Ant Colony Optimization for Detecting Communities from Bipartite Network

Yongcheng Xu
Department of Computer Science, Yangzhou University, Yangzhou, 225009, China
Email: ychxu08@126.com
Ling Chen and Shengrong Zou
Department of Computer Science, Yangzhou University, Yangzhou, 225009, China
State Key Lab of Novel Software Tech, Nanjing University, Nanjing, 210093, China
Email: yzulchen@163.com

***Abstract*—In this paper, an algorithm based on ant colony optimization for community detection from bipartite networks is presented. The algorithm establishes a model graph for the ants' searching. Each ant chooses its path according to the pheromone and heuristic information on each edge to construct a solution. Experimental results show that our algorithm can not only accurately identify the number of communities of a network, but also obtain higher quality of community detection.**

***Index Terms*—ant colony optimization, community detection, bipartite network**

I. INTRODUCTION

Most of the real world networks [1], such as the networks in biology [2], economy [3], sociology [4], and other fields [5], typically have community structure, namely, networks can often be divided into communities. Within community, the connections are dense, but connections between nodes of different communities are much sparser [6]. Identifying and analyzing such communities from a large network provides a means for functional dissection of the network and sheds light on its organizational principles.

Bipartite network is an important category of complex networks. Many real-world networks are naturally bipartite, such as scientists-paper cooperation network [7, 8], the actors-films network [1, 9], investors-company network [10, 11], disease-gene network [12], club member-activities network [13], audience-songs network [14], and computer terminals-data networks in P2P system [15] and so on. Communities are relatively independent in the structure, and it is believed that each of them may correspond to some fundamental functional unit. For example, a community in genetic networks often contains genes with similar functions and a community on the World Wide Web may correspond to web pages related to similar topics. Detecting communities in such bipartite network serves an important practical purpose in finding similar vertices and for analyzing overall community structures.

II. RELATED WORKS

One approach for detecting community of the bipartite network is to project the bipartite network into unipartite networks. However, the projection of a bipartite graph generally loses some information of the original network,. Guimera et al. [16] proved that the analysis of a projection can give incorrect results for community detection and even affect the properties including the community structures of the networks. Barber [17] demonstrated differences between communities detected in a real world bipartite network and its projection.

In recent years most of the community detecting algorithms directly analyze the original bipartite network. X Liu and T. Murata [18] proposed an algorithm called LP&BRIM for community detection in large-scale bipartite networks. The algorithm is based on label propagation, and employs BRIM algorithm [17] for generating better community structure. Lehmann et al. [19] extend the *k*-clique community detection algorithm to bipartite networks. The algorithm retains all of the advantages of *k*-clique algorithm, but it avoids discarding important structural information when performing a one-mode projection of the network. N. Du et al. [20] proposed an algorithm BiTector to mine the overlapping communities in large scale sparse bipartite networks. Rut Jesus et al. [21] investigate the bipartite network of articles linked by common editors in Wikipedia. They detect overlapping cliques of densely connected articles and editors, and cluster these densely connected cliques into larger communities of editors' flock around articles driven by interest. Peng Zhang et al. [22] defined a clustering coefficient for bipartite networks, and presented an algorithm which can detect communities by cutting the edges with the least clustering coefficient. Yajing Wu et al. [23] proposed a clustering algorithm on bipartite network. By analyzing the resource allocation on bipartite network, the algorithm uses fuzzy clustering method and F statistic to identify the best community structure. Wang Yang et al. [24] defined the attractive force between the node and the community in bipartite network. By analyzing the attractive force, they also presented a community detecting algorithm for bipartite network. Patrick Doreian et al. [25] extended relaxed

doi:10.4304/jsw.8.11.2930-2935

structural balance approach to large signed bipartite networks.

In this paper, a new algorithm for community detection for bipartite networks based on ant colony optimization [26] is presented. The algorithm transforms the problem of detecting communities into a combinatorial optimization problem. In the algorithm, we first define a model graph according to the bipartite network. The ants search the solution on such model graph. We define heuristic information on the basis of the degree of vertexes. In the model graph, each ant selects its path depending on the pheromone and heuristic information on each edge to construct a solution. The quality of solution obtained by each ant is measured by its bipartite modularity. Our algorithm can not only accurately identify the number of communities of a network, but also obtain higher quality of community partitioning.

III. Bipartite Modularity

To evaluate the quality of a particular division of a network into communities, modularity is broadly applied. A widely used and quite successful method for the identification of communities in unipartite networks is maximization of a modularity function. In recent years, modularity measures which can be applied to identify communities in bipartite networks are proposed. Guimera et al. [16] generalize the Newman's modularity metric [27] to the bipartite networks. Barber [17] defines the bipartite modularity matrix B as an extension of Newman's another work [28]. Murata [29] proposes a bipartite modularity, which gives consistent result as Newman's modularity when applied to unipartite networks. Wakita and Suzuki [30] advance a modified version of Murata's bipartite modularity, which can reflect the multi-facet correspondence among communities. Modularity has been applied as the quality function in many bipartite community detection algorithms. In this paper, we adopt the bipartite modularity proposed by Murata [29] to measure the quality of bipartite network community detection.

A bipartite network can be represented by a bipartite graph $G=(V, E)$, where $V = V^X \cup V^Y$. Here, V^X and V^Y are the two sets of vertices in G. E is the set of its edges. Let M be the number of edges in a bipartite network. Suppose the bipartite network is partitioned into X-vertex communities and Y-vertex communities, and the numbers of the communities are L^X and L^Y, respectively. Let $C^X = \{V_1^X, \dots V_{L^X}^X\}$ and $C^Y = \{V_1^Y, \dots V_{L^Y}^Y\}$ be the sets of the communities of X-vertices and Y-vertices respectively, A be the adjacency matrix of the network. Suppose $V_l \in C^X$ and $V_m \in C^Y$ are two communities, since the vertices in V_l and V_m are of different types, we can define e_{lm} and a_l as follows:

$$e_{lm} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V_m} A(i,j) \tag{1}$$

It can easily be seen that e_{lm} is the the fraction of all edges in the network that connect vertices in community V_l to vertices in community V_m. We further define a $k \times k$ symmetric matrix E composed of e_{lm} as its (l,m) element, and its row sums a_l:

$$a_l = \sum_m e_{lm} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V} A(i,j) \tag{2}$$

Then Murata's bipartite modularity Q_B is defined as follows:

$$Q_B = \sum_l Q_{B_l} = \sum_l (e_{lm} - a_l a_m),\ m = \underset{k}{argmax}(e_{lk}) \tag{3}$$

Here, Q_{B_l} is the deviation of the number of edges that connect the l-th X-vertex community and the corresponding (m-th) Y-vertex community, from the expected number of randomly-connected edges. A larger Q_{B_l} value means stronger correspondence between the l-th community and the m-th. Larger value of bipartite modularity Q_B indicates higher quality of a community division.

IV. Framework of Our Algorithm

A. Structure of a Solution

In order to utilize ant colony optimization for community detecting on bipartite networks, we construct a model graph, on which the ants search for the optimal solution. Suppose the numbers of X-vertices and Y-vertices in the bipartite network are n and m respectively. We label X-vertices with integers 1 to n, and label Y-vertices with n+1 to n+m, and $V^X = \{V_1, V_2, \dots, V_n\}$ is the set of X-vertices and $V^Y = \{V_{n+1}, V_{n+2}, \dots, V_{n+m}\}$ is the set of Y-vertices. Then the model graph of ants foraging is a directed graph as shown in Fig. 1.

In the directed model graph, there are n+m+1 nodes $u_1, u_2, \dots, u_{n+m+1}$, which represent vertices of bipartite network except the last one indicting the end of ants' foraging. Between each pair of neighboring nodes, there are n+m directed edges. Let the set of directed edges between nodes V_i and V_{i+1} is $E_i = \{E_{i\,1}, E_{i\,2}, \dots, E_{i\,n}, E_{i\,n+1}, \dots, E_{i\,n+m}\}$. Each ant chooses its path according to the pheromone and heuristic information on each edge. If an ant arrives at node u_i and chooses the edge $E_{i\,k}$, it means that the nodes V_k and V_i in the bipartite network are assigned into the same community. We use a vector $S = (S_1, S_2, \dots, S_n, S_{n+1}, \dots, S_{m+n})$ to denote the solution the ant constructed. If the value of component S_i is k, node v_i and node v_k in the bipartite network are in the same temporary community. In the algorithm, each ant selects a component S_i for every node V_i, so as to construct the solution vector S.

After constructing the vector S, we can merge the temporary communities to obtain the final result. If two temporary communities have common nodes, then we merge them into a larger temporary community. Repeating this process until there is no temporary communities can be merged. At this stage, the merged temporary communities are the communities in the final result.

For example, Fig. 2 (a) is a bipartite network, where the blue rectangles and the red circles represent two different types of nodes. There is no edge between the same types of nodes. We label these nodes as shown in Fig. 2 (b). Suppose an ant constructs the solution S = (5, 1, 4, 3, 1, 2, 3, 7) from the bipartite network in Fig. 2. Then, the temporary communities in this solution are as shown in Fig. 3, where each column indicates a temporary community. For instance, the first column in Fig. 3 is (1, 5), which means nodes v_1 and v_5 form a temporary community.

From Fig. 3, we can see that the eight temporary communities are (1, 5),(2, 1),(3, 4),(4, 3),(5, 1),(6, 2),(7, 3),(8, 7). Merging (1, 5) and (2, 1), we get another temporary sub-community (1, 2, 5). Repeating the process of merging until we obtain the final solution consisting of two communities: (1, 2, 5, 6), and (3, 4, 7, 8).

B. Implementation of the Algorithm

1). Pheromone

To construct an effective solution, pheromone information τ_{ij} is assigned on each path E_{ij} . The pheromone information influences the choices the ants in their searching. The larger amount of pheromone on an edge deposit, the higher probability an ant will select this edge. The intensity of pheromone on the edge is updated after each iteration. The more ants with high quality solutions pass an edge, the greater increment of pheromone will be offered on this edge. After each iteration the pheromone on each edge will be decreased by an evaporation rate. Communications and cooperation between individual ants by pheromone information enable the ant colony algorithm to have strong capability of finding the best solutions. In our algorithm, when all ants get their solutions after each iteration, then we update the pheromone on each edge using the following pheromone updating formula:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^k \tag{4}$$

Here, ρ is the evaporation rate, m is the number of ants, and $\Delta\tau_{ij}^k$ is the increment of pheromone laid on edge (i, j) by the k-th ant:

$$\Delta\tau_{ij}^k = \begin{cases} Q_B(S) & if\ V_j \in S \\ 0 & otherwise \end{cases} \tag{5}$$

where $Q_B(S)$ is the bipartite modularity of division communities constructed by the k-th ant.

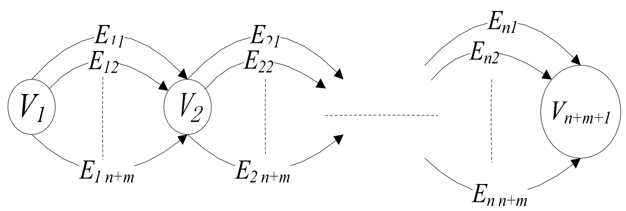

Figure 1. The Model Graph of Ants Searching

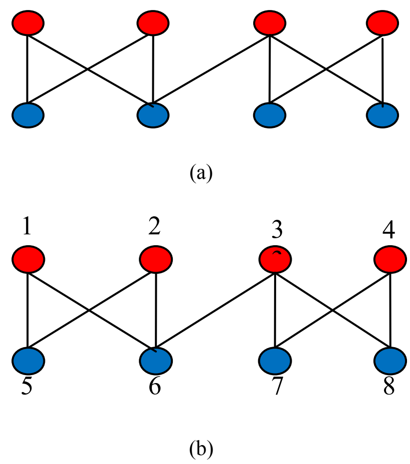

Figure 2. An Example of a Division of a Bipartite Network. (a) An bipartite network. (b) The numbering bipartite network.

1	2	3	4	5	6	7	8
5	1	4	3	1	2	3	7

Figure 3. Temporary communities of a solution

2). Heuristic information

In the algorithm, we also define the heuristic information η_{ij} to reflect the potential tendency for the ants to select the edge E_{ij} in the directed model graph. In the bipartite network, if two nodes v_i and v_k are both directly or indirectly connected with larger number of other nodes, they should be in the same community with a higher probability, and the heuristic function η_{ij} on edge E_{ij} will be assigned higher value. The value of heuristic function η_{ij} will be determined as follows:

(1) If both nodes v_i and v_j in the bipartite network are belong to the same type of vertices, namely, $i\ ,j \geq n$ or $i,j<n$, then:

$$\eta_{ij} = \frac{2C_{ij}}{d_i + d_j} \tag{6}$$

Here, d_i is the degree of node v_i, and C_{ij} is the number of nodes connected with both v_i and v_j.

(2) If nodes v_i and v_j in the bipartite network are belong to the different type vertices, namely, one of i and j is less than or equal to n, and the other is greater than n, then:

$$\eta_{ij} = \frac{1}{d_i + d_j}\left(\sum_{\substack{k\in V^X \\ k\neq i}} \frac{C_{ik}}{d_i + d_k - C_{ik}} + \sum_{\substack{g\in V^X \\ g\neq j}} \frac{C_{jg}}{d_j + d_g - C_{jg}}\right) \tag{7}$$

3). Transition probability

When constructing the solutions, the ants traverse on a model graph and select a directed edge at each node by a probabilistic decision. The transition probability for the k-th ant at the node of u_i choosing the path E_{ij} is given by:

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}.\eta_{ij}^{\beta}}{\sum_{g=1}^{n+m} \tau_{ij}^{\alpha}.\eta_{ij}^{\beta}} & \text{if } i \neq j \\ 0 & otherwise \end{cases} \quad (8)$$

Here, α and β are the parameters which control the relative importance of the pheromone and the heuristic information. If α assigned a larger value than β, the pheromone will has greater influence on the ants' searching, otherwise the heuristic information will have greater influence.

C. Framework of the Algorithm

Suppose the number of X-vertices in the bipartite network is n and the number of Y-vertices is m, we label these vertices from 1 to $n+m$. The framework of our algorithm ACOCD (ant colony optimization community detecting) is described in Fig. 5.

V. Experiment Results

To test the effectiveness and performance of our proposed algorithm ACOCD, we test it by a series of experiments. We compared the ACOCD algorithm to Brim algorithm [17], LP&BRIM algorithm [18] and Davis method [31]. The algorithms are coded using Java, and all the experiment studies are conducted on Corei3, Windows7 environment.

A. Southern Women network

To verify the accuracy of our algorithm, we use southern women network [38] collected by Davis et al. around Mississippi during the 1930s as part of an extensive study of class and race in the Deep South. As shown in Fig. 4, the network describes the participation of 18 women in 14 social events. If a woman attended an event, there will is an edge linking their nodes.

Firstly, we label the nodes of 18 women as v1 to v18, and the nodes of 14 events as v19 to v32 as shown in Fig. 4. We use our ACOCD algorithm to detect the communities on southern women network, and the experimental results are shown in Fig. 5. From the Fig. 5, we can see that two communities are obtained: {woman 1-9, event 19-26} and {woman 10-18, event 27-32}. The bipartite modularity of the result is 0.585964.

Murata propose LP&BRIM algorithm and also obtain two communities: {woman 1-7, 9, event 19-25} and {woman 8, 10-18, event 26-32} [25]. Davis, who collected the network, has ever used ethnographic knowledge to divide women into two communities: {women 1-9} and {women 9-19} (Woman 9 is a secondary member of both communities) [31]. In order to comparison, we must assign this individual to a specific group. Two partitioning schemes are implemented accordingly. One scheme, which is named Davis1, merges the 9th woman with women 1-8. The other scheme, which is named Davis2, merges the 9th woman with women 10-18. Comparison of bipartite modularity by different algorithms using southern women network is shown in Fig. 6.

From Fig. 6, we can see that our algorithm ACODC can get the highest modularity among all the methods. Noticing that, if we only look at women nodes, the final community division by our algorithm agrees with the one proposed by Davis. Since our algorithm does not require a predefined number of communities, and gets the results with the highest modularity, it can obtain higher quality of community partitioning without previously known parameters.

Algorithm ACOCD (*G*, *A*, *Q*)
Input: *G*: the bipartite network;
A: the adjacency matrix of *G*;
Output: S_{best}: solution community division;
Q_{best}: the bipartite modularity of the solution
Begin
Initialize the various parameters;
Initialize values of pheromone and heuristic information;
While not the terminate condition **do**
For k=1 **to** k **do** /*k ants*/
For i=1 **to** $m+n$ **do** /*$m+n$ nodes*/
Ant k selects S_i according to probability (9);
Endfor i;
Calculate the modularity Q_B of solution S;
If $Q_B > Q_{best}$ **then**
$Q_{best} = Q_B$; $S_{best} = S$;
/* Q_{best} is the highest modularity obtained so far;
S_{best} is the best solution obtained so far */
endif
Endfor k;
Update the pheromone on the edges;
End While;
End

Figure 4. Framework of algorithm ACOCD

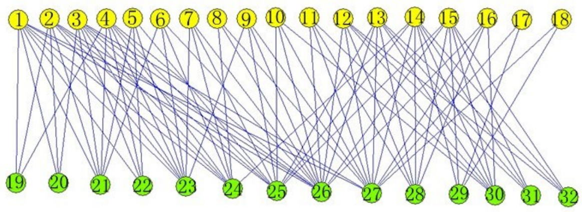

Figure 5. The numbering bipartite network. 18 yellow circles are 18 women, and 16 green circles are 16 events.

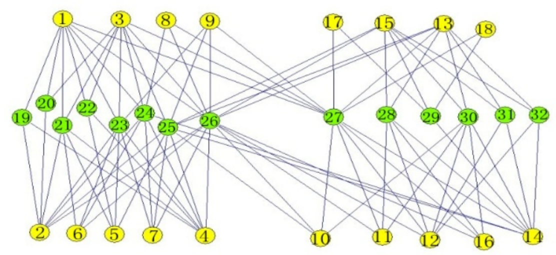

Figure 6. Final division obtained by our algorithm

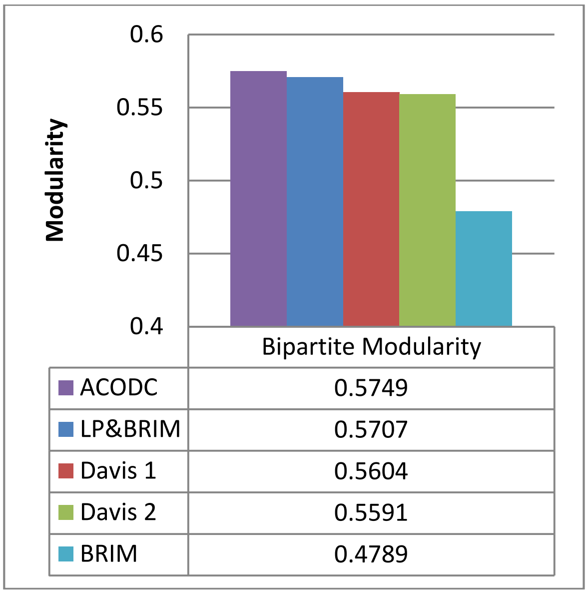

Figure 7. Comparison of bipartite modularity by the different algorithms using southern women network

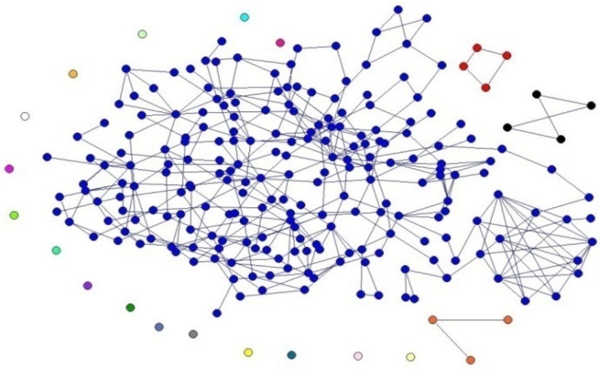

Figure 8. The network of Scotland. The largest component is composed of only the blue nodes. The nodes of the other colors are outliers.

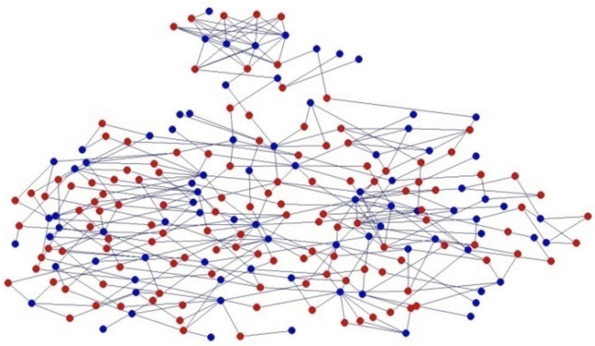

Figure 9. The largest component of the network of Scotland. The blue nodes are represented as 86 firms, and the red nodes are represented as 131 directors.

B. Scotland Corporate Interlock

As a second example, we also test on a network of corporate interlocks in Scotland in the early twentieth century [32]. The data set characterizes 108 Scottish firms during 1904-1905, detailing the corporate sector, capital, and board of directors for each firm. The dataset includes only those board members who held multiple directorships, totaling 136 individuals as shown in Fig. 7.

Here, we focus on the bipartite network of firms and directors, with edges existing between each firm and its board members. Unlike the Southern Women network, the Scotland corporate interlock network is not connected. We conduct two experiments on this bipartite network: one takes the whole dataset as the testing data, and the other uses only the largest component of the network containing 131 directors and 86 firms as shown in Fig. 8.

In the first experiment, the algorithm ACOCD divides the network into 49 communities, and gets a bipartite modularity 0.608412. By careful observation of the 49 communities, we find that there are 19 communities consisting of only one or more outlier nodes. In the second experiment which excludes the outliers, 30 communities are obtained by our algorithm ACOCD with a bipartite modularity 0.625256. Comparing of two cases, we find that the results obtained by the two experiments are almost identical except the outliers, and gets high bipartite modularity.

We also compare the result of our algorithm ACOCD with that of BRIM algorithm [17] by Baber. Baber's declares that if restricting the number of communities being less than thirty, BRIM algorithm can get the maximum value of modularity. The number of communities obtained by our algorithm ACOCD approximately agrees with Baber's conclusion, the community detecting results and their modularity are very close. But in contrast to BRIM algorithm, our algorithm can straightly obtain the number of communities and the specific division without any prior knowledge of the network.

VI. Conclusion

An algorithm for detecting communities from bipartite networks based on ant colony optimization is presented. The algorithm firstly transforms the problem of community detection into the one of combination optimization, and establishes a model graph for the ants' searching. Meanwhile we define heuristic information according the topological structure of the network. Each ant chooses its path according to the pheromone and heuristic information on each edge to construct a solution. The quality of solution obtained by each ant is measured by its bipartite modularity. Experiment results show that our algorithm can not only accurately identify the number of communities and true community structure from bipartite network, but also obtain higher quality of community division. Community detection is still an open and challenging problem, we will work to make further developments in this area.

Acknowledgements

This research was supported in part by the Chinese National Natural Science Foundation under grant Nos. 61070047, 61070133 and 61003180, State Key Fundamentals Research (973) Project under contract 2012CB316003.

REFERENCES

[1] D. Watts and S. Strogatz. Collective dynamics of small-world networks. Nature, 393(6684):440–442, June 1998.

[2] A. -L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell's functional organization Nature Reviews Genetics, vol. 5, pp. 101-113, 2004.

[3] N. Economides, “The economics of networks,” International Journal of Industrial Organization, vol. 16, no. 4, pp. 673-699, October 1996.

[4] L. C. Freeman, The Development of Social Network Analysis: A Study in the Sociology of Science. Vancouver, BC: Empirical Press, 2004.

[5] A. -L. Barabási, Linked: The New Science of Networks. Cambride, MA: Perseus Publishing, 2002.

[6] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences of the United States of America, vol. 99, Jun., 2002, pp. 7821-7826.

[7] Newman M E J. Scientific collaboration networks. I. network construction and fundamental results [J]. Physical Review E, 2001, 64: 016131.

[8] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [J]. Physical Review E, 2001, 64: 016132.

[9] LIU Ai-fen, FU Chun-hua, ZHANG Zeng-ping, CHANG Hui, HE Da-ren, An Empirical Statistical Investigation on Chinese Mainland Movie Network, Complex Systems and Complexity Science,2007,4(3): 10-16.

[10] Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs [J]. Computational & Mathematical organization Theory, 2004, 10: 69-94.

[11] Battiston S, Catanzaro M. Statistical properties of corporate board and director networks [J]. European Physics Journal B, 2004,38: 345-352.

[12] CHEN Wen-qin, LU Jun-an, LIANG Jia, Research in Disease-Gene Network Based on Bipartite Network Projection, 6(1), 2009, 13-19.

[13] Ergun G, Human sexual contact network as a bipartite graph[J]. Physica A, 2002,308: 483-488.

[14] Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite networks [J]. Physical Review E,2005, 72: 066107.

[15] Le Blond S, Guillaume J L, LatapyM. C lustering in P2P exchanges and consequences on performances [C]. CastroM, Renesse R. Peer- to-Peer Systems IV. Berlin: Heidelberg, 2005, 193- 204.

[16] Guimera R, Sales-Pardo M, Amaral L A. Module identification in bipartite and directed network [J]. Physical Review E, 2007, 76:036102.

[17] M. J. Barber, “Modularity and community detection in bipartite networks,” Physical Review E, vol. 76, no. 066102, pp. 1–9, 2007.

[18] Xin Liu, Tsuyoshi Murata, Community Detection in Large-scale Bipartite Networks, 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp 50-57.

[19] S. Lehmann, M. Schwartz, and L. K Hansen, “Biclique communities,” Phys. Rev. E, vol. 78, no. 1, 016108, 2008.

[20] N Du,B Wang, B Wu, et al. Overlapping Community Detection in Bipartite Networks[C].WI. 2008: 176-179.

[21] Rut Jesus, Martin Schwartz, Sune Lehmann, Bipartite Networks of Wikipedia’s Articles and Authors: a Meso-level Approach, WikiSym '09, October 25-27, 2009, Orlando, Florida, U.S.A.

[22] Zhang P, Wang J L, Li X J, et al. Clustering coefficient and community structure of bipartite networks [J]. Physica A, 2008,387: 6869-6875.

[23] WU Yajing DI Zengru FAN Ying, A Clustering Algorithm for Bipartite network based on distribution matrix of resources, Journal of Beijing Normal University(Natural Science), 2010, 46(5): 643-646.

[24] WANG Yang, DI Zeng-ru, FAN Ying, Comparative Definition of Community in Bipartite Network, Complex Systems and Complexity Science, 2009. 6(4): 40-44.

[25] Patrick Doreian, Paulette Lloyd, Andrej Mrvar, Partitioning large signed two-mode networks: Problems and prospects, Social Networks (2012) (In Press).

[26] Dorigo, M., Maniezzo, V., and Colorni, A., Ant System: Optimization by a colony of cooperating agents, IEEE Trans. Syst., Man, and Cybern. { PartB, 26(1), 29, 1996.

[27] M. Newman. Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, 2004.

[28] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

[29] T. Murata, “Detecting communities from bipartite networks based on bipartite modularities”, Proceedings of the 2009 International Conference on Computational Science and Engineering, pp.50-57.

[30] K. Wakita, and K. Suzuki, “Extracting multi-facet community structure from bipartite networks,” Proceedings of the International Symposium on Social Intelligence and Networking (SIN 09), Aug. 2009, in press.

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

[32] J. Scott and M. Hughes, The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital, 1900–1979, CroomHelm, London, 1980.