Community Division of Bipartite Network Based on Information Transfer Probability

Chunlong Fan(✉), Hengchao Wu, and Chi Zhang

College of Computer, Shenyang Aerospace University, Shenyang 110136, China
754713151@qq.com

Abstract. Bipartite network is a performance of complex networks,The divided of unilateral node of bipartite network has important practical significance for the study of complex networks of community division. Based on the diffusion probability of information and modules ideas in the network,this paper presents a community divided clustering algorithm (IPS algorithm) for bipartite network unilateral nodes.The algorithm simulates the probability of information transfer in the network,through mutual support value between the nodes in network,selecting the max value as the basis for merger different communities.Follow the module of the definition for division after mapping the bipartite network nodes as a single department unilateral network.Finally,we use actual network test the performance of the algorithm.Experimental results show that,the algorithm can not only accurate divided the unilateral node of bipartite network,But also can get high quality community division.

Keywords: Modularity · Bipartite network · Support value · Community division

1 Introduction

Complex networks is a rise in recent years involving physics, biology, mathematics and computer science and other fields of interdisciplinary research community structure for roughly classified into three types: Community found that evolutionary analysis community, and the community structure and network dynamics and network compression indicate other relations between functional features. Where the research community found mostly concentrated in a single unit network, the relative lack of research bipartite network of community division. Community research division of unilateral node bipartite network, focusing only on the use of two sub-networks to build a single unit network, using sophisticated algorithms have been divided communities. We can use the bipartite network partitioning algorithm community as a class of two types of processing nodes, perform community is divided and then extracted unilateral community nodes, such as the reality of bipartite network Scientist - Research Collaboration Network [2-3] of scientists collaborative research, movie - actor cooperative research actors network [4] actor, disease - gene networks [5] disease relevance, audience - songs network [6] audience community and computer terminals - P2P data network terminal groups [7].

M. Núñez et al. (Eds.): ICCCI 2015, Part I, LNAI 9329, pp. 133–143, 2015.
DOI: 10.1007/978-3-319-24069-5_13

2 Current Research Presentation

The bipartite networks are mapped into a single network using relatively conventional, relatively mature single network of community division algorithm is divided into bipartite networks of community common treatment. But whether it is entitled to a weighted projection or projection, each node can only be divided unilateral and projected results will lead to lack of bipartite network information. The division directly on the original two points of the network, often dichotomous types of nodes in the network as a class for processing, so far no one can recognize bipartite network community divided evaluation criteria. Therefore there is a need for further study two points online community divided.

According to the characteristics bipartite network, such as clustering coefficient, edge betweenness, degree and degree distribution, the number of overlapping topological potential, all kinds of bipartite network partitioning algorithm communities have been proposed. Yongcheng Xu et al. [8] proposed a bipartite network community mining transformed into an ant search graph model for the optimization problem, the algorithm based on the definition of the vertex topology heuristic information, construct a community divided result; but ants swarm algorithm uses random assignment process vertex belongs to the community with a large degree of randomness, so the whole algorithm run time overhead is relatively large. Chenbo Lun et al. [9] through the bipartite network diagram that corresponds to the matrix recursively split application matrix decomposition methods were divided community, also known as MP algorithm; but every time you want to split the result matrix remained intact Community Information very difficult, so the results may vary depending on the divided matrix decomposition process is different.Yajing Wu et al. [10] using clustering method original bipartite network, the bipartite network resource distribution matrix and fuzzy clustering method for vector combining clustering proposed determination F statistic most clustering methods, the algorithm The disadvantage is the need to know in advance the number of associations to be subdivided; do not know the number of associations in the case, this method is difficult to get an accurate division of the community. Gao et al. [11] Newman algorithm based on the idea of BRIM fast algorithm is proposed to improve the degree of aggregation algorithm module (MAB), although compared to the BRIM algorithm does not require additional number given community, but the algorithm complexity than BRIM Many high algorithms. Foreign studies have associated Raghavan et al. [12] Reference numeral propagation method (LPA), the algorithm first to each vertex is assigned a unique label, after selection of each vertex adjacent to them as labels, and finally with the same iterative vertex labels form a single community, but the algorithm is only suitable for small networks. Murata et al. [13] The LPA algorithm has been improved to the expense of accuracy obtained LPA algorithm suitable for large bipartite network and a large network of parallel real dichotomy community analysis. Italian scholars Dorigo et al. [14] proposed ant colony optimization algorithm has been successfully applied to multi combinatorial optimization problem, the experimental results show that the algorithm is universal, global, distributed computing, and robustness advantages; the same because the algorithm the combination is more, the algorithm complexity is still very high.

Newman Barber expand the definition of a single sub-network modularity presents a dichotomy modularity [15], combined with Adaptive maximization algorithm BRIM dichotomy modularity to get to divide the community, the algorithm can achieve better classification results. To get the optimal solution to try to go through the exchange community nodes defined in advance the number of community bipartite network. Therefore, the higher complexity of the algorithm, application limitations, does not apply to large-scale networks.

Based on information diffusion probabilistic thinking, combined with the definition of Newman et al modularity [16] and two points mapped network nodes unilateral proposed community bipartite network partitioning algorithm a unilateral nodes, the algorithm first two points of a network side of each vertex seen as a community, on both sides of the node distribution information to initialize. Six times the probability of information in accordance with the small world theory of six degrees of separation to get a rear diffuser exchange information resource matrix. Then the matrix values as a basis to judge the merger communities. After the last node will be unilateral projection mapping, combined with Newman's definition of modularity, as community standards division. Disconnect the combined maximum value at the Q to get the final result is divided community. Experimental results show that the algorithm can not only get an accurate number of communities, but also can get high-quality results divided communities.

3 Information Diffusion Probability (IPS) Model

Bipartite network can be represented as a bipartite graph G=(U,V,E), G, the nodes are divided into two parts, U and V, E is the edge of graph G . No edge is connected between the set U (or V) nodes, any edge of set E (u_i, v_j) , there must be $u_i \in U, v_j \in V$. Set up a collection of nodes in U is m, the number of nodes in the set V is n, two adjacency matrix can be expressed as:

$$\widetilde{A} = \begin{bmatrix} 0_{m\times m} & A_{m\times n} \\ A^T_{n\times m} & 0_{n\times n} \end{bmatrix} \tag{1}$$

Where matrix A is the sub-matrix of bipartite adjacency matrix , We call that matrices A is the relationship matrixof bipartite graph G , the relationship matrix A are:

$$A_{ij} = \begin{cases} 1 & \text{if} \quad i\ connect\ j \quad i \in U, j \in V \\ 0 & else \end{cases} \tag{2}$$

Each row of the matrix A is the connected case of a node of Set U,each row is the connected case of a node of Set V.Initialize set U in each node an information unit, the amount of information, given the set V of units of information in each node 0. To obtain set U of an initial unit information matrix $I^U_{m\times m}$:

$$I_{ij}^{U}=\begin{cases}1 & if \ \ i=j\\ 0 & else\end{cases} \tag{3}$$

Set U of the unit amount of information obtained as above matrix equation (3), each row of the unit matrix or the amount of information in each column represents a node in the set U,U has set the amount of information of the number of other nodes, initially only with its own node one unit of information, at this time it can be represented by a vector $\vec{x_i}$ owned by the node set U which is the amount of information to other nodes, the value of the information element value vector i matrix row or column i. Each node has a set amount of information $\vec{x_i}$, in accordance with rules such as the diffusion equation in equation (4).

$$\dot{R}_i=\begin{cases}\frac{1}{k_i}\vec{x_i} & if \ \ i \ connect \ j\\ 0 & else\end{cases} \tag{4}$$

Where k_i is the set of nodes of degree U, if there is an edge node u_i,v_j connecting, in this case the information of node u_i diffuse into the node v_j, each node vector v_j information obtained which are summed, the amount of information $R_i=\sum_{i=1}^{m}R_i^{'}$ obtained from set U in each node diffuse over , the combination of R_i of diffusion can be left matrix $R_{m\times n}$, in the node information set U diffusion matrix set V of nodes is expressed as:

$$S_{m\times n}=I_{m\times m}^{U}\bullet R_{m\times n} \tag{5}$$

Each row where $R_{m\times n}$ represents a node in set U diffusion to the collection of information for each node V, each column represents a set V of set U in node which receives information from each node. In this case each node in the set V is obtained from the amount of information for each node in the set U.

Finally, the collection information vector V on each node, according to the formula 6 and then spread to the set U in each node.

$$T_j=\begin{cases}\frac{1}{k_j}\vec{y_j} & if \ \ i \ connect \ j\\ 0 & else\end{cases} \tag{6}$$

Node $\vec{y_j}$ which has vector v_j information from each node in the set U, k_j is the degree of the node v_j. Combination of vector T_j information to get a n*m right information diffusion matrix $T_{n\times m}$, at this time, each node of set U itself diffuse to the collection of information of matrix U in the amount of information to other nodes that can be expressed as:

$$S_{m\times m}^{1}=I_{m\times m}^{U}\bullet R_{m\times n}\bullet T_{n\times m} \tag{7}$$

the elements R_{ij} of the matrix $R_{m\times n}$ can be obtained by corresponding elements of A_{ij} divided by relation matrix A where the degree of node i gived, $T_{n\times m}$ divided by the transposed matrix A^T may be the relationship of each element of the matrix A corresponding to A_{ij}^T degree of node i to give and therefore, the above-described steps n times the cycle information obtained after diffusion:

$$S_{m\times m}^n = (I_{m\times m}^U \bullet R_{m\times n} \bullet T_{n\times m})^n \tag{8}$$

Where S_{ij}^n represents the diffusion of information after the loop n times, node u_i to node u_j of the amount of information collected.

4 Select Determination Principle

4.1 Merge Determination Principle

Defines the degree of mutual support for the information size value of the diffusion matrix, when the information in the network for circulating diffusion times, In matrix S^n corresponding to the size of the value S_{ij} which represents the link between the two nodes tightness. Six Degrees principle states that the average distance between people to 6, after the information in the network will be able to meet the six diffusion determination result, therefore, the information diffusion cycle number n is 6.

Information Diffusion probabilistic algorithms of this paper is clustering algorithm, initialization set U of each node in a community, the diffusion of information on the diagonal matrix elements in addition to the traverse, select the elements in each row or column of the matrix S^6 of the maximum corresponding to merge the two communities. When communities merge, update the corresponding matrix S^6 corresponding i-th row and j-th row, the updated strategy for the community i or j is the largest value of other community support as a community updated external support, and delete information diffusion matrix j rows. Continue to traverse the updated next row until all communities merged into one community, then get the result of the merger of tree tree.

4.2 Principles of Defining Choice

The results of Single network and the bipartite networks division is determined by the merits of the community that have a lot of evaluation criteria, such as the clustering coefficient, modules, and so on. But there are two sub-types of the demarcation of the online community will be seen as the same type of node node treat, so this method is not applicable in this paper unilateral dichotomy network nodes in the network community is divided on the clustering. Bipartite network entitled to draw projection mapping and evaluation criteria divided community of single points on the network, so the community is divided on this paper, the results of Newman quality evaluation standard modules of the Q value [16] is an effective evaluation methods.

Physical meaning of modularity are: using a network connection belongs to the same community proportion edge node minus the proportion of community structures in the same random edges connecting these two nodes expectations. For obvious network community Q value is between 0.3 and 0.7. Module is defined as:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \| e^2 \| \tag{9}$$

Where e_{ii} represents the internal side of the community share in the proportion of all edges, which represents the ratio of the i-th community connected to the side edges of the proportion of all, $\| x \|$ represents all elements of the matrix and $Tre = \sum_i e_{ii}$ represents the matrix of and each element of the corner line.

4.3 Information Diffusion Example

For bipartite network shown in Figure 1, we study the situation on the community side of the divide nodes, for example, first get the dichotomous relationship matrix network A and A^T, according to the formula wherein R is a matrix of four changes from the information diffusion left matrix, the matrix T is changes come right information diffusion matrix A^T. After the diffusion node information to the upper edge of the network, the information diffusion matrix at this time becomes $I_{6\times6} \bullet R_{6\times5}$. After the current side node receives the information from each node, then the information on average each node after returning to the side connected to it, then get the information diffusion matrix $I_{6\times6} \bullet R_{6\times5} \bullet T_{5\times6}$ expression of which is $R_{6\times5}, T_{5\times6}$:

$$R_{6\times5} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} \tag{10}$$

$$T_{5\times6} = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix} \tag{11}$$

After the information diffusion six times to get a matrix, the matrix is a matrix of information diffusion between nodes on the network side, the resulting matrix diffusion six:

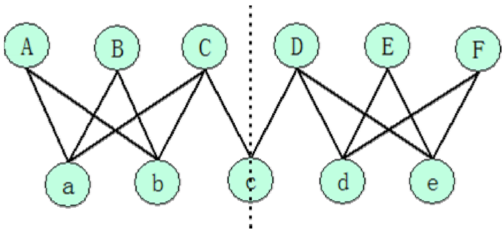

Fig. 1. bipartite network

Table 1. six iterative diffusion matrix

	A	B	C	D	E	F
A	0.2470	0.2470	0.3182	0.1101	0.0389	0.0389
B	0.2470	0.2470	0.3182	0.1101	0.0389	0.0389
C	0.2121	0.2121	0.2838	0.1451	0.0734	0.0734
D	0.0734	0.0734	0.1451	0.2838	0.2121	0.2121
E	0.0389	0.0389	0.1101	0.3182	0.2470	0.2470
F	0.0389	0.0389	0.1101	0.3182	0.2470	0.2470

Table is obtained after six iterations upper node information spread to other nodes on the side of the case. Column in the table indicates the information of other nodes similar to the volume diffusion node corresponding to the column, in the row corresponding to the line information indicating which node of the node spread to his information, the value of the diagonal line indicates after after six times the number of information diffusion itself still remaining information. For example, the amount of information of the node A diffusion node C to a value of 0.3182 on the amount of information diffusion node C to node A is 0.2121, the node A itself is 0.2470 in the remaining amount of information has experienced six times diffused.

A community of C community support is greater than the degree of support to other nodes, so the community A and C are combined into a single AC community, then updates the A and C corresponding line in the column, after updating the new community support for the selection of external a or C external support node greatest values, mutual support matrix updated as shown in Table 2, when the update is mutual support matrix, AC community's support for community-B maximum, perform a merge operation and updates matrix, and so on, until finally the entire unilateral nodes in the network are combined into a community until, as shown in Fig 2.

Table 2. AC merge support updated

	AC	B	D	E	F
AC	0.3182	0.2470	0.1451	0.0734	0.0734
B	0.3182	0.2470	0.1101	0.0389	0.0389
D	0.1451	0.0734	0.2838	0.2121	0.2121
E	0.1101	0.0389	0.3182	0.2470	0.2470
F	0.1101	0.0389	0.3182	0.2470	0.2470

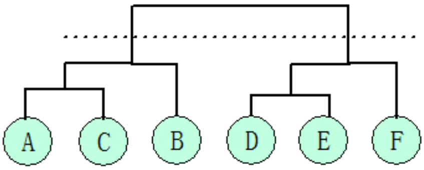

Fig. 2. unilateral node merging process

Figure 4 is a change in the value of Q in Fig. 3 node merging process, the Q value in the fourth step after completion of the merger maximum, where C and D after merging stops merge. To make the community is divided on the lower side in Figure 1 node, the same steps above. The results can be seen through, there is a loss of information, although after the mapping, but here as a community is divided on the basis of judgment is an effective method.

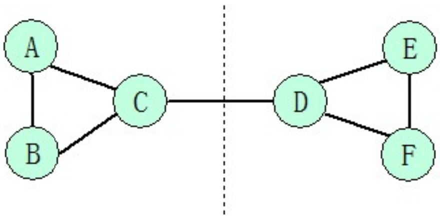

Fig. 3. unilateral node mapping

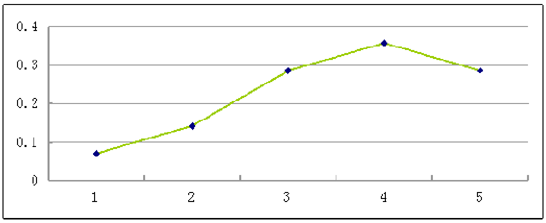

Fig. 4. Changes in the merger of Q value

4.4 Algorithm Complexity Analysis

The algorithm consists of: energy diffusion calculation, unilateral node mapping and Q value is calculated in three parts. Performed in the matrix multiplication energy dispersal process, this step algorithm complexity is O (n2 + m2), where m is the number of nodes in the bipartite graph node in the set U, n represents the number of nodes in the set V. After making six matrix multiplication algorithm complexity $O(6(m^2+n^2))$, to traverse the projection matrix unilateral node, this step of the algorithm complexity $O(mn)$, computational complexity Q value is approximately O (m2) so the final complexity of the algorithm is $O(k_1n^2+k_2m^2)$ where k1 and k2 are constants.

5 Experimental Results and Analysis

5.1 Tested on Southern Women Datasets

South African Women's Network (southern women) dataset [18] is the southern United States women's participation in the activities of the line into a real network, proposed by Stephen, the network consists of 18 women and 14 activities, numbered 1-18 points for the Women's Day, 19 -32 for the active node. As shown below:

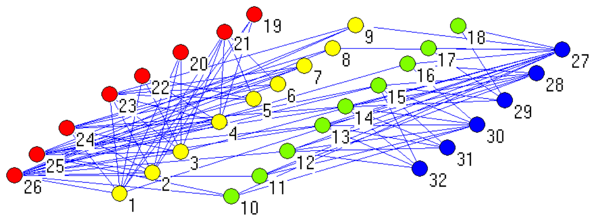

Fig. 5. South African women - Event Network

Experimental results show that the network is divided into four communities, including women 1-9 and 10-18 were each divided into a community, 19-26 and 27-32 are classified as another community, which is consistent with the original observations Davis . Women's network and the network event networks are projected, and dividing the result with several other algorithms divide the result of this algorithm compared with the module as a measure of the standard, to obtain results as shown in Table 1:

Table 3. Women - Event Network division results were compared with the Q value

Algorithm	Women's Network	Q Value	Event Network	Q Value
IPS	{1-9}{10-18}	0.3705	{19-26}{27-32}	0.3864
BRIM	{1-6}{7,9,10} {11-15}{8,16-18}	0.3455	{19-26}{27-32}	0.3864
LPA	{1-7,9}{8、 10-18}	0.3692	{19-24}{25-26} {27,29}{28,30-32}	0.3572
MP	{1-6}{7-10}{11-18}	0.3364	{19-24}{25-28} {29-32}	0.3487
ACODC	{1-9}{10-18}	0.3705	{19-26}{27-32}	0.3864

The figure shows the results of comparison of various algorithms division, the division of the results of the IPS module unilateral network algorithm to get the maximum value. Division results based on ant colony optimization algorithm for bipartite network of community division (ACODC algorithm) identical; the same division results Davis1 Davis did in the original literature. After mapping division of the women's network, obtain the following diagram shown in Figure 6, and FIG two are completely connected within the community, between the community is not completely connected. If there is no connection between whole numbers 2,4,5,6,7 and left women, although women and another one on the 1st of all women in the community who are now out of 21 in the same community, but if you put a woman on the 1st in another community, the module of the Q value of the entire network will be reduced, therefore the 1st women on the right side of the community can be better divided quality.

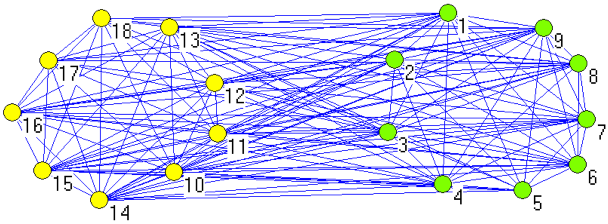

Fig. 6. Result of women network division

6 Conclusion

In order to divide the community bipartite network, we propose a bipartite network of community division algorithm based on information diffusion. The algorithm does not require additional parameters in the case, one side of the bipartite network node precise community division respectively.

The algorithm also exists insufficient; vertex mapping during unilateral calculate the Q value, the mapping information is missing presence. Ministry of network complexity and single community partitioning algorithm is also compared to the algorithm needs to be further improved. Therefore, the research division of bipartite network community needs to be further explored in solving practical problems.

References

1. Wang, X., Li, X., Chen, G.: Complex network theory and its applications. Tsinghua University Press, p. 9 (2006)
2. Newman, M.E.J.: Scientific collaboration networks. I. network construction and fundamental results. Physical Review E **64**, 016131 (2001)
3. Newman, M.E.J.: Scientific collaboration networks. II. network construction and fundamental results. Physical Review E **64**, 016132 (2001)
4. Liu, A., Chunhua, F., Zhang, Z.: Empirical statistical study Chinese mainland movie network. Complex Systems and Complexity Science **4**(3), 10–16 (2007)
5. Chen, W., Junan, L., Liang, J.: Disease gene network analysis bipartite graph projection. Complex Systems and Complexity Science **6**(1), 13–19 (2009)
6. Lambiotte, R., Ausloos, M.: Uncovering collective listening habits and music genres in bipartite network. Physical Review E **72**, 066107 (2005)
7. Le Blond, S., Guillaume, J.-L., Latapy, M.: Clustering in P2P exchanges and consequences on performances. In: van Renesse, R. (ed.) IPTPS 2005. LNCS, vol. 3640, pp. 193–204. Springer, Heidelberg (2005)

8. Xu, Y., Chen, J.: ACO based bipartite network community mining. Computer Science and Exploration **8**(3), 286–304 (2014)
9. Chen, B., Chen, J., Zou, S.: Mining algorithm based on matrix decomposition bipartite network community. Computer Science **41**(2), 55–101 (2014)
10. Zou, Y., Di, Z., Pan, Y.: Based on the distribution of resources bipartite network Fabric Matrix class methods. Beijing Normal University **46**(5), 643–646 (2010)
11. Gao, M.: Bipartite network community discovery algorithm. Hefei University of Technology, p. 04 (2012)
12. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**, 036106 (2007)
13. Liu, X., Murata, T.: Community detection in large-scale bipartite newworks. In: Proceedings of the 2009IEEE/WIC/ACM International Joint conference on Web Intelligence and Intelligent Agent Technology(WI-IAT 2009), pp. 5–8. IEEE Computer Society, Washington, DC (2009)
14. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system:optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics: Part B **26**(1), 29–41 (1996)
15. Barber, M.J.: Modularity and community detection in bipartite networks. Physical Review E **76**, 066102 (2007)
16. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National academy of sciences of the United states of America **103**(23), 8577–8582 (2006)
17. Milgram, S.: The small world problem. Psychology Today, 60–67, May 1967
18. Davis, A., Gardner, B.B., Gardner, M.R.: Deep south. University of Chicago Press, Chicago (1941)
19. Scott, J., Hughes, M.: The anatomy of Scottish capital: Scottish companies and Scottish capital. Croom Helm, London (1980). 1990–1979
20. Fan, C.L., Yu, L.: 2013 2nd International Conference on Mechatronic Sciences, Electric Engineering and Computer, (MEC2013) ShenYang, liaoning, China (2013)