ISSN : 0974 - 7435 Volume 10 Issue 6

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 10(6), 2014 [1353-1360]

A Novel community detecting algorithm in biological bipartite networks

Wei Liu*, Yueyang Zhang, Ling Chen
Information Science and Technology College, Yangzhou University, Yangzhou 225127, (CHINA)

ABSTRACT

The identification of communities is significant for the understanding of network structures and functions. In this paper, we propose a framework to address the problem ofcommunity detection in bipartite networks based on principal components analysis. We show that bipartite network can be conveniently represented as linear graph for module identification purposes. We apply the algorithm to real-world network data, showing that the algorithm successfully finds meaningful community structures of bipartite networks.

KEYWORDS

Bipartite network;
Community detecting;
Principal components analysis.

INTRODUCTION

In recent years, More and more researchers have been interested in the modularstructure identification[1] of the network. Informally, a modular structure, or community is a subgraph whose verticesare more likely to be connected to one another thanto the vertices outside the subgraph.

Now the detection and analysis of modular structures or communities have been intensively studied in relation to its applications in the analysis of networks[2-4] recently. Up to now, many algorithms have been proposed for detecting communities. Two classical algorithmsare the spectral bisection algorithm on the basis of the eigenvectors of the Laplacian matrix of a network[5] and the Kernighan-Lin algorithm that improves the initial division by optimizing the number of within-and between-community edges[6]. Recently, many algorithms based on modularity[7] are proposed. For example, Newman proposed a fast greedy algorithm[8] to maximize the modularity. The same algorithm implemented with a better data structure is proposed by Clauset et al.[9] which is typically thousands of times faster than the algorithm proposed by Kernighan and Lin[6]. An even faster and more accurate algorithm based on subgraph similarity is proposed by Xiang et al.[10] Ruan and Zhang[11] proposed an efficient heuristic algorithm which combines a spectral graph partitioning and a local searching to optimize the modularity. Duch and Arenas[12] presented a method to find community structure by extremal optimization subject to the modularity. Wang et al.[13] proposed a very fast algorithm for community detection based on local information. Newman proposed an algorithm using the eigenvectors of matrices.[14] Chen et al.[15] presented a fast and efficient algorithm by adding anode into a partial community recursively until obtaining a local optimal community.

However, most of the previous works are for unipartite networks. In real world situations, there are many bipartite networks composed of two types of non-overlapping nodes and the links which must have one end node from each set. In this paper, a new approach

FULL PAPER

named CDA_PCA (community detecting algorithm based on principal components analysis) is proposed to identify the communities in bipartite network. The main idea is firstly to transform the bipartite network into the linear graph, next to improve the graph's incidence matrix, then to use principal components analysis for the incidence matrix and finally to detect communities in the two parts of bipartite network. Experimental results show our algorithm is especially suited for module detection in bipartitenetworks.

METHOD

Basic properties of bipartite network

A bipartite network $G = (V, E)$ is a graph whose vertices can be divided into two disjoint sets V_1 and V_2 such that every edge connects a vertex in V_1 to one in V_2; that is, V_1 and V_2 are independent sets. Besides, there is no edge connected between any two vertices in V_1 or V_2. Supposed the partitions of the bipartite networkare of size $|V_1| = n_1$ and $|V_2| = n_2$, respectively. The adjacency matrix of a bipartite network can be represented as the form of block matrix. In the adjacency matrix, if the first n_1 rows and columns correspond to thevertices in V_1 and the firstrows and columns correspond to the vertices in V_2, the adjacency matrix of G can be obtained as: $A = \begin{bmatrix} O & A_1 \\ A_2 & O \end{bmatrix}$ where A_1 is a $n_1 \times n_2$ matrix, A_2 is a $n_2 \times n_1$ matrix and O is an all-zero matrix. Furthermore, A_1 and A_2 satisfy the following condition: $A_1 = A_2^T$.

Example1: Given a bipartite network H'shown as follows:

We can obtain its adjacency matrix as $A = \begin{bmatrix} 0&0&0&0&1&1&0 \\ 0&0&0&0&1&0&1 \\ 0&0&0&0&1&0&1 \\ 0&0&0&0&0&1&1 \\ 1&1&1&0&0&0&0 \\ 1&0&0&1&0&0&0 \\ 0&1&1&1&0&0&0 \end{bmatrix}$, We set, and $A' = \begin{bmatrix} 1&1&1&0 \\ 1&0&0&1 \\ 0&1&1&1 \end{bmatrix}$ $A = \begin{bmatrix} 0 & A'^T \\ A' & 0 \end{bmatrix}$. Note that the matrix A' uniquely represents the bipartite network.

Linear graph

For any graph e.g. Figure 2, we will take its edges as a set of vertices named V_1 and its vertices as another set of vertices V_2,which constitute a bipartite network. If a vertex of V_2 is related to a vertex of V_1,an edge will be added between the two vertices. After these conversion, graph G can be described as a bipartite networkH' shown in Figure 1.

As we noted, the bipartite networkH' has preserved all information of the original graph G. And the other way round, the bipartite networkH' also can be converted into the graph G which will keep the total information.

It can be seen that the submatrix A' of adjacency matrix A mentioned above is just the incidence matrix of the graph G as shown in figure 2 which indicates that figure 2 reserves all information of the bipartite networkH'.

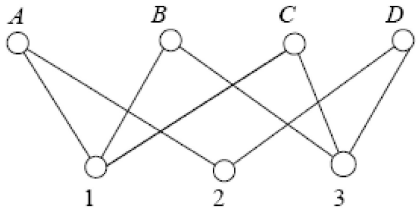

Figure 1 : Bipartite Network*H'*

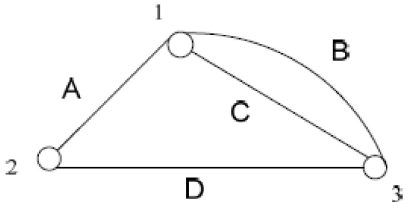

Figure 2 : Graph *G*

If we count A, B, C and D in figure 1 as s set of vertices while 1, 2 and 3 in it as a set of edges, we would obtain a graph named linear graph correspond-

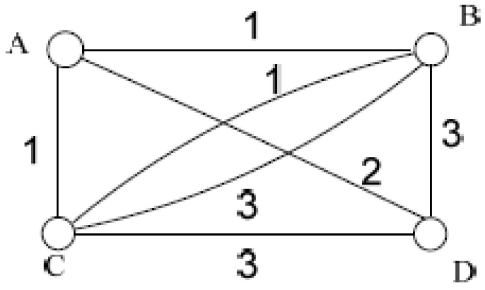

Figure 3 : Linear Graph *G'*

ing to the graph *G* shown as follows:

In accordance with Figure 3, we can find that the submatrix A^{π} of adjacency matrix *A* is just the incidence matrix of the linear graph *G'*. Therefore it can be concluded that G' has retained all information of the original graph *G* and the bipartite network*H'*. And each vertex in *G* would correspond to a clique in *G'*, that is, a complete subgraph.

The basic idea of our algorithm

Given a bipartite network, for example,gene-disease bipartite network*H* shown as Figure 4, we can view the nodes of disease phenome as vertices and the nodes of disease genome as edges which constitute the graph *G*. After clustering for the vertices of *G*, we can detect the communities in it by the use of communities detecting algorithms and then make clustering for disease phenomes.

Similarly, we can also consider the nodes of disease genome as vertices and the nodes of disease phenome as edges which constitute the linear graph *G'* of *G*. After mining communities for vertices of *G'*, we can make clustering for disease genome.

However, as mentioned above, the bipartite network*H*, graph *G* and linear graph *G'* are equivalent. In other words, as long as community detection for *G* or *G'*, we can both obtain the clustering results for disease phenome and disease genome at one time.

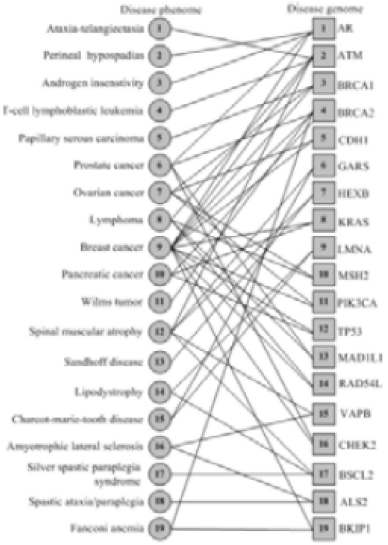

Figure 4 : The Diseasome Bipartite Network*H*

THE NETWORK COMMUNITY DETECTION ALGORITHM BASED ON PRINCIPAL COMPONENTS ANALYSIS

Supposed graph *G* is generated by the bipartite network*H* whose adjacency matrix is $A=\begin{bmatrix}0 & A^{\pi}\\ A' & 0\end{bmatrix}$, and the incidence matrix of *G* is A'. We use principal component analysis for the vectors of A' so as to detect communities in *G* based on the incidence matrix.

Principal components analysis

Assumed that there are data points $x_1, x_2, ...x_m$ of n-dimension space,which can be denoted as the matrix

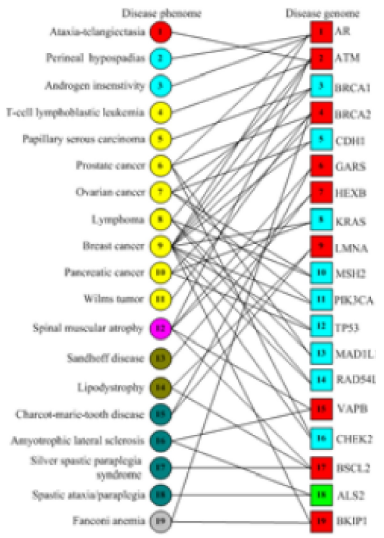

Figure 5 : Partitions of the disease-gene network obtained by our method

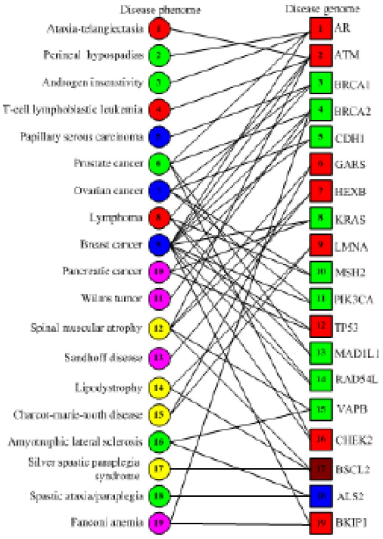

Figure 6 : Partitions of the disease-gene network obtained by Ref.[16]

FULL PAPER

as follows, $X=\begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1n} \\ x_{21} & x_{22} & \cdots\cdots & x_{2n} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ x_{m1} & x_{m2} & \cdots\cdots & x_{mn} \end{bmatrix}=\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_m \end{pmatrix}$

Where $\overline{x_j}=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$ and $\overline{x}=(\overline{x_1},\overline{x_2},\mathrm{L},\overline{x_n})$.

Let the denotation of the matrix be $\hat{x}=\begin{bmatrix} x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \cdots\cdots & x_{1n}-\overline{x_n} \\ x_{21}-\overline{x_1} & x_{22}-\overline{x_2} & \cdots\cdots & x_{2n}-\overline{x_n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{m1}-\overline{x_1} & x_{m2}-\overline{x_2} & \cdots\cdots & x_{mn}-\overline{x_n} \end{bmatrix}$,Thus the covariance matrix is $S=\frac{1}{m-1}\hat{x}\hat{x}^T=\frac{1}{m-1}\begin{bmatrix} \sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{1k}-\overline{x_1}) & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \sum_{k=1}^{n}(x_{ik}-\overline{x_i})(x_{jk}-\overline{x_j}) & \cdots \\ \cdots & \cdots & \cdots \\ \sum_{k=1}^{n}(x_{mk}-\overline{x_k})(x_{1k}-\overline{x_k}) & \cdots & \cdots \end{bmatrix}$

Where S is a $m\times m$ symmetrical matrix and itcan be rewritten as the form of matrix-vector:

$$S=\frac{1}{m-1}\sum_{k=1}^{m}\left(x_k-\overline{x}\right)\left(x_k-\overline{x}\right)^T$$

Next we'll computer eigenvalues of S, namely $\lambda_1\geq\lambda_2\geq\lambda_3\cdots\geq\lambda_m\geq 0$ and its corresponding eigenvectors $l_1l_2\cdots l_m$ which are regarded as orthogonalization, that is $l_i\cdot l_i^T=1,\ l_i\cdot l_j=0$,.

We select p bigger eigenvalues, viz.$\lambda_1\lambda_2\cdots\lambda_p$ whose corresponding eigenvectors are $l_1l_2\cdots l_p$. Given a $m\times p$ matrix comprised of $u=\left(l_1l_2\cdots l_p\right)$,the data x (a m-dimension vector)can be denoted as x' in the new space such that $x'=x\cdot u$ where x' is a p -dimension vector.

Principal components analysis of theincidence matrix A'

In order to make analysis of principal components for convenience, we have to reconstruct the incidence matrix as follows:

(1) when the number of "1" in each column in excess of 2(normally, in the incidence matrix, "1"only appears twice in each column,but there may be multiple "1"emerged here which is determined by the properties of the bipartite network),we should change "1" of each column into $\frac{1}{the\ number\ of\ "1"\ in\ the\ column}$ and then make pairwise combination for them. After these transformations, we can get C_k^2 columns,for instance: The matrix $A=\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$ will be converted into $\begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{2} \end{bmatrix}$ and then continuously be changed into $A'=\begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{2} \end{bmatrix}$.

(2) We use two directed edges instead of each undirected edge denoted by each column, namely"1"stands for the head of the edge and "-1"denotes the tail, thus the matrix A' mentioned above will be changed into:

$$A''=\begin{bmatrix} 1/3 & -1/3 & 0 & 0 & 1/3 & -1/3 & 1/2 & -1/2 \\ -1/3 & 1/3 & 1/3 & -1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1/3 & 1/3 & -1/3 & 1/3 & -1/2 & 1/2 \end{bmatrix}$$

Denote the transformed matrix as B which is a $m\times N$ matrix, then its covariance matrix is a $m\times m$ square matrix shown as:

$$S=\frac{1}{N-1}BB^T\quad(*)$$

The proof of the formula(*)is omitted due to the limited space.

The property of covariance matrix S

(1) S is a symmetrical square matrix.

(2) Assumed that $i\neq j$, s_{ij} is the covariance in different dimensions between the vertex$i$and j. If there is a common vertex connected between i and j in the bipartite network, the covariance is negative; otherwise it is positive. In the network community detection, there is positive correlation between the tail and the head of the edge, so here we should calculate all eigenvalues of S and sort them ascendingly, then select the minimal eigenvalue and its eigenvec-

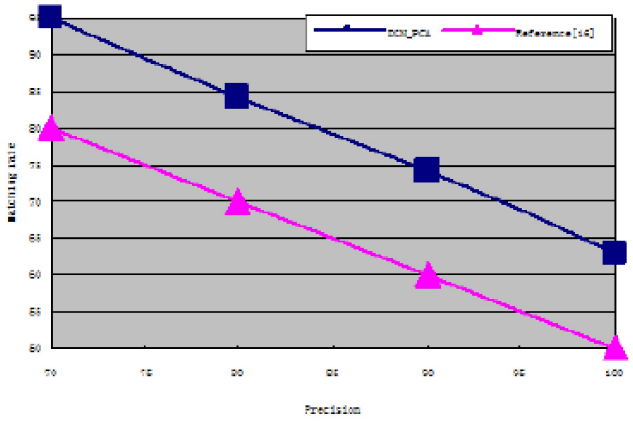

Figure 7 : Comparison of the matching rate of two algorithms with the increases of precision(clustering for Disease Phenome Network)

tor as discriminative vector.

(3) The minimal eigenvalueof S is 0.

(4) Supposed that G is generated by the incidence matrix A_1 of bipartite network H, the covariance matrix S and the Laplacian matrix of G are equal only up to a constant factor.

The proof is omitted due to the limited space.

THE FRAMEWORK OF THE ALGORITHM

Based on the analysis mentioned above, we can obtain the framework of our algorithm shown as follows:

Algorithm CDA_PCA

Input: Abipartite network W and its adjacency matrix $A=\begin{bmatrix} 0 & A_1^T \\ A_1 & 0 \end{bmatrix}$. Two sets of vertices in W, namely V_1 ,V_2 whose size are n_1 and n_2 respectively.

Output: Community groups of V_1 and V_2 .

Begin

if $n_1 > n_2$, $A_2 = A_1$,else $A_2 = A_1^T$;

Reconstructing A_2 and obtaining the matrix A_3 ;

Computing $S = A_3 A_3^T$;

Calculating eigenvalues of S, viz. $\lambda_1 \le \lambda_2 \le \ldots\ldots \le \lambda_n$, and the corresponding eigenvectors, namely $u_1, u_2, \ldots, u_n$;

Eliminating the minimal eigenvalue $\lambda_1 = 0$;

for i=2 to n_1 - 1 do $l_i = \lambda_{i+1} - \lambda_i$ End for

Searching for the maximal l_i denoted as l_k among $l_2, \ldots, l_{n_1-1}$,and then choosing $u_2, u_3, \ldots, u_k$ to build a $n \times (k-1)$ matrix, that is $u = [u_2, u_3, \ldots, u_k]$;

Supposed the row vector of u are $p_1, p_2, \ldots, p_n$, namely $u = \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ \cdot \\ \cdot \\ p_n \end{bmatrix}$, k-means algorithm is made use of clustering for $p_1, p_2, \ldots, p_n$;

Hence the clustering for vertices can be obtained and from that the corresponding edges clustering also can be got.

End

EXPERIMENTAL RESULTS AND ANALYSIS

All the experiments were conducted on a 3.0GHzPentium with 2G memory. All codes were compiled using MATLAB 7.0.

Performance comparison on real-world data

Now let us turn to real-world datasets. We extract a small subset from OMIM and construct a diseasome bipartite network named H as shown in Figure 4. We apply our algorithm to community detection in Figure 4 and investigate correlations between elements from the disease phenome and the disease genome, respectively.

In order to make further analysis for convenience, we compare our algorithm with the algorithm in[16] and the results are shown in Figure 5-Figure 8. From Figure 5 and Figure 6,we cansee that our algorithm ob-

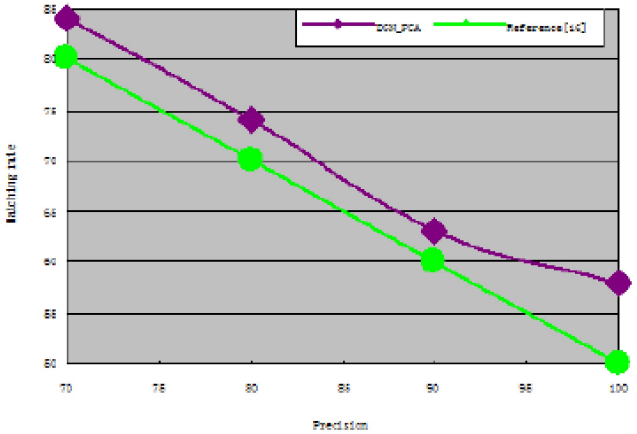

Figure 8 : Comparison of the matching rate of two algorithms with the increases of precision(clustering for Disease Genome Network)

FULL PAPER

tains seven modules about disease phenome and three modules about disease genome. The colors of circles and rectangles correspond to the community to which the disorders or disease genes belongs.

In accordance with the visual representation of disease-gene network, it can be concluded that our result accords with the fact better than Ref.[16] which indicates that our algorithm can deeply reveal the relativities between the diseases and the disease genes which is helpful to study the pathogenesis of inherited disease and make gene diagnosis and therapy.

We also making clustering for the two projections of the Diseasome: the projection on diseases, called HDN(Human Disease Network, left parts of Figure 4), and the projection on genes, called DGN (Disease Gene Network, right parts of Figure 4). The comparative resultsof the matching rate of two algorithms are shown as Figure 7-Figure 8.

According to Figure 7 and Figure 8, we notice that accompany with the increase of the precision, no matter clustering for HDN or DGN, our algorithm is of higher matching rate which indicate our algorithm is more efficient.

Southern women network

As the second experiment, we consider southern women network[17] to verify the accuracy of CDA_PCA. There are 18 women nodes and 14 events nodes, with 89edges linking to woman nodes and event nodes if the women attended the corresponding events[18-21].

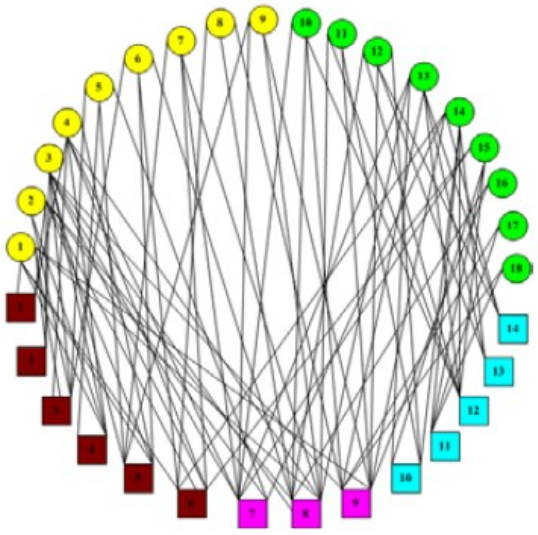

Figure 9 : Partitions of the Southern women network obtained by our method

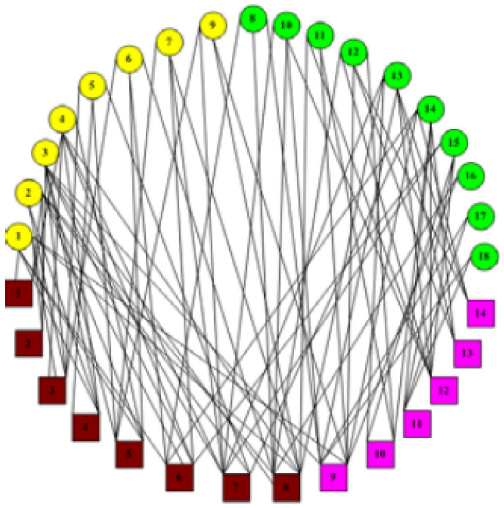

Figure 10 : Partitions of the Southern women network obtained by Guimera[19]

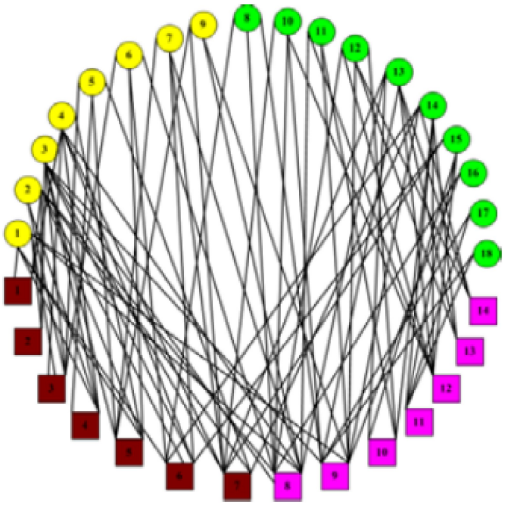

Figure 11 : Partitions of the disease-gene network obtained by Murata[20]

The community partitions obtained by our method and some other approaches are shown in Figure 9-Figure 13. Women are indicated as circle symbols located at the upper camber, while events are indicated as square symbols located at the lower camber. Nodes in the same community are painted in the same color. It can be observed that our method divides the events into three communities and 18 women into two groups. It is satisfying to find that only our partition for women (the circle symbols) is consistent with that proposed by Freeman. Our partition of events into three communities is also reasonable, as it conforms to the criteria of "good". We can see that event communities {1-6} and {10-14}, respectively, correspond to

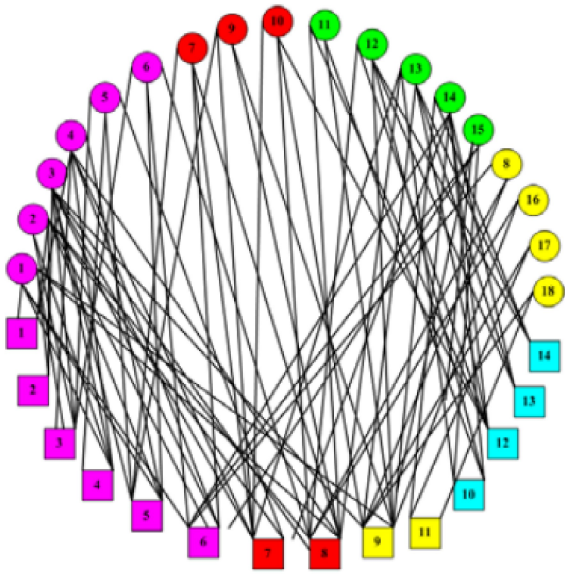

Figure 12 : Partitions of the disease-gene network obtained by Barber[18]

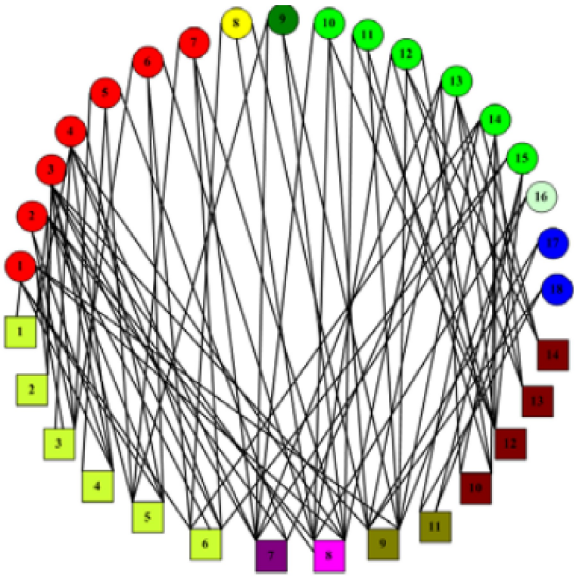

Figure 13 : Partitions of the disease-gene network obtained bySuzuki[21]

woman communities {1-9} and {10-18}, while event community {7-9} corresponds to both woman communities which indicate that the correspondence between communities obtained by our method is clear. We can also find that our method detect communities of many to-many correspondence well. However, Figure 10-Figure 12 can only detect communities of one-to-one correspondence. Figure 13 can detect communities of many to-many correspondence, but this partition seems somewhat irrational, since several communities contain only one node.

CONCLUSION

In this paper, we propose a novelcommunity detection method CDA_PCA based on principal components analysis. We convert the origin bipartite network into the equivalent graph or linear graph and make reconstruction for the graph's incidence matrix, and then detect communities by the use of principal components analysis method. Experimental results show our algorithm can successfully identify the modular structure of bipartite networks.

ACKNOWLEDGEMENTS

This research was supported in part by the Natural Science Foundation of Education Department of Jiangsu Province under grant No.12KJB520019, Natural Science Foundation of Jiangsu Province under contracts No. BK20130452.

REFERENCE

[1] R.Guimer'a, M.Sales-Pardo, L.A.N.Amaral; Classes of complex networks defined by role-to-role connectivity profiles [J]. Nature Phys., **3**, 63-69 **(2007)**.

[2] S.Maslov, K.Sneppen; Specificity and Stability in Topology of Protein Networks[J]. Science, **296(5569)**, 910-913 **(2002)**.

[3] V.Colizza, A.Flammini, M.A.Serrano, A.Vespignani; Detecting rich-club ordering in complex networks[J]. Nature Phys., **2**, 110-115 **(2006)**.

[4] S.Fortunato; Community detection in graphs [J]. Physics Reports-Review Section of Physics Letters, **486(3–5)**, 75-176 **(2010)**.

[5] M.Fiedler; Algebraic connectivity of Graphs[J]. Czechoslovak Mathematical Journal., **23(98)**, **(1973)**.

[6] B.W.Kernighan, S.Lin; An efficient heuristic procedure for partitioning graphs[J]. Bell Systems Technical Journal., **49** 291–307 **(1970)**.

[7] M.E.J.Newman, M.Girvan; Finding and evaluating community structure in networks[J]. Phys.Rev.E, **69**, 026113 **(2004)**.

[8] M.E.J.Newman; Fast algorithm for detecting communitystructure in networks[J]. Phys.Rev.E, **69**,066133, 2004.

[9] A.Clauset, M.E.J.Newman, C.Moore; Finding com-

munity structure in very large networks [J]. Phys.Rev.E, 70066111 **(2004)**.

[10] B.Xiang, E.H.Chen, T.Zhou; Finding community structure based on subgraph similarity[J]. Stud.Comput.Intell., **207**, 73–81 **(2009)**.

[11] J.H.Ruan, W.X.Zhang; Identifying network communities with a high resolution[J]. Phys.Rev.E, 77016104 **(2008)**.

[12] J.Duch, A.Arenas; Community detection in complex networks using extremal optimization[J]. Phys.Rev.E, 72027104 **(2005)**.

[13] X.T.Wang, G.R.Chen, H.T.Lu; A very fast algorithm for detecting community structures in complex networks[J]. Physica A: 384(2),pp 667-674,2007.

[14] M.E.J.Newman; Finding community structure in networks using the eigenvectors of matrices[J]. Phys.Rev.E, 74036104 **(2006)**.

[15] D.B.Chen, Y.Fu, M.S.Shang; A fast and efficient heuristic algorithm for detectingcommunity structures in complex networks[J]. Physica A, **388(13)**, 2741–2749 **(2009)**.

[16] Chen Wenqin, Lu Junan, Liang Jia; Research in Disease-Gene Network Based on Bipartite Network Projection[J]. Complex Systems and Complexity Science., **6(1)**, **(2009)**.

[17] A.Davis, B.B.Gardner, M.R.Gardner; Deep South[M]. University of Chicago Press, Chicago, IL, **(1941)**.

[18] M.J.Barber. Modularity and community detection in bipartite network[J]. Phys.Rev.E, **76**, 066102 **(2007)**.

[19] R.Guimera, M.S.Pardo, L.A.N.Amaral. Module identification in bipartite and directed networks[J]. Phys.Rev.E, **76**, 036102 **(2007)**.

[20] T.Murata, T.Ikeya; A new modularity for detecting one-to-many correspondence of communities in bipartite networks[J]. Advances in Complex Systems, **13(1)**, 19–31 **(2010)**.

[21] K.Suzuki, K.Wakita; Extracting multi-facet community structure from bipartite networks[J]. In Proc. Of International Conference on Computational Science and Engineering, Vancouver, BC, Canada, 312–319 Aug **(2009)**.