The Application Research of Social Networks Community Detection with Correlation Coefficients

Fuqiang Zhao[1, a], Guijun Yang[1, b], Enjun Xing[1, c], Li He[1, d]

[1]Department of Information Science and Technology, Tianjin University of Finance & Economics, Tianjin, 300222,China

[a]email: fqzhao@126.com, [b]email: yangguijun@tjufe.edu.cn, [c]email: xing_enjun@126.com, [d]email: renkeheli@163.com

Keywords: algebraic connectivity; matrix reordering; laplascian matrix; correlation coefficients; social networks

Abstract. In complex network graph, cut model based on edge centrality doesn't apply to overlapping community detection by minimizing the algebraic connectivity of complex networks. The problem can be resolved by calculating node Correlation coefficients. It cuts one edge at one time, which is the fastest decline in the algebraic connectivity, until to divide into two communities. When components of fielder vector is less than threshold level 'a' and the difference of node adjacent edges which belong to two groups is less than 2, correlation coefficients is calculated. This node's community can be detected by the correlation coefficients. We concluded that elapsed time by ordering matrix is less than before. The advanced cut model can be used in overlapping community detection with higher efficiency and accuracy.

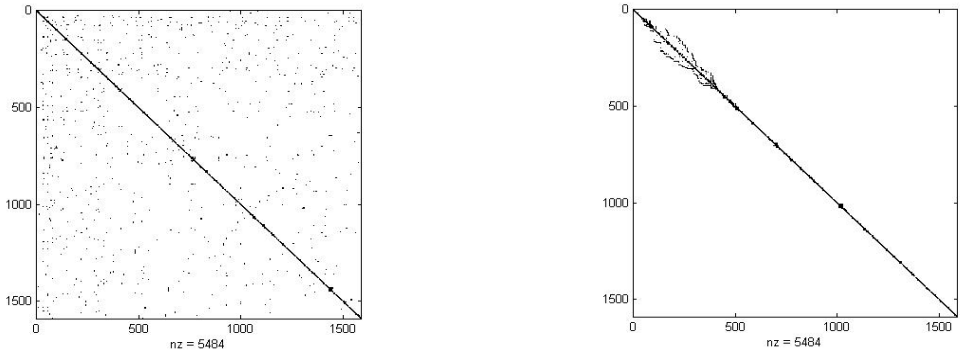
Introduction

With the rapid development of many social networks in the last decade, the social network has a huge amount of data. The number of network nodes can reach millions or billions[1].The processing of large data also promotes the development of complex network models and methods. Although some community detection algorithms can effectively identify communities in complex networks, the algorithm complexity is still high, and the efficiency is not high. For example, GN (Girvan-Newman, 2002) [2] is the classical community detection algorithm. Its main drawback is that each iteration can only delete one edge the time complexity of GN algorithm is $O(m^2n)$. For sparse networks, its time complexity is $O(n^3)$. Where n is the number of nodes and m is the number of edges. Fortunato et al. proposed information centrality method to judge the edge between the communities. The nodes with high information centrality are link edge in different community. This algorithm has good classification accuracy in the case of network community structure, but the time complexity is $O(m^3n)$. It is $O(n^4)$ for sparse network. Some scholars put forward the community cut models, but the time complexity is relatively high. The [3] (Clauset et al., 2004)time complexity of is $O\left(nlog^2n\right)$.The [4] (Duch & Arenas,2005) time complexity is $O(n^2logn)$. The [5] (Eckmann & Moses, 2002) time complexity is $O(m<k^2>)$. The [6] (Capocci et al.,2005) time complexity is $O(n^2)$. The [7] (Zhou & Lipowsky, 2004) time complexity is $O(n^3)$. A cutting edge model based on the edge centrality measure is proposed in the [18](Edge Centrality Cut Model, ECCM, Zhang, 2012). Although the model in a certain extent reduces the time complexity, but it did not consider the overlapping community nodes. The cut model uses the Lanczos algorithm [8] for calculating the second smallest eigenvalue (graphs by algebraic connectivity) of Laplacian matrix. The Related software packages are ARPACK[9] and Irbleigs[10]. The methods based on trace minimization include TraceMin-Fiedler[11], MC73_FIEDLER[12] and NewTraceMIN[13]. Different solutions lead to differences in computational speed. The speed of matrix reordering before and after is worth further study. In the reference [14, 15, 16], community is detected by the laplascian matrix and the module matrix. Their eigenvalues are used to reflect the significance of each eigenvector, and the

gap between the two eigenvalues can be used as an important basis for the multi-scale community structure detection.

Related Work

Given a network or graph $G = (V, E)$, V is a set of n nodes and E is a set of m edges. G can be represented by adjacency matrix or incidence matrix. If graph G is sparse networks, non-zero elements in a Matrix is called nz for short. l_{ij} is an edge connecting nodes i and j, $l \sim (i, j)$. The sparse matrix reordering [17] is shown in Fig. 1 (a). Where the black point is non-zero elements. The elements in the matrix are scattered and the bandwidth is very large. So the convergence speed can be reduced by using the iterative method. The elements of the matrix are reordered to be shown in Figure 1 (b), the node distribution is more compact, the bandwidth is significantly reduced.



(a) no matrix reordering (b) matrix reordering

Fig. 1. Scientific collaboration network sparse matrix node distribution

The incidence matrix $A \in R^{n \times m}$ of the graph G is the matrix with lth column a_l .For an edge l connecting nodes i and j, we define the edge vector $a_l \in R^n$ as $a_{li} = 1$, $a_{lj} = -1$ and all other entries 0.

The Laplacian L of G is the n×n matrix:

$$L = AA^T = \sum_{l=1}^{m} a_l a_l^T \tag{1}$$

Let $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of L. The second smallest eigenvalue $\lambda_2(L)$ is called the algebraic connectivity of the graph G, and the corresponding normalized eigenvector is called the Fiedler vector. The algebraic connectivity is considered to be a measure of how well-connected a graph is.

That is, the more connected graph has the greater algebraic connectivity on the same vertex set. The algebraic connectivity reflects the degree of connectivity between the nodes in the connected graph. $\lambda_2(L) > 0$ if and only if G is connected. $\lambda_2(L)$ is monotone increasing in the edge set. The algebraic connectivity function of complex networks is a monotone convex function, which is defined as:

$$L(x) = L - \sum_{l=1}^{m} x_l a_l a_l^T \tag{2}$$

Where $x_l = 1$ if edge l belongs to the edge subset, and $x_l = 0$ otherwise. We study the following problem. Choose k edges from candidate edges that lead to the greatest decrease in algebraic connectivity when cut from G. That is, we want to solve the problem that the following convex function is minimal:

$$\begin{aligned} &\text{minimize} \quad \lambda_2(L - \sum_{l=1}^{m} x_l a_l a_l^T) \\ &\text{subject to} \quad 1^T x = k, \\ &\qquad\quad x \in [0,1]^m, \end{aligned} \tag{3}$$

The Application of Correlation Coefficients in Community Detection Algorithm

The reference [18] Edge Centrality Cut Model (ECCM), which is different from the traditional social network Community Detection algorithm, is proposed to deal with the large-scale community structure based on the edge centrality measure. The model is taken by the method of spectral analysis, and the edge centrality of each edge is calculated. The edge which is the greatest decrease in algebraic connectivity is cut. And then the spectral center of the updated graph is calculated iteratively. Once cut an edge, until to $\lambda_2(L) = 0$. The model has a great speed improvement in the community division of the social network. Its results are satisfactory, but the model did not consider the overlapping community nodes.

In contrast to the traditional algorithms used to detect communities, the ECCM defines edge centrality by spectral analysis. This cut algorithm is better for medium-size networks with higher efficiency and accuracy. It deletes the node with the highest spectral centrality, and then recalculates the spectral centrality of the remaining nodes.

The algebraic connectivity function of the cut edge model based on the edge centrality measure is

$$\lambda_2(L - x_l \frac{1}{e^{\max(|v_i|,|v_j|)}} a_l a_l^T) \tag{4}$$

the gradient is

$$\frac{1}{e^{\max(|v_i|,|v_j|)}} (v_i - v_j)^2 \tag{5}$$

However the values of the components of the fiedler eigenvector are not well distinct overlapping Community. Thus one has to look at correlations between corresponding components of different eigenvectors, rather than the value of the components itself.

$$r_{ij} = \frac{\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle}{\left[\left(\langle x_i^2 \rangle - \langle x_i \rangle^2 \right) \left(\langle x_j^2 \rangle - \langle x_j \rangle^2 \right) \right]^{\frac{1}{2}}} \tag{6}$$

Where the average $\langle \cdot \rangle$ is over the first few nontrivial eigenvectors. The quantity r_{ij} measures the community closeness between node i and j.

ECCM with correlation coefficient is as follows:

(1) Calculate the spectral centrality for each edge of graph G based on the edge centrality function and sort them.

(2) Find k edges with the highest spectral centrality ($k \geq 1$), and then, delete them; renew the complex network to G^{new}. The choice of k is based on the edge sparse degree of complex networks. When G is a sparse graph, k is equal to one. Otherwise, k is more than one.

(3) Calculate algebraic connectivity of a graph. If $\lambda_2 = 0$, the algorithm goes to step (4), otherwise step (1).

(4) When the corresponding component value of the Fiedler vector of a node is less than the thresholdα, we should analyze the necessity of computing the correlation coefficient. Moreover, we should calculate the correlation coefficient of the node if the difference in the number of edges that it connects to the two communities is less than 2.

(5) Use formula (3) to calculate the correlation. By comparing the two correlation coefficient of the nodes in the two communities, we can determine to which community the node belongs to.

(6) Update the complex network new graph G_1^{new} and G_2^{new}. The graph G is divided into two

communities.

Test results

Matrix reordering experiment

Matrix reordering experiments use 4 matrices: Dolphin, Football, Netscience and rail_1357. The second smallest eigenvalue and consume time are calculated respectively for matrix no reordering and matrix reordering. As shown in Table 1, four methods are used for the sparse matrix decrease to a certain degree.

Table 1 Comparison of the consume time of matrix no reordering and matrix reordering for the sparse matrix

Dataset		Fiedler	lanczos	Eigifp	JDCG
Dolphin (nz=318)	no reordering	0.09806	1.6357	0.3953	0.33909
	reordering	0.00345	0.53	0.05741	0.0711
Football (nz=1226)	no reordering	0.03113	1.47629	0.04936	0.07079
	reordering	0.00843	1.34305	0.02599	0.05329
Netscience (nz=5484)	no reordering	0.02059	1.88928	1.03403	0.43511
	reordering	0.01925	1.68309	1.00863	0.36573
rail_1357 (nz=8985)	no reordering	0.10761	6.67079	3.10416	0.5837
	reordering	0.04432	5.83562	1.81269	0.29575

Test cut model on real network

The real network is the American college football network, a well-known graph regularly used as a benchmark to test community detection algorithms. There are 115 vertices and 613 edges, representing the teams, and two vertices are connected if their teams play against each other. Four cut edge methods are used in the experiments: Random, GN, GACEM(Greedy Spectral Optimal Cut edge Model)[18] and ECCM. GN method is used to delete the edge with largest betweenness. Fig.2(a) shows the test results. The spent time of ECCM is shorter than GN and GACEM, as shown in Table 2.

Table 2 Comparison of the consume time of cutting models

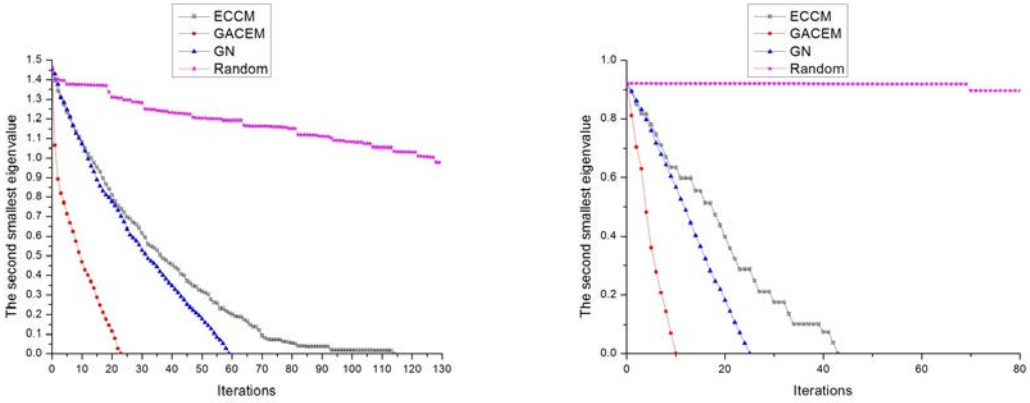
Time(s) Cutting sequence / Cutting models	1	2	3	4	5	6	7	8	9
ECCM	0.178579	0.281273	0.40904	0.533574	0.651089	0.783801	0.915594	1.028059	1.171516
GACEM	0.43202	0.69765	0.88325	1.094732	1.390213	1.62319	2.018373	2.31063	2.518721
GN	13.15193	26.14901	39.39728	52.525	65.18353	78.0974	91.15705	104.2828	117.2086

Test cut model on artificial network

Artificial network consists of 1000 vertices and 7572 edges. Its pout is 0.1. Fig.2(b) shows the test results. The initial value of λ_2 of artificial network is 0.919444 and bigger than 0.8 after deleting 80 edges randomly. λ_2 equals to zero after deleting 25 edges with the highest edge betweenness by GN algorithm, and the cutting time is too long. λ_2 equals to zero after deleting 23 edges by GACEM method, and the cutting time is shorter than GN algorithm. ECCM has the same result as GN and GACEM with the shortest time.

The real network is Zachary's network of karate club members[19], a well-known graph regularly used as a benchmark to test community detection algorithms. It consists of 34 vertices and 78 edges, the members of a karate club in the United States, who were observed during a period of three years. The result of the ECCM is same as for the Laplacian matrix by calculating karate network. Compared to the results using the actual network, the result is wrong. The modularity method is the same as for the actual network. As shown in Figs.3, the eigenvalues of nodes 3, 15 and 21 are close to zero. It is difficult to distinguish which community the nodes 3, 15 and 21

belong to. When absolute values of the difference between the eigenvalue and 0 are less than the threshold value a. But the difference in the number of edges 15 and 21 that it connects to the two communities is not less than 2. Therefore, correlation coefficient between the node 3 and other is calculated. Node 3 should have same community with node 1 and 2. The result of community detection by the improved ECCM is shown in Fig.4.



(a)Football network (b) Artificial network

Fig. 2. λ_2 calculated by different cut edge model and iterations

Conclusion

The Edge Centrality Cut Model with correlation coefficient can apply to overlapping community detection by minimizing the algebraic connectivity of complex networks. When components of fielder vector is less than threshold level a and the difference of Node adjacent edges Belong to two groups is less than 2, correlation coefficient is calculated. Although the ECCM time complexity is relatively low which solve the second smallest eigenvalue of Laplacian matrix by Lanczos algorithm, calculating correlation coefficient also need spend extra time. At the same time, how to detect the large-scale complex network by this model and calculate the second smallest eigenvalue by parallel processing method, these require further study in the future.

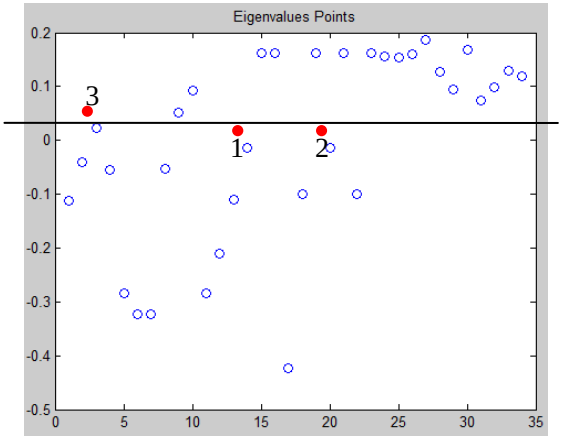


Fig. 3. Karate node eigenvalue scatter graph

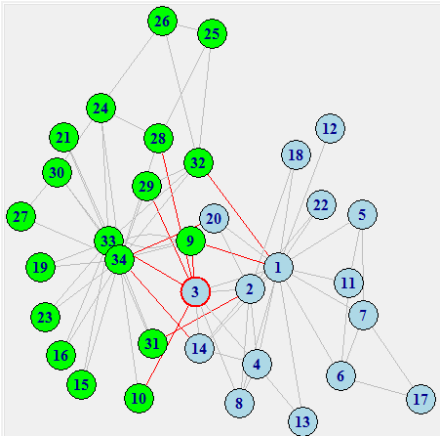


Fig. 4. Karate correct bisect result diagram

Acknowledgement

In this paper, the research was sponsored by the Natural Science Foundation of China (Project No. 11471239), Tianjin Natural Science Foundation of China (Project No. 15JCYBJC16000) and Tianjin Philosophy and Social Science Research Program Foundation Project China (Project No. TJTJ15-002).

References

[1] Charu, C., Aggarwal. Social Network Data Analytics[M]. Springer Science + Business Media, LLC (2011).

[2] Girvan M, Newman M E J. Community structure in social and biological networks[J].Proceedings of the National Academy of Sciences of the United States of America, 2002,99(12):7821-7826.

[3] Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (6), 2004, 066111.

[4] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2), 2005, 027104.

[5] J.-P. Eckmann, E. Moses, Curvature of co-links uncovers hidden thematic layers in the World Wide Web, Proc. Natl. Acad. Sci. USA 99 (2002) 5825-5829.

[6] Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Detecting communities in large networks[J], Physica A 352, 2005, 669-676.

[7] H. Zhou, R. Lipowsky, Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities[J], Lect. Notes Comput. Sci. 3038, 2004, 1062-1069.

[8] C. Lanczos. An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integtal Operators[J]. Journal of research of the National Bureau of Standards. 1950,45(4):255-282.

[9] R.B. Lehoucq, D.C. Sorensen, C. Yang, ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, SIAM, Philadelphia, PA, USA, 1998.

[10] Baglama, J., Calvetti, D., and Reichel, L. 2003. Algorithm 827: irbleigs: A MATLAB program for computing a few eigenpairs of a large sparse Hermitian matrix. ACM Transactions on Mathematical Software 29, 3 (Sept.), 337–348.

[11] A. Sameh, Z. Tong, The trace minimization method for the symmetric generalized eigenvalue problem, J. Comput. Appl. Math. 123 (1–2) (2000)155–175.

[12] Y.F. Hu and J.A. Scott. HSL MC73: a fast multilevel Fiedler and profile reduction code. TechnicalReport RAL-TR-2003-036, 2003.

[13] A. Klinvex, F. Saied, A. Sameh.Parallel implementations of the trace minimization scheme TraceMIN for the sparse symmetric eigenvalue problem[J].Computers and Mathematics with Applications.2013(65):460-468.

[14] H. W. Shen, X. Q. Cheng and B. X. Fang. Covariance, correlation matrix and the multiscale community structure of networks[J], Phys. Rev. E, 2010, 82:016114.

[15] H. W. Shen and X. Q. Cheng, Uncovering the community structure associated with the diffusion dynamics on networks[J]. Stat. Mech. (2010) P04024.

[16] H. W. Shen, X. Q. Cheng and Y. Z. Wang. A Dimensionality Reduction Framework for Detection of Multiscale Structure in Heterogeneous Networks[J]. Journal of Computer Science & Technology,2012,(02):341-357.

[17] Kyle V. Camarda , Mark A. Stadtherr. Matrix ordering strategies for process engineering: graph partitioning algorithms for parallel computation[J]. Computers and Chemical Engineering 23 (1999):1063–1073.

[18] Shuo Zhang. The Research of the community detection model in complex networks based on algebraic connectivity [D]. Tianjin University (2012).

[19] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75-174.