a projection loses a lot of information and leads to an oversimplified and less useful representation [6, 19, 24]. Therefore, we refine it in an alternative way below.

We define for each type-a vertex a_i the N_b vector [19]:

$$\sigma^i = (\ldots, 1, \ldots, 0, \ldots, 1, \ldots), \quad i = 1, 2, \ldots, N_a, \tag{5}$$

where $\sigma^i_\ell = B_{i\ell}$ is equal to 1 if there exists one edge between a_i and b_ℓ; otherwise, it is 0. Then, we calculate the correlation between vertices a_i and a_j using the cosine similarity [25], which is a symmetric correlation measure. That is,

$$C_{ij} = \frac{\sigma^i \cdot \sigma^j}{|\sigma^i||\sigma^j|} = \cos\theta_{ij}, \tag{6}$$

where $\sigma^i \cdot \sigma^j$ denotes the scalar product between σ^i and σ^j. Besides,

$$\left|\sigma^i\right| = \sqrt{\sum_{\ell=1}^{N_b} \left(\sigma^i_\ell\right)^2} = \sqrt{k_i}, \tag{7}$$

where k_i is the degree of vertex a_i. This measure of correlation, which corresponds to the cosine of the two vectors in an N_b-dimensional space, is equal to 1 when their entries are strictly identical and vanishes when they have no common entries. Specifically, for each pair of type-a vertices, $C_{ij} = 0$ when a_i and a_j have no common edges with any type-b vertices, and C_{ij} will become 1 when they have identical edges. We call the $N_a \times N_a$ matrix a similarity matrix, with its element C_{ij}, and the unipartite weighted network with similarity matrix C is a similarity network.

In [19], the authors revealed the emergence of social communities and music genres by filtering similarity matrices. However, the threshold of filtering coefficient was selected arbitrarily by the authors, and the community structures they found were not unique. To avoid this issue, we firstly would like to know the relation between the similarity correlation and community membership of vertices. We now make an analysis of relations among edge existence, similarity correlation, and community membership of the unipartite network vertices.

2.3.2. Relation between Similarity Correlation and Community's Edges. According to the definition of a community [6, 26, 27], there are many edges within communities but few edges between communities. Modularity [17, 28] is the most popular function to measure the division quality of a network. Given a particular network with an $N \times N$ adjacency matrix $A = \{A_{ij}\}$, its modularity is defined as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta\left(g_i, g_j\right), \tag{8}$$

where k_i is the degree of vertex i, m is the number of edges, and g_i denotes the community to which vertex i is assigned. The δ function yields 1 if vertices i and j are in the same community $(g_i = g_j)$ and is 0 otherwise. Therefore, each pair of vertices with an edge between them is more likely to be in the same community than in a different community. This is because it will increase the value of Q if they are in the same community but makes no contribution to Q otherwise.

Now, we investigate whether a pair of vertices with a higher similarity correlation is more likely to be in the same community rather than a different community. We let the ith row of A be the ith vertex's N-vector σ^i and use the cosine of the two vectors σ^i and σ^j in the N_b-dimensional space, to quantify the correlation C_{ij} between vertices i and j. It is obvious that there are more coneighbors between a pair of vertices with a higher correlation. Moreover, we have proven in the previous paragraph that two ends of an edge are more likely to be in the same community than in a different community. Therefore, a pair of vertices with a higher correlation is more likely to be in the same community than in a different community.

We test our inference on two widely used unipartite networks, the "karate club" network of Zachary [29] and the network of political blogs assembled by Adamic and Glance [30]. Both the two networks have a known community structure. We define the average similarity correlation of the ith vertex with the vertices in the same community as $C_i^{\text{aver\_s}} = (\sum_{j,g_j=g_i} C_{ij})/n_{g_i}$ and the average similarity correlation of the ith vertex with the vertices of different communities as $C_i^{\text{aver\_d}} = (\sum_{j,g_j\neq g_i} C_{ij})/(n-n_{g_i})$, where n is the total number of vertices and n_{g_i} denotes the vertices number of the community vertex I belongs to. In Figure 1, we plot the average correlation of each of the vertices with the vertices in the same community $C_i^{\text{aver\_s}}$ against those in the different communities $C_i^{\text{aver\_d}}$ for each vertex of the two networks, respectively. Figures 1(a) and 1(b) show that one vertex's average similarity correlation with the vertices in the same community is greater than that in different communities.

Therefore, we form an edge between each pair of vertices when its similarity correlation value is higher than a given value to construct a binary network from the similarity network of the bipartite network. We will discuss how to select such a threshold in the following section.

2.4. Filtering Procedure. To derive such a binary network from the similarity network of a bipartite network, i.e., transform the correlation values in the continuous range [0, 1] to an edge valued 1 or 0 between a pair of vertices, we define a filtering coefficient $\phi \in [0,1]$ as in [19]. We filter the similarity matrix elements using ϕ, so that $C_{ij} \mid \phi = 1$ if $C_{ij} > \phi$ and is equal to 0 otherwise. We call the unweighted unipartite network $C \mid \phi$, obtained by filtering the similarity matrix, a filtering network, whose adjacency matrix $C \mid \phi$, with one element denoted as $C_{ij} \mid \phi$, is named a filtering matrix.

We take the Southern women network as an example and plot the total degree $\text{Degree}(C \mid \phi) = \sum_{ij} C_{ij} \mid \phi$ of the filtering network as a function of the filtering coefficient ϕ on the women similarity network and events similarity network both projected from the Southern women dataset. As shown in Figure 2, the total degree $\text{Degree}(C \mid \phi)$ of the filtering network $C \mid \phi$ reaches a maximum when $\phi = 0$, and the number decreases or remains unchanged with increasing ϕ, reaching a minimum when $\phi = 1$. We find a total degree value of 88, which is nearest to the exact number of 89, at $\phi = 0.614$ on the women filtering network and at $\phi = 0.464$ on the events filtering network.

This raises a question of how do we know when the filtering is good? To answer this, we first introduce the concept of null model. A null model is a random network which matches the original in some of its structural features but does not have any community structure. The most popular null model is known as the standard null model of modularity [17]. It consists of a randomized version of the original graph, where the edges are rewired at random, under the constraint that the expected degree of each vertex matches the degree of the vertices in the original graph. We call the original network the real network, which is assumed to be a unipartite unweighted network projected from a bipartite network.

Let $\overline{A}$ be the $N_a \times N_a$ adjacency matrix of the real network of type-a vertices, then $\sum_j \overline{A}_{ij} = k_i$ and $m = \sum_i k_i$ is the total degree of the real network. Now, we build a null mode R as in [17]:

$$R_{ij} = \frac{k_i \times k_j}{m}. \tag{9}$$

We would like the degree of each vertex of the filtering network to approximate the degree of the vertices in the original graph. Firstly, we define a measure of degree difference between the filtering network and the null model, and we call it degree difference (DD):

$$\text{DD}(\phi) = \sum_{ij} \left(C_{ij} \mid \phi - R_{ij}\right)^2. \tag{10}$$

From equation (7), we know the degree of each vertex of the filtering network approximately matches the degree of the vertices in the original graph when DD is minimized. By taking a derivative with respect to $C_{ij} \mid \phi$ in equation (10) and let it equals to 0, we have

$$\sum_{ij} C_{ij} \mid \phi = \sum_{ij} R_{ij} = m, \tag{11}$$

where $\sum_{ij} C_{ij} \mid \phi$ is the total degree of the filtering network and m is the edge number of the bipartite network. Now, we define a measure of the quality of a filtering network of a bipartite network, which we call QFC based on equation (11):

$$\text{QFC}(\phi) = \left|\sum_{ij} C_{ij} \mid \phi - m\right|, \tag{12}$$

where $\text{QFC}(\phi)$ is actually the absolute value of $\sum_{ij} C_{ij} \mid \phi - m$. That is to say, the best filtering network, whose total degree matches or is closest to that of the original network, occurs when the QFC reaches a minimum with $\phi = \phi^{\text{best}}$. We call the best filtering network the approximation network, whose adjacency matrix $C \mid (\phi = \phi^{\text{best}})$ is named an approximation matrix.

Next, we perform experiments to test QFC criterion on the Southern women network. We plot the $\text{QFC}(\phi)$ of the filtering network as a function of the filtering coefficient ϕ on the type-a similarity network and type-b similarity network projected from the real-world bipartite networks mentioned above. The approximation networks can be obtained when $\text{QFC}(\phi)$ reaches its minimum value at the bottom of the curve, as shown in Figure 3.

In this procedure, we construct the $\text{QFC}(\phi) = \left|\sum_{ij} C_{ij} \mid \phi - m\right|$ as a function of the filtering coefficient

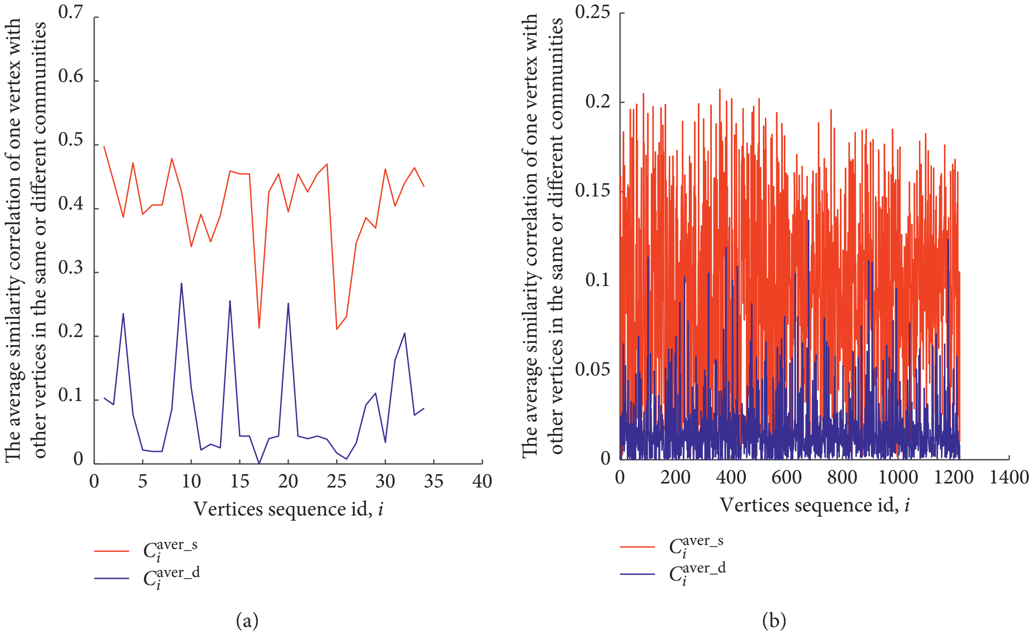


Figure 1: Average correlation value of each of the vertices with vertices in the same community $C_i^{\text{aver\_s}}$ against those in different communities $C_i^{\text{aver\_d}}$. Tests on (a) the "karate club" network with 34 vertices and (b) the "political blogs" network with 1222 vertices.

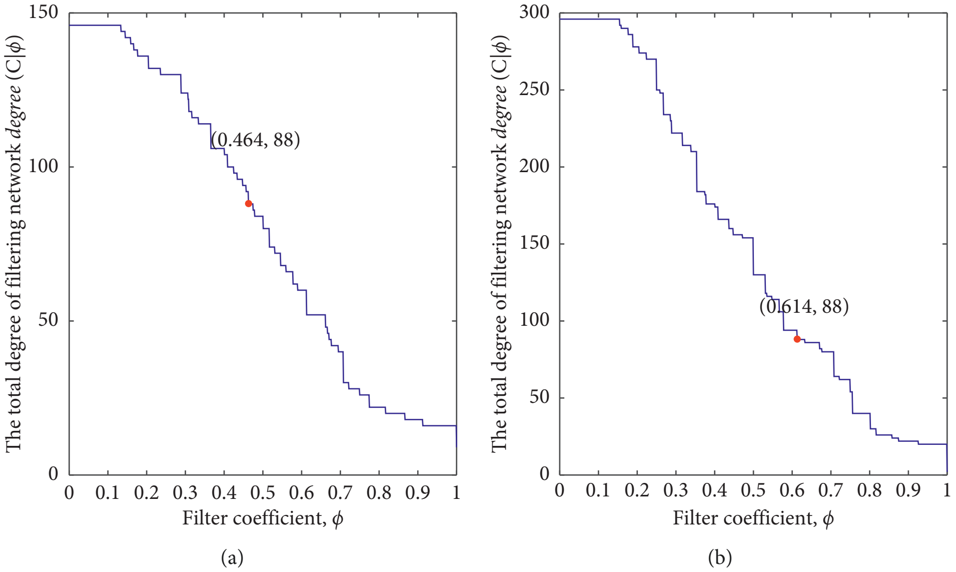
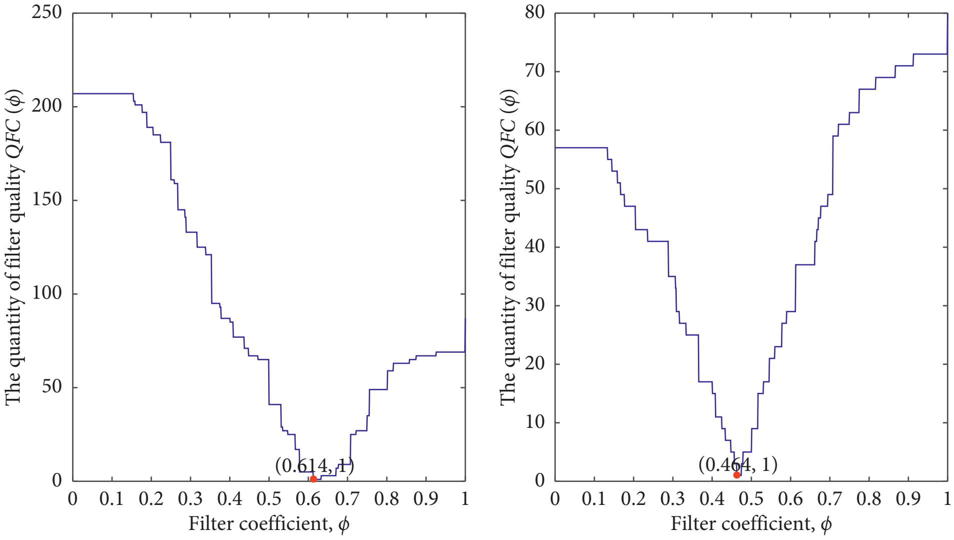


Figure 2: Total degree Degree$(C \mid \phi)$ of filtering network as a function of the filtering coefficient ϕ. (a) Total degree Degree$(C \mid \phi)$ of the events filtering network as a function of ϕ. We can get the best total degree at $\phi = 0.464$. (b) Total degree Degree$(C \mid \phi)$ of the women filtering network as a function of ϕ. We can see that when $\phi = 0.614$ at the red point, and the total degree is 88 with one degree lost, which is nearest to the exact total degree of 89.

$\phi \in [0, 1]$ and find the minimum value of QFC(ϕ) and the corresponding $\phi = \phi^{\text{best}}$. Then, we get the approximation matrix A^{appro} with elements $A_{ij}^{\text{appro}} = C_{ij} \mid (\phi = \phi^{\text{best}})$ of type-a vertices.

2.5. Estimation Procedure. In the work of [11], the authors introduced a method for estimating the number of communities in a unipartite network. We can use this method to determine the number of communities in the approximation network of a bipartite network.

Riolo et al. [11] employed a more sophisticated approach, the degree-corrected stochastic block model, to overcome the shortcomings of the stochastic block model [31], which gives substantially better results for real-world network data. With the model specified, they find the probability P that a particular network with adjacency matrix $A = \{A_{ij}\}$ is found by the following equation:



(a)

(b)

Figure 3: QFC(ϕ) of filtering network as a function of the filtering coefficient ϕ. (a) QFC(ϕ) of the women filtering network. We get the minimum QFC$(\phi) = 1$ when $\phi^{\text{best}} = 0.614$ at the red point. (b) QFC(ϕ) of the events filtering network. We get the minimum QFC$(\phi) = 1$ when $\phi^{\text{best}} = 0.464$ at the red point.

$$\begin{aligned} P(A \mid \omega,\theta,g,k) &= \prod_{i<j}\left(\theta_i\theta_j\omega_{g_ig_j}\right)^{A_{ij}} e^{-\theta_i\theta_j\omega_{g_ig_j}} \\ &\quad \times \prod_i \left(\frac{1}{2}\theta_i^2\omega_{g_ig_i}\right)^{A_{ii}/2} e^{-\left(\theta_i^2\omega_{g_ig_i}/2\right)} \\ &= \prod_i \theta_i^{d_i} \prod_{r<s} \omega_{rs}^{m_{rs}} e^{-n_rn_s\omega_{rs}} \prod_r \omega_{rr}^{m_{rr}} e^{-\left(n_r^2\omega_{rr}/2\right)}, \end{aligned} \tag{13}$$

where k is the number of groups, g_i denotes the group to which vertex i is assigned, r and s are the groups to which the vertices belong to, and m_{rs} is the number of edges running between groups r and s. The parameter θ_i is used to independently control the average degree of each node and hence match any desired distribution. The parameter ω_{rs} is the expected value of the adjacency matrix entry A_{ij} for vertices i and j belonging to groups r and s, respectively, and they control the community structure. The parameters above have been discussed in detail in [31].

After integrating the parameters ω and θ, from equation (13), we have

$$\begin{aligned} P(A \mid g,k) &= \prod_r n_r^{\kappa_r} \frac{(n_r-1)!}{(n_r+\kappa_r-1)!} \times \prod_{r<s} \frac{m_{rs}!}{(\rho n_r n_s + 1)^{m_{rs}+1}} \\ &\quad \cdot \prod_r \frac{m_{rr}!}{((1/2)\rho n_r^2 + 1)^{m_{rr}+1}}, \end{aligned} \tag{14}$$

where n_r is the number of vertices in group r and κ_r is the sum of the degrees of the vertices in group r.

We use equation (13) to derive the probability $P(g,k \mid A)$:

$$P(g,k \mid A) = \frac{P(g,k)P(A \mid g,k)}{P(A)}, \tag{15}$$

where

$$P(g,k) = (n-2)^{-k} \prod_{r=1}^{k} n_r!. \tag{16}$$

The values k and g define the "state" of a statistical mechanical system with the probability $P(g,k \mid A)$. States of this system are sampled in proportion to the probability using Markov chain Monte Carlo sampling. Then, an estimate of the probability $P(k \mid A)$ of having k communities given the observed network A is found using the histogram of values of k over the Monte Carlo sample. Then, the most likely value of k is the one for which $P(k \mid A)$ is greatest. For one network, we performed 10000 Monte Carlo sweeps, each one of which include n individual nodes moves of two types [11]. After one sweep, the values k and g may change, if $k = 5$, $k = 3$, and $k = 2$ show out, respectively, 5000, 3000, and 2000 sweeps, then fraction of community numbers $P(k = 5/A) = 5000/10000 = 0.5$, $P(k = 4/A) = 0.3$, and $P(k = 2/A) = 0.2$. Thus, the most likely value of k is 5.

We set A^{appro} as the input to the unipartite network communities number estimating method of Riolo et al. [11] to estimate the number of communities in the network of type-a vertices. Then, by transposing the affiliation matrix B and using the same method as above, we can estimate the number of communities in the network of type-b vertices.

Now, let us analyze the time complexity of our method for type-a network. In the projection procedure, we take $O(N_a^2 N_b)$ times to calculate cosine similarity. Then, we take time approximately $O(N_a^2)$ to finish the filtering procedure, mainly finding the total degree $\sum_{ij} C_{ij} \mid \phi$ of one filtering matrix. Finally, we take $O(N_a k)$ to move n nodes to k communities and $O(N_a^2 k^2)$ to calculate the complete probability $P(g, k \mid A)$, so the estimation procedure takes time $O(N_a^2)$. Therefore, the complete time complexity of our method is $O(N_a^2 N_b + N_b^2 N_a)$.

3. Results

In this section, we compare the partitions generated by other one-mode projections with the performance of the proposed biCNEQ. There are two types of projections, which we call classical unweighted projection (CUP) and classical weighted projection (CWP) in order to distinguish from our method. An unweighted projection of a bipartite network onto its type-a vertices is obtained by letting two type-a vertices i and j be connected if they share any type-b neighbor k. Each edge of a weighted projection has a weight equal to the number of shared neighbors. Given an adjacency matrix $B = N_a \times N_b$, the classical weighted projection matrix P and the classical unweighted projection matrix $\overline{P}$ are, respectively, given by $P = B^2$ and $\overline{P}_{ij} = \begin{cases} 1, & \text{if } P_{ij} \geq 1, \\ 0, & \text{if } P_{ij} = 0, \end{cases}$ where the diagonal blocks of $N_a \times N_a$ and $N_b \times N_b$ correspond to the projections onto type-a and type-b vertices, respectively. The matrix P is equivalent to a "two-step" adjacency matrix, with each entry weighted by the number of length-2 paths between each pair of vertices [4]. Then, we set P or $\overline{P}$ as the input to the unipartite network communities number estimating method of Riolo et al. [11]. We will demonstrate that our method performs better than CUP and CWP in the following sections.

3.1. Synthetic Network: The Easy Case. As the mixed parameter λ increases, i.e., the level of noise is decreased, the fraction of correct community numbers of the approximation network of type-a vertices and type-b vertices calculated by the biCNEQ increases as a whole (blue line in Figure 4). However, CUP and CWP only give correct community numbers of the network in the noise-free situation ($\lambda = 1$), as the red and green lines in Figure 4. Then, we use our method to derive the approximation networks of synthetic mixed networks generated with $\lambda = 0.6$ and $\lambda = 0.65$ (red circles in Figure 4) and show posterior probabilities of the number of communities in the approximation network in Figure 5. As shown in Figures 4 and 5, when $\lambda \geq 0.65$, our method can estimate the correct number $k_a = 4$ of communities in the type-a vertices approximation network with the adjacent matrix A_a for this easy case. Analysis of the type-b vertices approximation network is carried out in the same way.

Next, we test whether the biCNEQ scales well when the parameters of synthetic networks were set as Table 1. Firstly, we define success estimation rate (SER) as

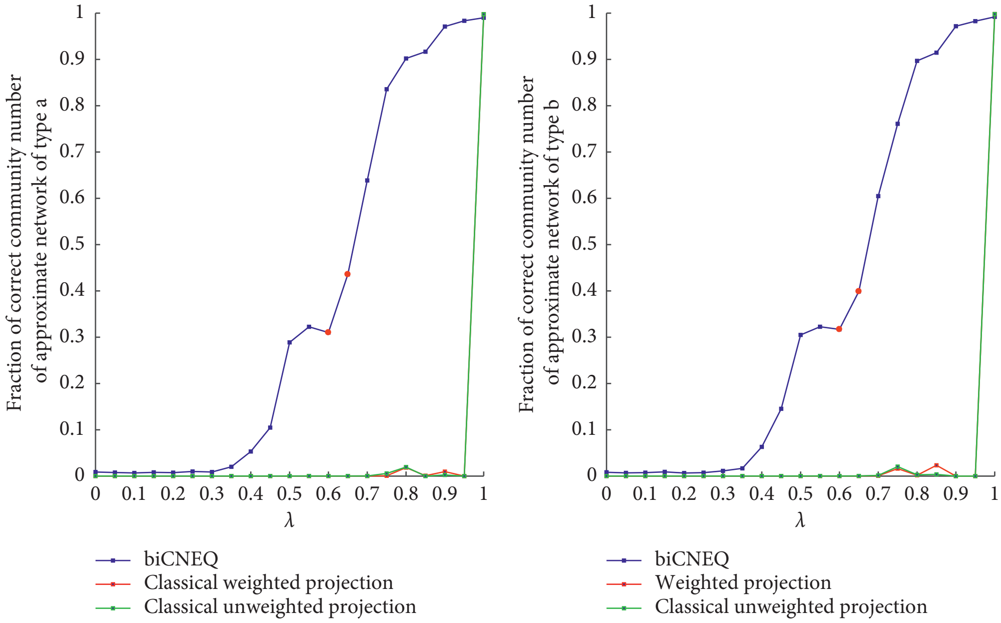
$$\text{SER} = \frac{\text{the runs of Monte Carlo sweeps that estimates correctly the highest average likelihood}}{\text{the total runs of Monte Carlo sweeps performed}}, \quad (17)$$

where each run includes 50000 Monte Carlo sweeps. The greater the SER is, the better the biCNEQ performs. As can be seen from Figure 6(a), our method performs reliably when $\lambda \geq 0.65$ and $k_a = k_b = 4$ whereas CUP and CWP can only deal with the network when noise free. As can be seen from Figure 6(b), the biCNEQ performs less well when the size of communities n_g grows. The biCNEQ performs less well when the level of noise increases, and the number K of planted communities grows and hardly gives right community number when $K = K_a + K_b = 20$ and $n_g = 300$.

3.2. Synthetic Network: The Difficult Case. For our method, as shown in Figure 7 (blue line), when the level of noise is decreased, the fraction of correct estimates of the number of communities of the approximation network type-a vertices remains stable with small fluctuations while $\lambda < 0.8$. It increases sharply when $\lambda \geq 0.8$. That of type-b vertices remains stable with small fluctuations when $\lambda < 0.75$ and increases sharply when $\lambda \geq 0.75$. We used our method to derive the approximation networks of synthetic mixed networks generated with $\lambda = 0.8$ and 0.75 (see red circles in Figure 7) and show posterior probabilities of the number of communities in the approximation network in Figure 8. As seen from Figures 7 and 8, our method can estimate the correct number $k_a = 2$ of communities in the type-a vertices approximation network with the adjacent matrix A_a when $\lambda \geq 0.8$ and $k_b = 3$ of communities in the type-b vertices approximation network with the adjacent matrix A_b when $\lambda \geq 0.75$ for the difficult case.

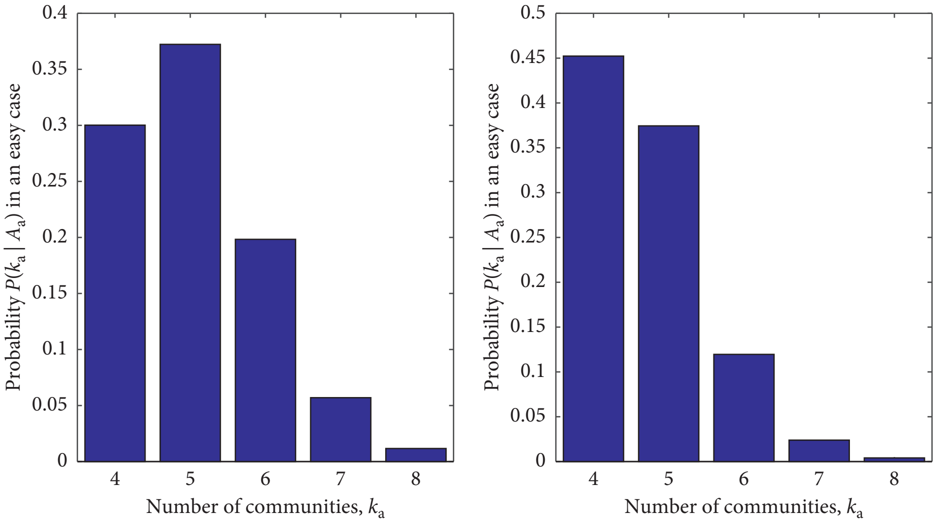
However, CUP and CWP can only correctly identify three communities in the type-b unipartite network without noise ($\lambda = 1$) and fail to estimate the correct community number in testing on the type-a unipartite network, as the red and green lines shown in Figure 7, respectively. The reason is the average size of type-a communities (35 nodes) is bigger than that of type-b communities (10 nodes). Furthermore, we found that biCNEQ performs less well when the size of communities n_g grows and fails to work even when n_g reaches 40 nodes with $\lambda = 1$ and $k_a = 2, k_b = 3$, which is a very small network.

3.3. Empirical Networks. We use our method, with a filtering coefficient of $\phi = 0.614$, as shown in Figure 3, to create the Southern women approximation network, whose adjacency matrix is denoted as A_w. Then, we use A_w as the input to the method of Riolo et al. [11]. As shown in Figure 9(a), we can estimate the correct number $k_w = 2$ of Southern women communities which matches the one determined in [4, 21, 22]. Figure 9(b) shows that, for the events, $k_e = 3$



(a) (b)

Figure 4: Test of biCNEQ against classical projections on synthetic networks in the easy case. Each point shows the median of the corresponding values on 100 networks. (a) The fraction of correct number $k_a = 4$ of communities in the approximation network of type-a vertices. (b) The fraction of correct number $k_b = 4$ of communities in the approximation network of type-b vertices.

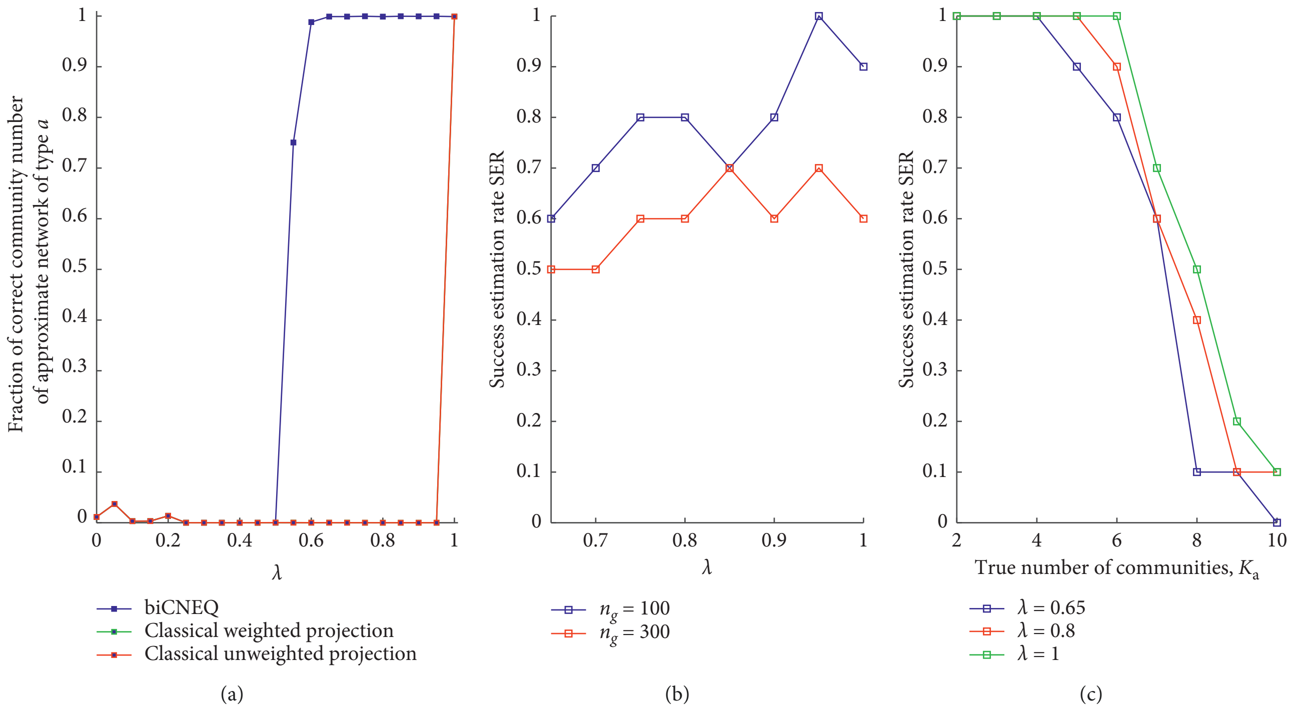


(a) (b)

Figure 5: Posterior probabilities calculated for the approximation networks of the synthetic mixed networks in the easy case. The synthetic networks were generated with $\lambda = 0.6$ (a) and $\lambda = 0.65$ (b). Each bar shows the median of the corresponding values on 100 networks.

Table 1: Parameters set for synthetic networks.

	K	n_g	λ	$m_{rs}, r = 1, \dots, K_a$, $s = r + K_a$
Figure 6(a)	$K_a = K_b = 4$	100	0 to 1 with one step is 0.05	2000
Figure 6(b)	$K_a = K_b = 7$	100, 300	0.65 to 1 with one step is 0.05	2000, 18000
Figure 6(c)	$K_a = K_b = 2, \dots, 10$	300	0.65, 0.8, 1	18000

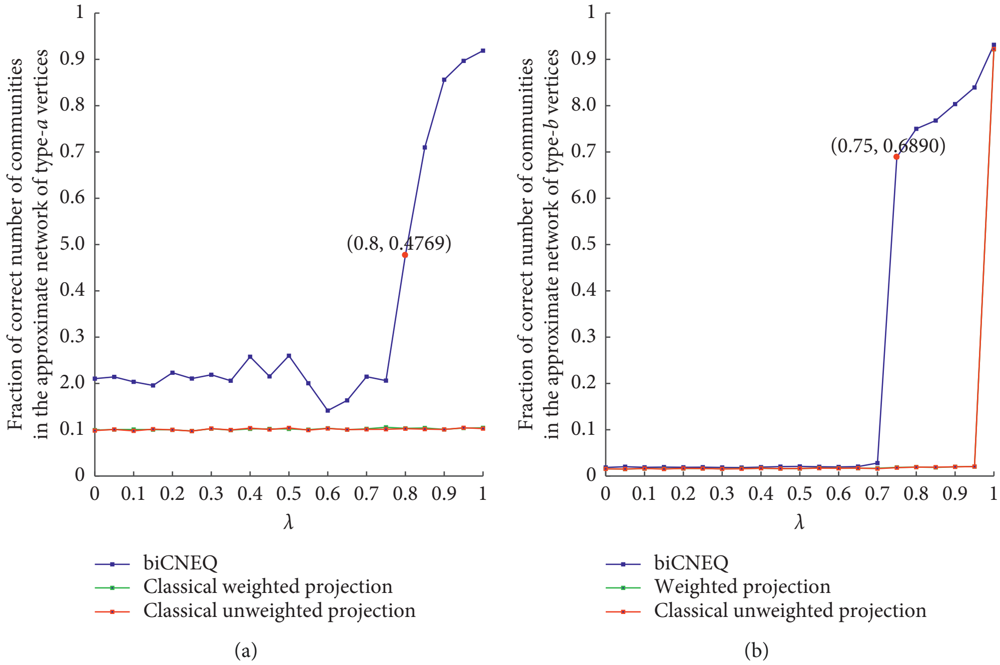


(a)

(b)

(c)

Figure 6: Tests of biCNEQ on synthetic networks with parameters set as Table 1. For each network, we performed 10 runs of 50 000 Monte Carlo sweeps each. (a) The fraction of correct number $k_a = 4$ of communities in the approximation network of type-a vertices found by biCNEQ against classical projection as a function of λ. Each point shows the results from the run that finds the highest average likelihood. (b) The success estimation rate SER of biCNEQ tested on the approximation network of type-a of synthetic network with $n_g = 100$ against $n_g = 300$ as a function of λ. (c) The success estimation rate SER of biCNEQ tested on the approximation network of type-a of synthetic network with different values of λ as a function of the true number of communities K_a.



(a)

(b)

Figure 7: Test of biCNEQ against classical projections on synthetic networks in the difficult case. Each point shows the median of the corresponding values on 100 networks. The fraction of correct number (a) $k_a = 2$ of communities in the approximation network of type-a vertices and (b) $k_b = 3$ of communities in the approximation network of type-b vertices.

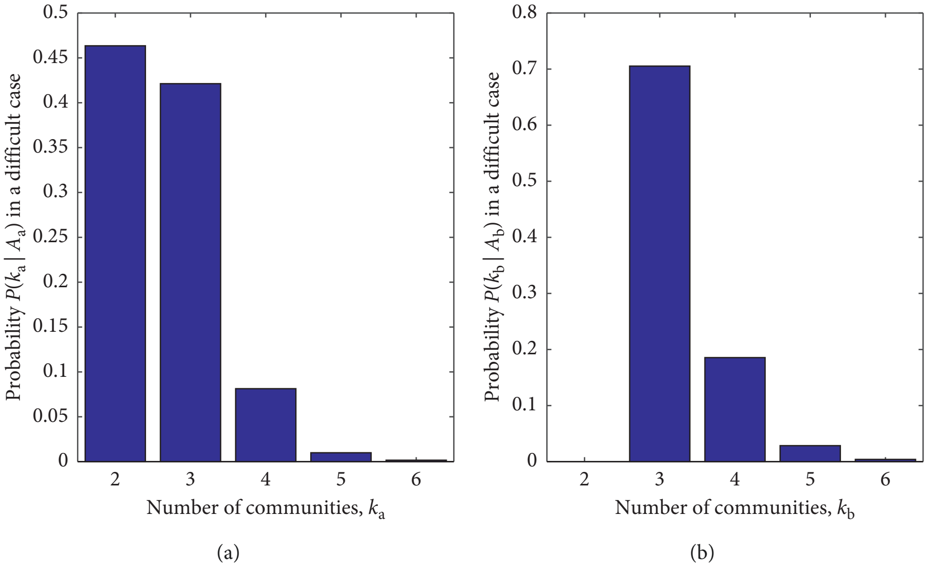


Figure 8: Posterior probabilities calculated for the approximation networks of the synthetic mixed networks in the difficult case. The approximation network of (a) type-*a* vertices generated from a synthetic mixed network with $\lambda = 0.8$ and (b) type-*b* vertices generated from a synthetic mixed network with $\lambda = 0.75$. Each bar shows the median of the corresponding values on 100 networks.

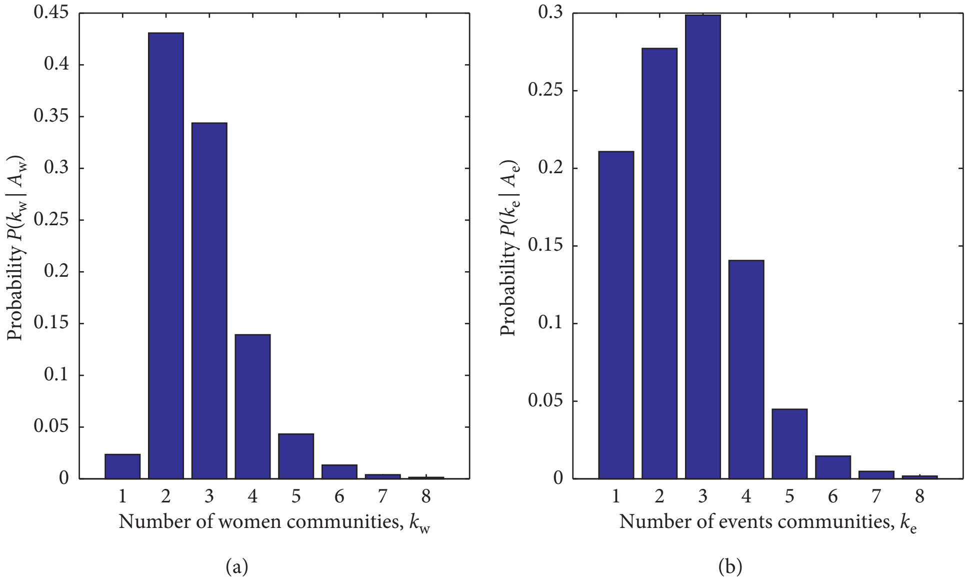


Figure 9: Posterior probabilities calculated for the unipartite approximation network for the Southern women dataset using our method. For the approximation network of women and events, we performed 10 runs of 50, 000 Monte Carlo sweeps, respectively.

receives the most weight but $k_e = 2$ comes a close second. It is interesting that the number of events groups in [21] is 2 but in [4], it is 3. However, the community numbers of women and events calculated by both the classical projection methods are 1 which is not plausible.

4. Conclusions

In this paper, we developed a method called biCNEQ for inferring the number of pure-type communities into which a bipartite network can divide. We designed a measure of the filtering quality named QFC to select a threshold of filtering coefficient to filter a weighted similarity network projected from a bipartite network to obtain a binary unipartite network. Then, we used the method of [11] to estimate the number of communities in the approximation network of each type of vertices. Via tests, biCNEQ gives correct answers and performs better than the classical unweighted and weighted projection methods on an empirical network with a known community structure and mixed synthetic networks including an easy case and a difficult case.

As discussed in the last section, the performance of our method degrades when the community size grows, especially in the difficult case synthetic network. This shortcoming makes biCNEQ hard to scale well on the real-world networks where there exists community structure. The reason for this may be due to information loss in the projection and filtering procedure or other stages of the proposed method. Thus, in the future work, the following two issues can be investigated: (1) an improved projection approach to minimize the information lost in the biCNEQ method can be developed and (2) a projection-free approach using a bipartite degree-corrected stochastic block model and Markov chain Monte Carlo sampling may be proposed.

Data Availability

The community and edge data of the "karate club" network and the "political blogs" network are obtained from Newman's web pages (http://www-personal.umich.edu/~mejn/dcsbm/ZacharyCorrectOutput/DegreeCorrected/ActualComms.tsv, http://www-personal.umich.edu/~mejn/dcsbm/ZacharyCorrectOutput/DegreeCorrected/EdgeLists.tsv, http://www-personal.umich.edu/~mejn/dcsbm/PolBlogsCorrectOutput/DegreeCorrected/ActualComms.tsv, and http://www-personal.umich.edu/~mejn/dcsbm/PolBlogsCorrectOutput/DegreeCorrected/EdgeLists.tsv).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61773250), the Shanghai Municipal Education Commission of China (no. 2019PD1-2-37), and the Program for Shanghai Youth Top-Notch Talent.

References

[1] K. Yang, Q. Guo, and J.-G. Liu, "Community detection via measuring the strength between nodes for dynamic networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 509, pp. 256–264, 2018.

[2] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, article 011047, 2014.

[3] C. Zhou, L. Feng, and Q. Zhao, "A novel community detection method in bipartite networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 492, pp. 1679–1693, 2018.

[4] D. B. Larremore, A. Clauset, and A. Z. Jacobs, "Efficiently inferring community structure in bipartite networks," *Physical Review E*, vol. 90, no. 1, article 012805, 2014.

[5] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, article 208701, 2002.

[6] L. Wu, L. Dong, Y. Wang et al., "Uniform-scale assessment of role minimization in bipartite networks and its application to access control," *Physica A: Statistical Mechanics and Its Applications*, vol. 507, pp. 381–397, 2018.

[7] T. Wang, L. Yin, and X. Wang, "A community detection method based on local similarity and degree clustering information," *Physica A: Statistical Mechanics and Its Applications*, vol. 490, pp. 1344–1354, 2018.

[8] M. Tasgin and H. O. Bingol, "Community detection using preference networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 495, pp. 126–136, 2018.

[9] K. R. Žalik and B. Žalik, "A framework for detecting communities of unbalanced sizes in networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 490, pp. 24–37, 2018.

[10] J. Xiao, Y.-J. Zhang, and X.-K. Xu, "Convergence improvement of differential evolution for community detection in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 503, pp. 762–779, 2018.

[11] M. A. Riolo, G. T. Cantwell, G. Reinert, and M. E. J. Newman, "Efficient method for estimating the number of communities in a network," *Physical Review E*, vol. 96, no. 3, article 032310, 2017.

[12] X. Zhou, X. Zhao, Y. Liu, and G. Sun, "A game theoretic algorithm to detect overlapping community structure in networks," *Physics Letters A*, vol. 382, no. 13, pp. 872–879, 2018.

[13] H.-L. Sun, E. Ch'ng, X. Yong, J. M. Garibaldi, S. See, and D.-B. Chen, "A fast community detection method in bipartite networks by distance dynamics," *Physica A: Statistical Mechanics and Its Applications*, vol. 496, pp. 108–120, 2018.

[14] S. Han, M. Sun, B. C. Ampimah, and D. Han, "Epidemic spread in bipartite network by considering risk awareness," *Physica A: Statistical Mechanics and Its Applications*, vol. 492, pp. 1909–1916, 2018.

[15] E. Corel, R. Méheust, A. K. Watson, J. O. McInerney, P. Lopez, and E. Bapteste, "Bipartite network analysis of gene sharings in the microbial world," *Molecular Biology and Evolution*, vol. 35, no. 4, pp. 899–913, 2018.

[16] Q. Cai and J. Liu, "Hierarchical clustering of bipartite networks based on multiobjective optimization," *IEEE Transactions on Network Science and Engineering*, 2018.

[17] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, article 026113, 2004.

[18] A. Miyauchi and N. Sukegawa, "Maximizing Barber's bipartite modularity is also hard," *Optimization Letters*, vol. 9, no. 5, pp. 897–913, 2014.

[19] R. Lambiotte and M. Ausloos, "Uncovering collective listening habits and music genres in bipartite networks," *Physical Review E*, vol. 72, no. 6, article 066107, 2005.

[20] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South*, University of Chicago Press, Chicago, IL, USA, 1941.

[21] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, "Module identification in bipartite and directed networks," *Physical Review E*, vol. 76, no. 3, article 036102, 2007.

[22] P. Zhang, D. Wang, and J. Xiao, "Improving the recommender algorithms with the detected communities in bipartite networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 471, pp. 147–153, 2017.

[23] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[24] L. Bai, J. Liang, H. Du, and Y. Guo, “A novel community detection algorithm based on simplification of complex networks,” *Knowledge-Based Systems*, vol. 143, pp. 58–64, 2018.

[25] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[26] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

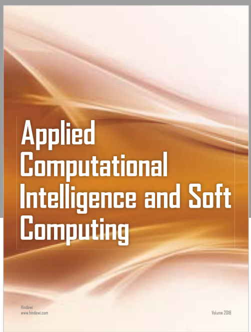
[27] S. Fortunato and D. Hric, “Community detection in networks: a user guide,” *Physics Reports*, vol. 659, pp. 1–44, 2016.

[28] M. E. J. Newman, “Mixing patterns in networks,” *Physical Review E*, vol. 67, no. 2, article 026126, 2003.

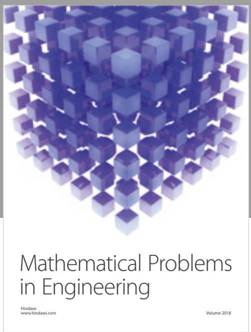
[29] W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[30] L. A. Adamic and N. Glance, “The political blogosphere and the 2004 US election,” in *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, Washington, DC, USA, February 2005.

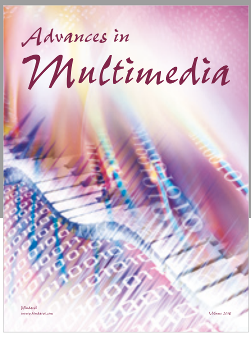
[31] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Physical Review E*, vol. 83, no. 1, article 016107, 2011.

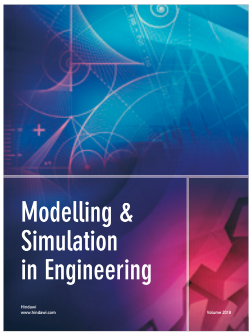

Applied Computational Intelligence and Soft Computing

The Scientific World Journal


Mathematical Problems in Engineering

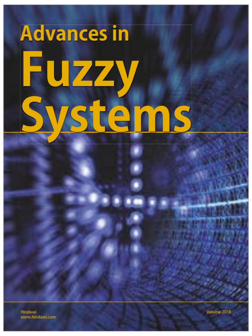
Journal of Engineering

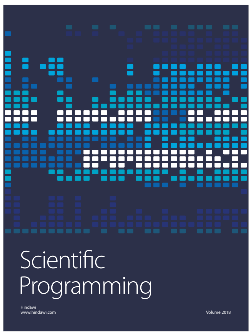

Advances in Multimedia

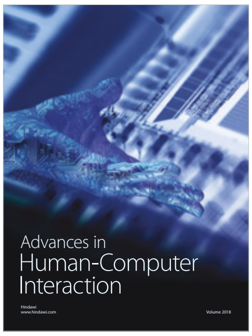

Modelling & Simulation in Engineering

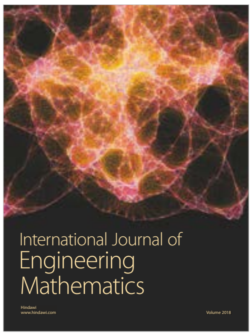
Advances in Artificial Intelligence

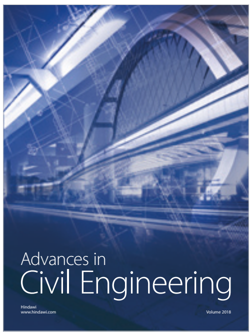
International Journal of Reconfigurable Computing

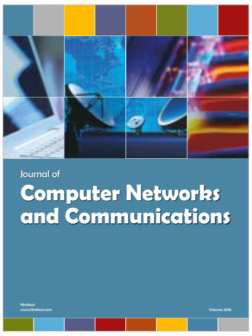

Advances in Fuzzy Systems


Scientific Programming


Advances in Human-Computer Interaction

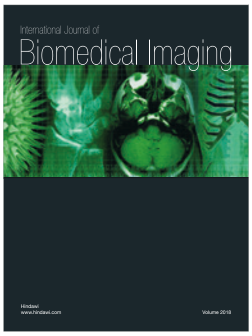

International Journal of Engineering Mathematics


Advances in Civil Engineering

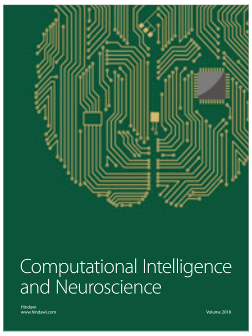

Journal of Computer Networks and Communications

Journal of Robotics

International Journal of Computer Games Technology


International Journal of Biomedical Imaging

Journal of Electrical and Computer Engineering


Computational Intelligence and Neuroscience