Survey of Community Structure Segmentation in Complex Networks

Tingrui Pei
College of Information Engineering, Xiangtan University, Xiangtan, China
Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education,Xiangtan University, Xiangtan, China
Email: peitr@163.com

Hongzhi Zhang
College of Information Engineering, Xiangtan University, Xiangtan, China
Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education,Xiangtan University, Xiangtan, China
Email: zhz317387928@163.com

Zhetao Li*
College of Information Engineering, Xiangtan University, Xiangtan, China
Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education,Xiangtan University, Xiangtan, China
School of Computer, National University of Defense Technology, Changsha, China
Email: chu5044130@sohu.com

Youngjune Choi
Department of Information and Computer Engineering, Ajou University, Suwon, Korea
Email: choiyj@ajou.ac.kr

***Abstract*—Community structure analysis is a hot research spot in social networks and complex networks. In order to summarize recent research progress, this paper reviews the background, the motivation, the advantages and disadvantages of existing works related to community structure discovering. A comprehensive outline was obtained by analysis of different clustering algorithms**.

***Index Terms*—Social networks, Community structure, Segmentation analysis**

I. INTRODUCTION

Social network is to reflect the relationship between the nodes which are in the intra-group or in the inter-group. It uses nodes to represent individuals in the network, and attachment between nodes to represent the relationship among individuals, the community to represent groups which share the same characteristic. Thus, what we define for social network is that to project the complex relationship in society to nodes and attachment in network.

In the past decade, there has been a surge of interest in both empirical studies of networks [1], and development of mathematical and computational tools for extracting insight from network data [2-5]. The best-studied form of large-scale structure in networks is modular or community structure [6, 7]. The main reason is that community structure may correspond to functional units within a networked system. An example of this kind of link between structure and function drives much of the present excitement about networks: In a metabolic network, a community might correspond to a circuit. In a social network, a community might correspond to a group of people brought together by a common interest. Discovery of communities also has great significance in recognition terrorist organization or prevention infectious diseases and so on.

Although, scholars have made outstanding achievements in the past years, the existing research results are still not enough to discover the relationship between the function and structure in complex network. Besides, there are no uniform optimal community concept and measurement criteria on dividing the community, so it is difficult to have a recognized measure to identify the quality of the discovered communities.

II. CLASSIFICATION ALGORITHMS

At present, the complex network clustering algorithms can be divided into the following 3 types:

The first type: Based on optimization methods, it transforms the complex network clustering problem into a

*Corresponding author: Zhetao Li , Email: chu5044130@sohu.com

doi:10.4304/jsw.9.1.89-93

quadratic optimization problem, through the calculation of the matrix characteristic vector to optimize the predefined "cut" function, such as: spectrum divide method, etc.

The second type: Based on heuristic methods, it transforms the complex network clustering problem into predefined heuristic planning design problems, such as: GN algorithm, CPM algorithm, and so on.

The third type: other clustering methods, such as: the method based on similarity, etc.

III. Based on the Optimization Clustering Method

The two principal methods based on optimization clustering method are Spectrum method and local optimal method. Spectrum method is the application of "cut" function which is minimizing predefined by quadratic optimization technology. The lowest "cut" division is considered to be the best network division. Local optimal algorithm includes three basic parts: the objective function, the optimal solution and the candidate search strategy. The differences existing in the different algorithm are the objective function and the search strategy of the optimum solution.

(1) Traditional spectrum divide method [8] based on the Laplace matrix. This algorithm based on the node that is in the same club is approximately equal to division the structure of community in each element that is in the eigenvectors of non-zero eigenvalues. In most cases, the actual Laplace matrix is a sparse matrix, so it can be fast calculated the characteristic vector though the Lanczos method, which has advantage of high speed and accuracy.

The biggest disadvantage of traditional spectrum method is that it only divides the community into two communities once. To have a network divided into more communities, one has to repeat the algorithm many times. As to this problem, Wu and Huberman put forward the rapid resistance spectrum segmentation method [9] based on the network voltage spectrum. The basic idea is: if one treats two nodes which are not in the same club considered as source nodes (voltage is 1) and the end node (whose voltage is 0), treats each side as a resistance whose value is 1,then, they will have a similar voltage value in the same community. This algorithm has low complexity ($O(m+n)$), but it needs to know the number of the communities in advance.

In order to overcome the defects, Capocci proposed another spectrum divide algorithm [10] on the foundation of traditional spectrum method and on standard matrix $N = K^{-1}A$. Using standardized transformation conversion, the maximum eigenvalue of the matrix N is always more than 1, and the corresponding feature vector is called the first the ordinary characteristic vector. In a community structure with obvious network, if the number of community is K, then the first ordinary eigenvalue of the matrix N will be K-1 which is very close to 1, while the other values are obviously different from 1. Besides, in this K-1 eigenvalue vector, the nodes in the same club will be much closed. Therefore, in the network with obvious structure, the elements distribution in the vector is the obvious form of steps, and the level of the ladder is equal to the number of community K, so the spectrum divide algorithm can get a good effect.

(2) K-L algorithm [11], fast Newman algorithm [12] and GA algorithm [13] in chapter 3 is typical clustering algorithm methods which based on local search optimization technology. This kind of algorithm includes three basic parts: the objective function, the candidate of the search strategy and optimal solution search strategy. Almost all of them have the same candidate solution search strategy, but the target function and the optimal solution search strategy are different.

In the process of K-L algorithm, it only accepts better candidate solution, so the solution that finds can always be the local optimal but not the global. The biggest limitation of K-L algorithm is that it needs prior experience to produce good initial cluster structure, otherwise the bad initial solution may lead to slow convergence speed and the worse of the final solution.

In order to improve the algorithm rate, in 2004, Newman put forward the FN [12] algorithm, which is mainly realized by maximizations module degrees Q. The function Q is defined as follows:

$$Q = \sum_{s=1}^{K} [\frac{m_s}{m} - (\frac{d_s}{2m})^2] \quad (1)$$

The complexity of the algorithm is $O(mn)$, m, n is representing the all number edges and nodes in network respectively, but FN algorithm compared with GN algorithm has less accuracy.

Through the same goal with FN, Guimera and Amaral put forward clustering algorithm GA [13] based on the simulated annealing algorithm. This algorithm has jumped out local optimal solution, and has the ability of finding optimal solutions, so its cluster precision is high. However, because of its sensitive to input parameter, different parameter setting often leads to bigger difference clustering results and running time.

Take optimization method to identify the network structure totally depends on optimizing targets, so "biased" goal function will cause "biased" solution. But, the vast majority of optimization algorithm is based on the maximized Q value to do cluster analysis. However, the research found that Q function is biased itself. It can't characterize the optimal cluster structure accurately, which means this algorithm may not find all network cluster structures which are real properly.

IV. Heuristic Clustering methods

GN algorithm [6], improved GN algorithm [4, 14], CPM algorithm [15] and PAK algorithm [16] are classic heuristic complex network clustering algorithm. The common characteristics of this kind is: for most network, to design heuristic algorithm on certain intuitive hypothesis, they can find the optimal solution or time optimal solution quickly, but they can't ensure that it can get satisfactory solution to all input network from theory.

In 2002, Girvan and Newman put forward GN algorithm, using repeatedly recognition and cluster strategy clustering complex network which delete the

connections in between [6]. It mainly base on the theory that the intra-group's betweenness should be bigger than inter-group eventually once can build a hierarchical clustering tree to realize community division. The biggest drawback of GN algorithm is the low calculation speed, which is suitable for small and medium-sized network. But, it doesn't need prior conditions, and possess high accuracy.

In 2003, Tyler introduced the basic statistical methods into GN algorithm, and put forward approximate algorithm GN [4]. The monte carlo algorithm estimates part of the approximate number of node betweenness, rather than calculates all the exact number of node betweenness. Obviously, this algorithm improves the speed, but reduces the cluster precision.

Considering that the GN algorithm's low efficiency is caused by excessive spending count edge betweenness, Radicchi propose connection clustering coefficient to replace the edge betweenness [14] in 2004. Based on the theory that cluster of connections between connected clustering coefficients should be less than that in connected clustering coefficient to divide community. The complexity of the algorithm is $O(m^3 / n^2)$, in which m is an iteration time, and n is total node. The biggest limitation of this algorithm is that it is not suitable for network with little processing circuit or no circuit.

At present, overlap community structure is not considered in most algorithms, however, has more practical significance in most applications. For example, in the semantic web, polysemous are allow to be appeared in network cluster with different meaning. In 2005, Palla put forward CPM algorithm [15] which can identify overlap network structure. The fundamental assumption of the algorithm is: network constituted by more neighbor cluster of k-group, and the next-to two k groups share k-1 nodes at least, with each k group only belonging to a cluster, but the k-groups belonging to different network may share some nodes. This algorithm is the first one which can calculate the overlap community structure. In actual application, parameter k is difficult to be determined, the selection of different k value often gets greater difference between network cluster structure, so it difficult to judge quality.

In order to solve the problem of excavation overlap community structure, Gregory, put forward PAK algorithm [16] in 2010 to find overlapping community. This method is based on the advanced modular design and use the professional label algorithm as its basic theory. It is the best overlap community found algorithm at present, as it can recover the overlap community effectively and is suit for large-scale and intensive network, but it also has higher complexity.

V . Other Clustering Method

Besides the two principal methods mentioned above, there are other complex network clustering methods, for example: hierarchical clustering method based on similarity, in which the similarity of nodes is defined based on the structure of network topology, similar coefficient based on structure congruent [17], similarity based on the random walk [18], joint center [19] and node number based on the sharing neighbor [20], etc.

VI. The Comparison between the Clustering Algorithms

TABLE I.
HEURISTIC CLUSTERING ALGORITHM

Algorithm	Complexity	Accuracy	Applicable condition
GN algorithm	high	high	Medium and small network
Akin-GN algorithm	low	low	Complex network
Improve GN algorithm	low	high	More contour network
CPM algorithm	low	low	Network have obvious hierarchical structure
PAK algorithm	low	high	Mass and intensive network

The heuristic algorithm it is generally to designing algorithm rules based on the intuitive assumption, so from the table 2, it can generally draw an ideal accuracy, and its algorithm complexity is also improved compared with optimization algorithm. But, it is because the intuitive assumption that its usable range is reduced and optimization algorithm present the conditions more obviously. Hence, seeking for universal characteristics of the complexity of network or the nature of community is also a challenge to heuristic algorithm.

TABLE II.
BASED ON THE OPTIMIZATION CLUSTERING ALGORITHM

Algorithm	Complexity	Accuracy	Applicable condition
Laplace Matrix algorithm	low	high	Two community network
Resistance - voltage spectrum segmentation method	low	high	known the number of community
The standard of the spectrum matrix method	low	high	Network have obvious hierarchical structure
K-L algorithm	high	high	Good initial cluster structure
FN algorithm	low	low	Complex network
GA algorithm	high	high	Network have obvious hierarchical structure

Table 1 show when the complexity and the accuracy of the algorithm are assured, the usable range will greatly cut or need more prior experience. But for a wide useable algorithm, its complexity and accuracy will become a pair of contradictory. The reason may be that optimization algorithms are not able to fully understand the relationship between community structure and the complex network itself, thus leading to the "deviation"

for the objective function, which eventually leads to the restrictions of the accuracy and applicability. So, how to find the reasonable target function is a challenge to optimize clustering algorithms.

VII . Comparison of Accuracy and Stability

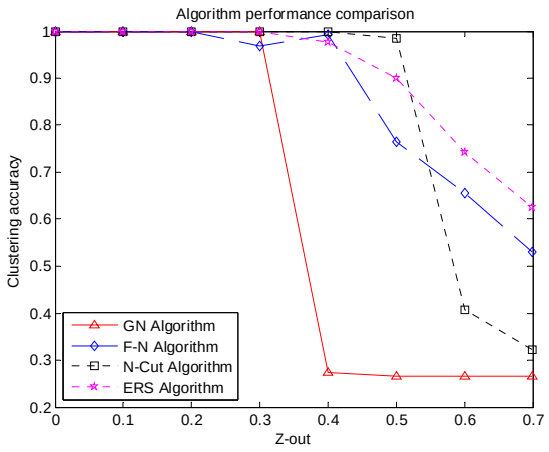

Figure 1. Algorithm performance comparisons..

Most of the existing algorithm of complex network clustering can be divided into two categories: Optimization based methods and heuristic methods. This paper shows their advantages and disadvantages respectively through the analysis of two kinds of social networks community partition algorithm. In this experiment, we use the network which is put forward by Newman and Girvan. This network is regarded as a standard network when testing the performance of the algorithm. Each graph was constructed with 128 vertices, which divided into four communities of 32 vertices each, and the node degree can be adjusted freely. Change of the node degrees would affect the artificial networks hierarchy structure, thus the networks can test the performance of community partition algorithm better. In the experiment, we compared the performance by testing GN, FN, N-Cut and ERS algorithm.

At first, it is a truth for a community that the nodes internal degree is larger than the nodes external degree necessarily in the networks, which means that out-degree can reflect the close degree of internal nodes in the networks. Thus it is concluded that the entire network was of the degree of obvious of hierarchy. The larger of the out-degree of the network, the less obvious structure of the network was. This paper uses Z-out as a variable to divide community. We can see it clearly in Fig.1 that the bigger the value of Z-out, the less accuracy it will be. Also, it shows all algorithms perform high accuracy when Z-out is less than 0.3, but the accuracy of the algorithms begins to fall as the Z-out grows. When the Z-out is less than 0.3, GN algorithm drops faster than any other algorithms, while ERS and N-Cut algorithm still perform high accuracy. As we can see that Z-out=0.5 is a turning point. All algorithms perform a lower accuracy when Z-out is bigger than 0.5, but the ERS algorithm still keeps stable relatively. So ERS has a high stability than other algorithms, and it is more suitable for complex network.

VIII . Conclusion

The complex network clustering is one of the most important complex network analysis methods, and has a broad prospect of application. This paper mainly focuses on the existing complex network clustering algorithms. According to the basic solution strategy, the existing complex network clustering algorithms are divided into two categories: optimization-based method and heuristic-based method; From analyzing the basic principles and relative merits of the existing methods, it shows that encouraging results have been achieved in decade years, but the complex network clustering problem is far from been well resolved. It is embodied in the following respects:

The first problem is that what is the inevitable relationship between the network cluster structure and the else complex phenomenon of the network? Namely, from the network "inner" attributes, can we promote an optimization objective function, an objective reflect and characterizations of the cluster structure.

Secondly, most of the existing discover algorithm for complex networks consider the complexity and precision, or due to the needs of transcendental experience, its scope of application is greatly limited. Therefore, how to design a clustering algorithm which is fast, high precision and no supervision is also a burning question.

Lastly, along with the development of social development and expansion in application field, network clustering problem are diversified. The existing algorithm has been difficult to satisfy present demand, so developing the new clustering algorithm for the special network is necessary. The typical problem during the process includes the division of the overlapped community, dynamic complex network of clustering, distributed network clustering and so on [21-25]. In addition, the unmetered theory and technology has also become one of the major problems in current complex network clustering.

It is believed that these three aspects will be the developing direction of the complex network research in future.

Acknowledgment

This research was supported by National Natural Science Foundation of China with Grant No. 61173036, Hunan Province College Key Laboratory Open Foundation Project with Grant No. 2009GK3016 and Science and Technology Planning Project of Hunan Provincial Science & Technology Department with Grant No.2011GK3200.

References

[1] Newman M E J. The structure and function of complex networks[J]. SIAM review, 2003, 45(2): 167-256.
[2] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics reports, 2006, 424(4): 175-308.
[3] Newman M. Networks: an introduction[M]. Oxford University Press, 2009.

[4] Tyler J R, Wilkinson D M, Huberman B A. E-mail as spectroscopy: Automated discovery of community structure within organizations[J]. The Information Society, 2005, 21(2): 143-153.
[5] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements[J]. Advances in Physics, 2007, 56(1): 167-242.
[6] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.
[7] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3): 75-174.
[8] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430-452.
[9] Wu F, Huberman B A. Finding communities in linear time: a physics approach[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 331-338.
[10] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting communities in large networks[J]. Physica A: Statistical Mechanics and its Applications, 2005, 352(2): 669-676.
[11] Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 321-330.
[12] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.
[13] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900.
[14] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663.
[15] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.
[16] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(7): 1493-1500.
[17] Wasserman, Stanley, and Joseph Galaskiewicz, eds. Advances in social network analysis: Research in the social and behavioral sciences. Sage, 1994.
[18] Pons P, Latapy M. Computing communities in large networks using random walks[M]//Computer and Information Sciences-ISCIS 2005. Springer Berlin Heidelberg, 2005: 284-293.
[19] Yang B, Liu J. Discovering global network communities based on local centralities[J]. ACM Transactions on the Web (TWEB), 2008, 2(1): 9.
[20] Sun P G, Gao L, Shan Han S. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks[J]. Information Sciences, 2011, 181(6): 1060-1071.
[21] Gan X, Wang J. The synchronization problem on a class of supply chain complex network[J]. Journal of Computers, 2013, 8(2): 267-271.
[22] Ma R, Deng G, Wang X. A cooperative and heuristic community detecting algorithm[J]. Journal of Computers, 2012, 7(1): 135-140.
[23] Wang Y, Gao L. Detecting protein complexes by an improved affinity propagation algorithm in protein-protein interaction networks[J]. Journal of Computers, 2012, 7(7): 1761-1768.
[24] De Lay N, Gottesman S. A complex network of small non-coding RNAs regulate motility in Escherichia coli[J]. Molecular microbiology, 2012, 86(3): 524-538.
[25] Barthwal R, Misra S, Obaidat M S. Finding overlapping communities in a complex network of social linkages and Internet of things[J]. The Journal of Supercomputing, 2013: 1-24.

Tingrui Pei, born in 1970. PhD, professor, Doctor Supervisor. He is graduated from Beijing University of Posts and Telecommunications, His main research interests include wireless sensor network (WSN) and Multimedia communication.

He is a professor of Dept. Information and Communication Engineering Xiangtan University.

Zhetao Li, born in 1980. PhD,Associate professor, Master Supervisor. His main research interests include internet of things (IOT), compressive sensing, social computing. He is Associate professor of Dept. Information and Communication Engineering Xiangtan University.

Hongzhi, Zhang born in 1987. Master graduate student. His main research interests include internet of social computing and complex large-scale data information processing.