Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (13): 2421-2428, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2421.2428

Community Detecting in Bipartite Network Based on Principal Components Analysis

Wei Liu and 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 study, we propose a framework to address the problem of community detection in bipartite networks based on principal components analysis. We apply the algorithm to real-world network data, showing that the algorithm successfully finds meaningful community structures of bipartite networks.

Key words: Bipartite network, community detecting, principal components analysis

INTRODUCTION

The last few years have witnessed tremendous activity devoted to the applications to physical, chemical, biological, technological and social networks (Costa *et al*., 2007). Of great current interest is the modular structure identification of the network. Informally, a modular structure, or community (Fortunato, 2010) is a subgraph whose vertices are more likely to be connected to one another than to the vertices outside the subgraph.

Up to now, many algorithms have been proposed for detecting communities. For example, Newman proposed a fast greedy algorithm (Newman, 2004) to maximize the modularity. The same algorithm implemented with a better data structure is proposed by Clauset *et al.* (2004) which is typically thousands of times faster than the algorithm proposed by Kernighan and Lin (1970). An even faster and more accurate algorithm based on subgraph similarity is proposed by Xiang *et al.*(2009) and Ruan and Zhang (2008) proposed an efficient heuristic algorithm which combines a spectral graph partitioning and a local searching to optimize the modularity. Duch and Arenas (2005) presented a method to find community structure by extremal optimization subject to the modularity. Wang *et al.* (2007) proposed a very fast algorithm for community detection based on local information. Newman (2006) proposed an algorithm using the eigenvectors of matrices. Chen *et al.* (2009a) presented a fast and efficient algorithm by adding a node into a partial community recursively until obtaining a local optimal community.

In this study, a new approach 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 equivalent graph or 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 bipartite networks.

METHOD

Basic properties of bipartite network: A bipartite network H = (VE) 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. Supposed the partitions of the bipartite network are of size $|V_1| = n_1$ and $|V_2| = n_2$, respectively. The adjacency matrix of a bipartite network H can be represented as the form of block matrix:

$$A = \begin{bmatrix} O & A_1 \\ A_2 & O \end{bmatrix}$$

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:

$$A_1 = A_2^T$$

Example 1: 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:

Corresponding Author: Wei Liu, Information Science and Technology College, Yangzhou University, Yangzhou, 225127, China

$$A' = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

and:

$$A = \begin{bmatrix} 0 & A'^{T} \\ A' & 0 \end{bmatrix}$$

Note that the matrix A' uniquely represents the bipartite network.

Linear graph: For any graph G,e.g., Fig. 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 H'. And if we take {A, B, C, D} of H' as s set of vertices while {1, 2, 3} as a set of edges, we would obtain a graph G' named linear graph corresponding to G shown as Fig. 3.

In accordance with Fig. 1-3, we can find that the submatrix A' of adjacency matrix A mentioned above is

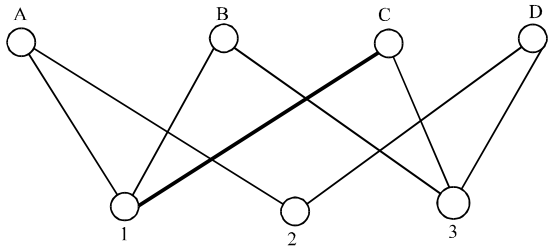

Fig. 1 Bipartite Network H'

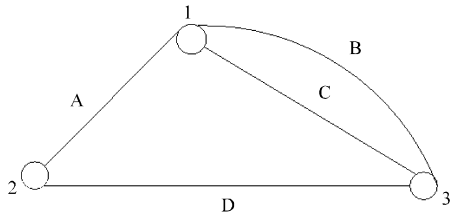

Fig. 2: Graph G

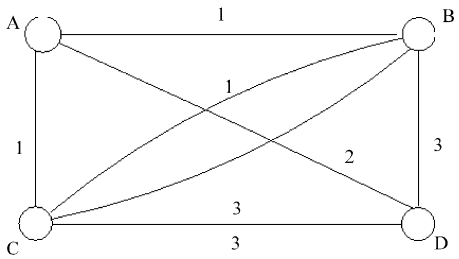

Fig. 3: Linear graph G'

just the incidence matrix of G and the sub matrix A'^{T} of adjacency matrix A is just the incidence matrix of the G'which indicate that H',G and G' are equivalent. In other words, as long as community detection for G or G', we can both obtain the clustering results for the node set {A, B, C, D} and the set {1, 2, 3} of H'.

THE NETWORK COMMUNITY DETECTION ALGORITHM BASED ON PRINCIPAL COMPONENTS ANALYSIS

Next we use principal component analysis for the vectors of incidence matrix A' so as to detect communities in G.

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 as follows:

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

Where:

$$\overline{x_j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$$

and $\overline{x} = (\overline{x}_1, \overline{x}_2,..., \overline{x}_n)$.

Let the denotation of the matrix be:

$$\hat{x} = \begin{bmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & ..... & x_{1n} - \overline{x_n} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & ..... & x_{2n} - \overline{x_n} \\ ..... & ..... & ..... & ..... \\ x_{m1} - \overline{x_1} & x_{m2} - \overline{x_2} & ..... & x_{mn} - \overline{x_n} \end{bmatrix}$$

Thus the covariance matrix is:

$$S = \frac{1}{m-1}\hat{x}\hat{x}^{T} = \frac{1}{m-1} \begin{bmatrix} \sum_{k=1}^{n}\left(x_{1k} - \overline{x_k}\right)\left(x_{1k} - \overline{x_k}\right) & ... & ... \\ ... & ... & ... \\ ... & \sum_{k=1}^{n}\left(x_{ik} - \overline{x_k}\right)\left(x_{jk} - \overline{x_k}\right) & ... \\ ... & ... & ... \\ \sum_{k=1}^{n}\left(x_{mk} - \overline{x_k}\right)\left(x_{1k} - \overline{x_k}\right) & ... & ... \end{bmatrix}$$

where, S is a m×m symmetrical matrix and it can be rewritten as the form of matrix-vector:

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

Next we'll computer eigenvalues of S, namely $\lambda_1 \geq \lambda_2 \geq \lambda_3 ... \geq \lambda_m \geq 0$ and its corresponding eigenvectors $l_1 l_2 ... l_m$ which are regarded as orthogonalization, that is $l_i.l_i^T = 1$, $l_i.l_j = 0$.

We select p bigger eigenvalues, viz., $\lambda_1 \lambda_2 ... \lambda_9$ whose corresponding eigenvectors are $l_1 l_2 ... l_p^a$, Given a m×p matrix comprised of $u = (l_1 l_2 ... l_p)$, the data x (a m-dimension vector) can be denoted as x' in the new space such that x' = x.u where x' is a p-dimension vector.

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

- when the number of "1" in each column in excess of 2 (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:

$$\frac{1}{\text{the number of "1"in the column}}$$

  and then make pairwise combination for them. After these transformations, we can get C^2_k columns
- 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

Denote the transformed matrix as B which is a m×N matrix, then its covariance matrix is a m×n square matrix shown as:

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

Assumed that $S = [S_{ij}]$, then:

$$S_{ij} = \frac{\sum_{k-1}^{n} a_{ik}a_{jk}}{N-1}$$

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

The property of covariance matrix S:

- S is a symmetrical square matrix
- 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
- The minimal eigenvalue of S is 0
- 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: A bipartite network H' and its adjacency matrix:

$$A = \begin{bmatrix} 0 & A_1^T \\ A_1 & 0 \end{bmatrix}$$

Two sets of vertices in H', 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., $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ and the corresponding eigenvectors, namely $u_1, u_2, ..., u_n$
- 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 $l_2, ..., l_{n1-1}$ and then choosing $u_2, u_3, ...u_k$ to build a n×(k-1) matrix that is $u = [u_2, u_3, ...u_k]$
- Supposed the row vector of u are $p_1, p_2, ..., n_1$, namely:

$$u = \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ . \\ p_n \end{bmatrix}$$

  k-means algorithm is made use of clustering for $p_1, p_2, ..., p_n$
- Hence, the clustering for vertices can be obtained and from that the corresponding edges clustering also can be got.

End

EXPERIMENTAL RESULTS AND ANALYSIS

Here, we conduct a set of experiments to compare the performance of the CDA_PCA algorithm with other

typical algorithms. All the experiments were conducted on a 3.0 GHzPentium 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. At first, we extract a small subset from OMIM and construct a diseasome bipartite network named H as shown in Fig. 4 or 5. It can be seen that there are 19 hereditary diseases, 19 disease genes and linkages associated with them, where circles and rectangles correspond to disorders and disease genes, respectively. A link is placed between a disorder and a disease gene if mutations in that gene lead to the specific disorder. We apply our algorithm to community detection in diseasome 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 developed by Chen *et al.* (2009b) and the results are shown in Fig. 4 and 5. 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 the results obtained by Chen *et al.* (2009b) 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 notice that both of the results demonstrate that the disease occurrence is not isolated but relevant to many other diseases.

Southern women network: As the second experiment, we consider southern women network (Davis *et al.*, 1941) to verify the accuracy of CDA_PCA. This is because the network has been broadly analyzed by social network researchers and its community structure is known. Many researchers use this network as a touchstone for testing their community detection method (Barber, 2007; Guimera *et al.*, 2007; Murata and Ikeya, 2010; Suzuki and Wakita, 2009).

The community partitions obtained by our method and some other approaches are shown in Fig. 6-10. Women are indicated as circle symbols located at the

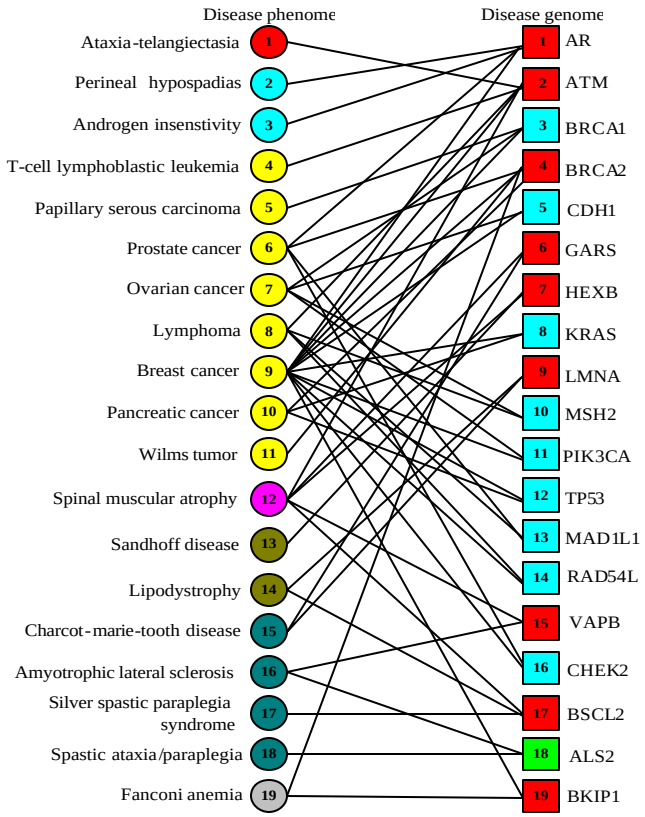

Fig. 4: Partitions of the disease-gene network H obtained by our method

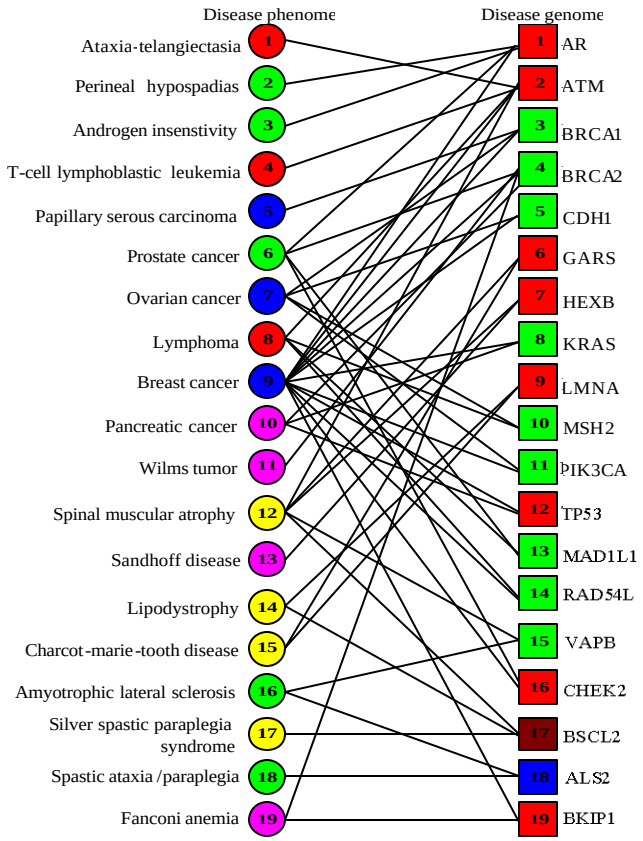

Fig. 5: Partitions of the disease-gene network H obtained by Chen *et al*. (2009b)

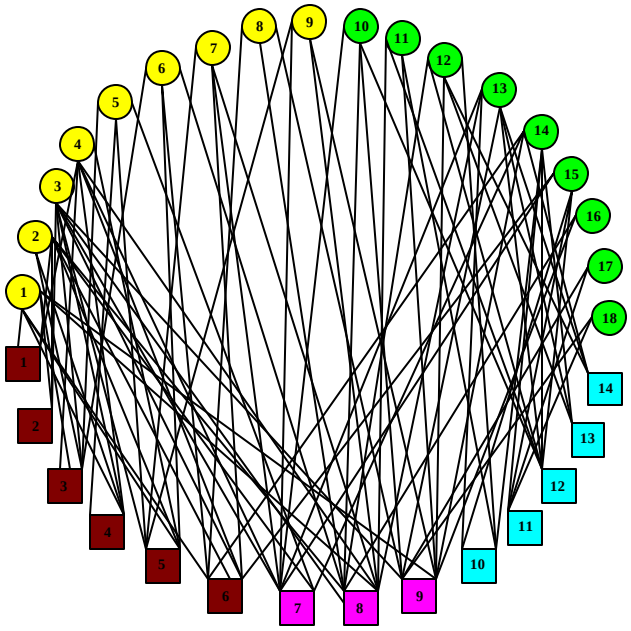

Fig. 6: Partitions of the Southern women network obtained by our method

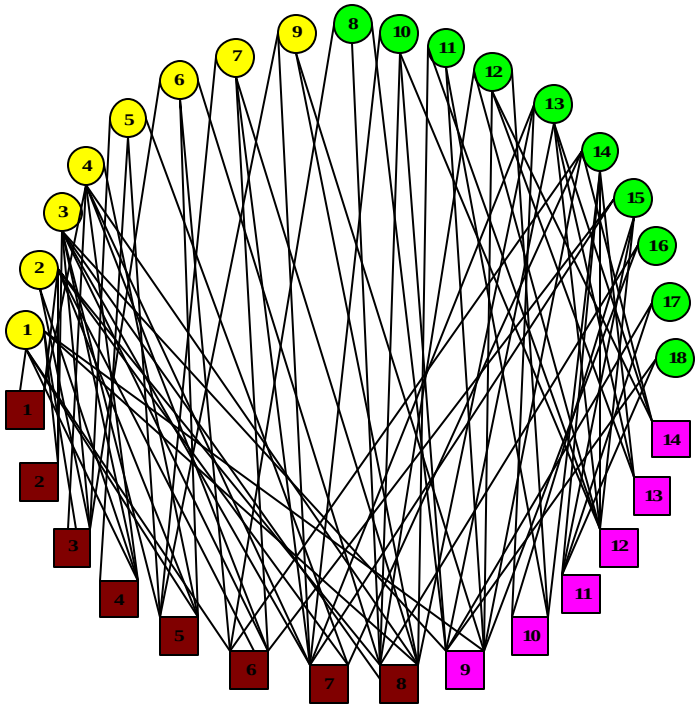

Fig. 7: Partitions of the Southern women network obtained by Guimera *et al*. (2007)

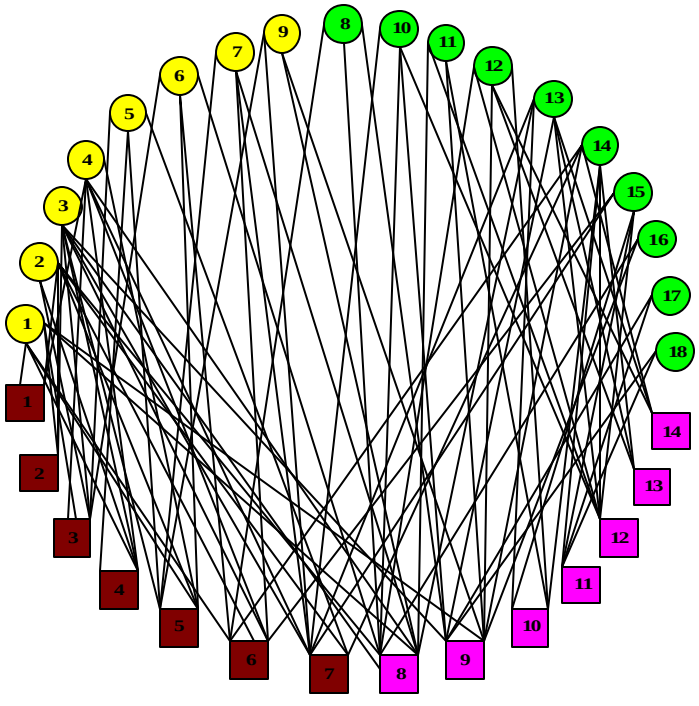

Fig. 8: Partitions of the disease-gene network obtained by Murata and Ikeya (2010)

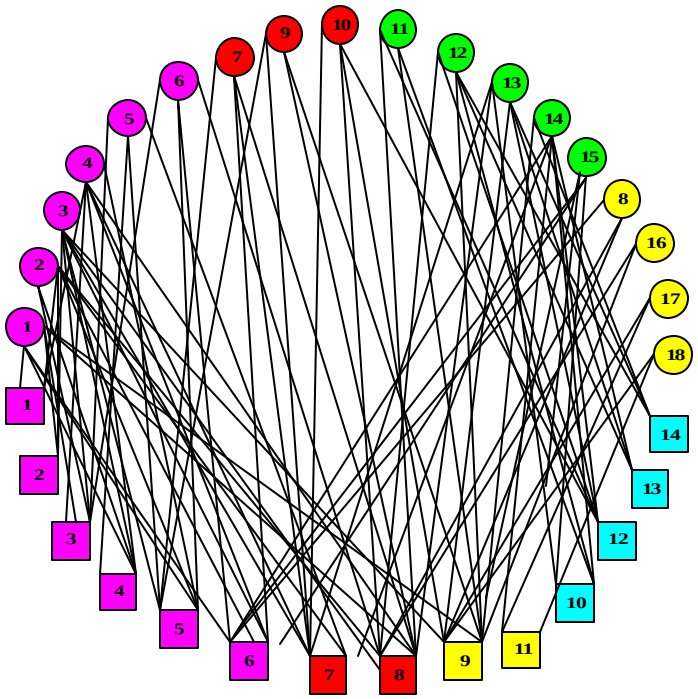

Fig. 9: Partitions of the disease-gene network obtained by Barber (2007)

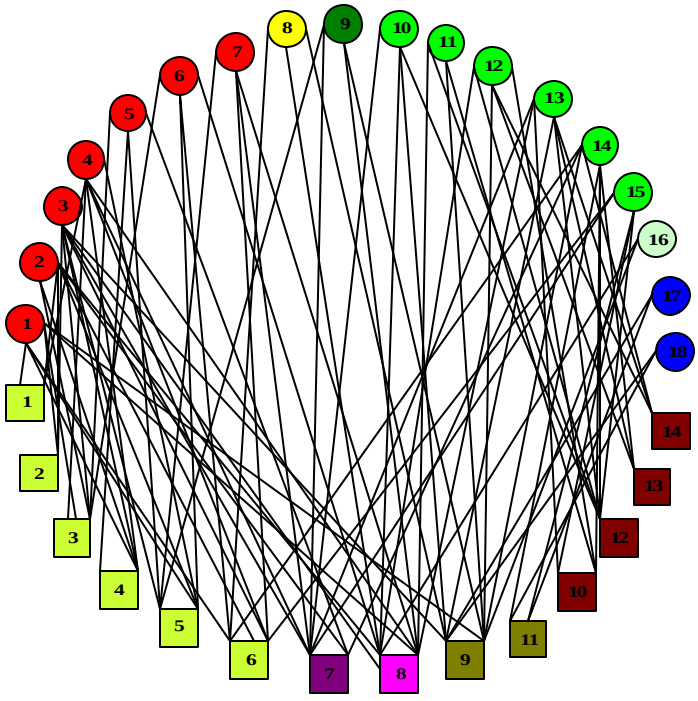

Fig. 10: Partitions of the disease-gene network obtained by Suzuki and Wakita (2009)

upper camber wihle events are indicated as square symbols located at the lower camber. Nodes in the same community are painted in the same color. 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 woman communities {1-9} and {10-18} wihle 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, Fig. 7-9 can only detect communities of one-to-one correspondence. Figure 10 can detect communities of many to many correspondence but this partition seems somewhat irrational, since several communities contain only one node.

CONCLUSION

In this study, we propose a novel community detection method CDA_PCA and triumphantly apply it to some real-world networks. We convert the origin bipartite network into the 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, Science and Technology Innovation Foundation of Yangzhou University under grant No.2012CXJ026, Chinese National Natural Science Foundation under grant No.61070047.

ACKNOWLEDGMENT

This research was supported in past by the Natural Science Foundation of Jiangsu Provinec under Grant No. BK20130452, Natural Science Foundation of Education Department of Jiangsu Provinec under Grant No. 12KJB520019, Science and Technology innovation foundation of Yangzhou University under grant No. 2012CXJ026, Chinese National Natural Science Foundation Under Grant No. 61070047.

REFERENCES

Barber, M.J., 2007. Modularity and community detection in bipartite networks. Phys. Rev. E, 76: 1-9.

Chen, D.B., Y. Fu and M.S. Shang, 2009. A fast and efficient heuristic algorithm for detecting community structures in complex networks. Phys. A, 388: 2741-2749.

Chen, W.Q., J.A. Lu and J. Liang, 2009. Research in disease-gene network based on bipartite network projection. Complex Syst. Comp. Sci., 6: 13-19.

Clauset, A., M.E.J. Newman and C. Moore, 2004. Finding community structure in very large networks Phys. Rev. E, 70: 066111-066116.

Costa, L.F., F.A. Rodrigues, G. Travieso and P.R.V. Boas, 2007. Characterization of complex networks: A survey of measurements. Adv. Phys., 56: 167-242.

Davis, A., B.B. Gardner and M.R. Gardner, 1941. Deep South. University of Chicago Press, USA.

Duch, J. and A. Arenas, 2005. Community detection in complex networks using extremal optimization. Phys. Rev. E, Vol. 72

Fortunato, S., 2010. Community detection in graphs. Phys. Rep., 486: 75-174.

Guimera, R., M. Sales-Pardo and L.A. Amaral, 2007. Module identification in bipartite and directed network. Phys. Rev. E, Vol. 76. 10.1103/PhysRevE.76.036102

Kernighan, B.W. and S. Lin, 1970. An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J., 49: 291-307.

Murata, T. and T. Ikeya, 2010. A new modularity for detecting one-to-many correspondence of communities in bipartite networks. Adv. Complex Syst., 13: 19-31.

Newman, M.E.J., 2004. Fast algorithm for detecting community structure in networks. Phys. Rev. E, Vol. 69. 10.1103/PhysRevE.69.066133

Newman, M.E.J., 2006. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev., Vol, E74,

Ruan, J. and W. Zhang, 2008. Identifying network communities with a high resolution. Phy. Rev., Vol. 77.

Suzuki, K. and K. Wakita, 2009. Extracting multi-facet community structure from bipartite networks. Proceedings of the International Conference on Computational Science and Engineering, Volume 4, August 29-31, 2009, Vancouver, BC., pp: 312-319.

Wang, X.T., G.R. Chen and H.T. Lu, 2007. A very fast algorithm for detecting community structures in complex networks. Phys. A, 384: 667-674.

Xiang, B., E.H. Chen and T. Zhou, 2009. Finding community structure based on subgraph similarity. Stud. Comput. Intell., 207: 73-81.