A weight's agglomerative method for detecting communities in weighted networks based on weight's similarity*

Shen Yi(沈 毅)†

College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, *China*

(Received 28 August 2010; revised manuscript received 27 October 2010)

This paper proposes the new definition of the community structure of the weighted networks that groups of nodes in which the edge's weights distribute uniformly but at random between them. It can describe the steady connections between nodes or some similarity between nodes' functions effectively. In order to detect the community structure efficiently, a threshold coefficient κ to evaluate the equivalence of edges' weights and a new weighted modularity based on the weight's similarity are proposed. Then, constructing the weighted matrix and using the agglomerative mechanism, it presents a weight's agglomerative method based on optimizing the modularity to detect communities. For a network with n nodes, the algorithm can detect the community structure in time $O(n^2 \log_2^n)$. Simulations on networks show that the algorithm has higher accuracy and precision than the existing techniques. Furthermore, with the change of κ the algorithm discovers a special hierarchical organization which can describe the various steady connections between nodes in groups.

Keywords: complex networks, weight's similarity, community structure, weight's agglomerative method

PACS: 05.45.Xt, 89.75.Hc **DOI:** 10.1088/1674-1056/20/4/040511

1. Introduction

Complex network is an effective tool to study the complex large systems in the real world (the Web,[1,2] social networks,[3] protein interactions,[4] citation networks,[5] etc.). With the study in physical significance and mathematical properties of complex networks, many researches have found that the complex networks are structured in terms of communities which are groups of nodes defined on the basis of different concepts such as node similarity or comparison of initial graph with random graphs and so on.[6] Because communities in networks closely correspond to functional units in various fields, ranging from social groups in social networks[7] to sets of pages on related topics on the world-wide web (WWW),[1,2] the property of community structure is indeed of considerable importance. It allows us to build a bridge to understand the relationship between single node in microscopy and groups in macros-copy. Therefore, community structure detection is significant in many practical scenarios.

The criteria for defining community structure correspond to meaningful understandings of functions of groups.[8] In recent years, many efforts have been made to define the community structure. In the unweighted networks, the community structure is defined as groups of nodes in which edges are denser than those on the external between them.[9] This community structure reflects well the close connections between nodes and there is a large volume of researches having been devoted to the development of algorithms for detecting community structure in the unweighted networks.[10] However, apart from the topological issues that the interaction between two nodes is characterized by the existence of a link, it is important to realize that many complex networks are intrinsically weighted. For example, in a social network there are stronger and weaker social ties among individuals[7,11,12] and in WWW there are more and less data traffic among websites.[2,13] Moreover, in many cases the weights affect significantly the properties or functions of these networks, e.g., virus and disease spreading,[14−18] synchronization dynamics of

*Project supported by the Fundamental Research Funds for the Central Universities (Grant Nos. KYZ200916, KYZ200919 and KYZ201005) and the Youth Sci-Tech Innovation Fund, Nanjing Agricultural University (Grant No. KJ2010024).

†Corresponding author. E-mail: shen_yi1979@126.com

© 2011 Chinese Physical Society and IOP Publishing Ltd

oscillators,[19−23] and motif statistics.[24,25] It is clear that edge weights, as strength of the interaction represented by networks, are believed to be an important variable in networks. A good deal of information about the network's structure would be thrown away if we discard the weights of edges.

By considering that weights on edges taking greater values mean stronger connections for node pairs, Newman defined the community structure of the weighted networks as groups of nodes in which the edge's weights are relatively larger than those on the external between them.[26] The community structure can be detected by optimizing the weighted modularity Q^w proposed by Newman[26]

$$Q^w = \frac{1}{2T}\sum_{ij}\left[w_{ij} - \frac{T_i T_j}{2T}\right]\delta(c_i, c_j), \qquad (1)$$

where w_{ij} represents the weight on the edge between nodes i and j, T_i is the weight of node i: $T_i = \sum_j w_{ij}$, T is the sum of all edges' weights in networks, δ is the Kronecker delta symbol $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise and c_i is the label of the community to which node i is assigned.

Recently, there have been several algorithms[26−28] for detecting community structure in weighted networks proposed. The random walk based method[27] relies on a transition probability matrix that is achieved by evaluating the conductance between any two nodes, so the problem of slowness and inaccuracy inevitably exists especially in dense weighted networks.[27] The GN (Girvan and Newman) algorithm[8] and EO (extremal optimization) algorithm[29] are first developed for un-weighted networks. They are generalized to weighted version WGN (weighted GN) and WEO (weighted EO) by adding the extra operation of dividing the un-weighted edge betweenness[8] by the link weight[26] and replacing the modularity function Q[8] by Q^w[26] respectively. The WGN and WEO belong to the dividing algorithms, so they have high speed.[6] However one common point within them is that they delete only one edge each time to split the network, the modularity for evaluating the divided results must be calculated after each dividing process.[6] This makes them unable to directly use the information contained in edges' weights. Furthermore, based on optimizing the Q^w proposed by Newman, all of the three methods[26−28] are intended to divide a weighted network into several groups in which the edges' weights are relative larger than those on the external between them.

It is worthwhile pointing out that weights not only indicate particularly close connections but also the similarity between nodes' functions. The distribution of weights on edges contains more useful information we should strive to find out. Hence, the criteria for defining and evaluating community structure in weighted networks are not only limited to the groups in which the edges' weights are relatively larger than those on the external between them. In some cases, for example in communication networks, the data traffic between units described by the weights on the edges between nodes is relatively more uniform in the same group than the external between groups, which describes some of the common functions between units in the same group. In this paper, the community structure of the weighted networks is defined as groups of nodes in which the edge's weights distribute uniformly but at random between them. In order to detect such community structure effectively, a threshold coefficient κ to evaluate the equivalence of edges' weights and a new weighted modularity based on the weight's similarity are proposed. A weight's agglomerative method which can directly make use of the information contained in edge weights is presented. For a network with n nodes, the algorithm can detect the community structure in time $O(n^2 \log_2^n)$. The algorithm is analysed by random weighted networks with known community structure and a real network. Compared with WGN and WEO algorithms, the algorithm has the advantages of higher accuracy and precision. Furthermore, the method can find a special hierarchical organization which can describe the various steady connections between nodes in groups by changing the threshold coefficient κ.

2. The modularity functions based on weight's similarity

It is well known that the final dividing results based on optimizing modularity depend on the configuration models.[9] In the weighted networks, if nodes i and j have more edges with the same weights, a weighted edge is more likely to be assigned between them. Otherwise, if they have no edge with the same weight, then there should be no edge existing between them. Hence, if we observe in a real network that there is an edge between i and j if they have no edge with the same weight, it should be considered a big surprise. We can use this fact to construct a configuration model. We define w_{id} and w_{je} as the weights

of the d-th edge of node i and the e-th edge of node j respectively and define $\Phi(a,b) = I_{(||a|-|b||\leq\kappa)}$ as an event indicator function which means $\Phi(a,b) = 1$ when $||a| - |b|| \leq \kappa$ and 0 otherwise. The κ is a threshold and is used to evaluate the equivalence of edges' weights. We write the configuration model as

$$P_{ij}^{W} = \frac{\sum_{ij}\sum_{d\in k_i}\sum_{e\in k_j}\Phi(w_{id}, w_{ie})}{2m}. \tag{2}$$

Equation (2) describes the expected fraction of edges that have the similar weights. Then we give the weighted modularity in a similar form to Eq. (2)

$$Q^{W} = \frac{1}{2m}\sum_{ij}\left[A_{ij} - P_{ij}^{W}\right]\delta(g_i, g_j). \tag{3}$$

Optimizing Eq. (3) means that we should find the groups within which there are more edges with the similar weights than those between groups, which will lead to the partition of the weighted network that within groups the edges' weight distribute uniformly but at random between them. Obviously, such community structure can reveal the steady connections between nodes or some similarity between nodes' functions. Take social networks for example: the strength of social relationships between individuals, described by the weights on the edges between nodes, is relatively more uniform within groups than those between groups, which describe the various steady relationships between individuals in groups or some similarity between their functions. In some cloud-computing networks,[30] the data traffic between units described by the weights on the edges between nodes is relatively more uniform in the same group than that on the external between groups, which describes the units cooperating with each other to accomplish functions in the same group. Furthermore, the configuration model described by Eq. (2) is constructed on the basis of the edges' characteristic. It can describe the topologies and edges' information of networks simpler than on nodes. Equation (3) can be considered as a general expression of the un-weighted networks, the directed networks and the weighted networks. For the un-weighted networks, the edges' weights can be simply treated as being equivalent to each other. For the directed networks, we only need to add the information of direction to w_{id} and transform it into w_{id}^{in} or w_{id}^{out}.

3. The weight's agglomerative method

The aim of community structure detection on the basis of our proposed definition is to divide the network into several groups in which the edge's weights distribute uniformly but at random between them. So if we can agglomerate the edges with the same weight to form a connected group, the communities is obviously found. The weight's agglomerative method works as follows.

(i) Initially, each node is considered as a group. The weighted matrix $\boldsymbol{W}$ with element w_{ij} is constructed. The value of w_{ij} is the weights on the edges connecting groups i and j.

(ii) If the difference of more than one edge's weights in different groups are less than κ, the corresponding groups are agglomerated into one group and the Q^W is calculated.

(iii) If the Q^W obtained after Step (ii) increases, then we accept the agglomerating operation in Step (ii) and recalculate the $\boldsymbol{W}$. Go to Step (ii). If Q^W decreases, we go to Step (ii) to agglomerate other alternative groups.

The process is repeated until the maximal Q^W is obtained. Then the corresponding community structure is found.

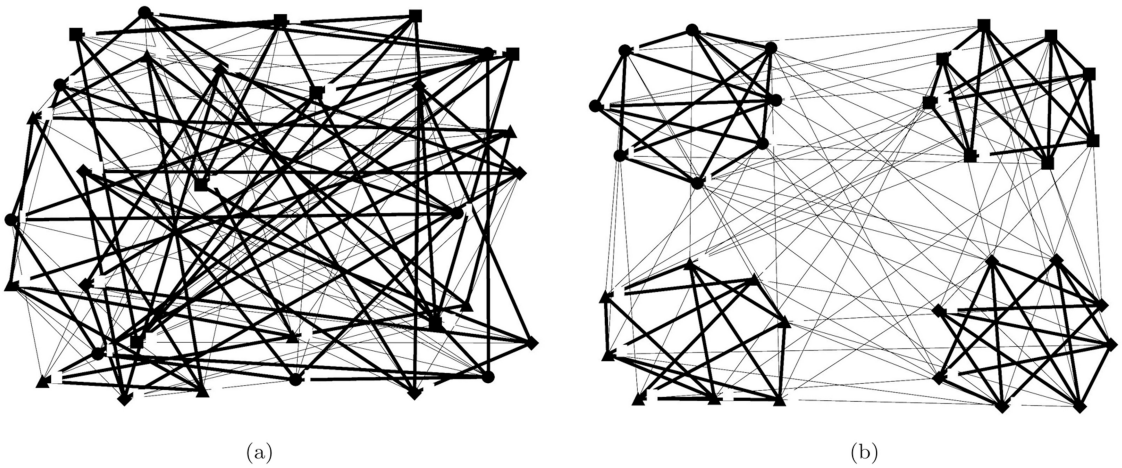
The agglomerating operation can be run in parallel, so for a network with n nodes, it takes at most $O(\log_2^n)$ time to agglomerate all groups. Once two groups are agglomerated, the Q^W should be calculated to make a judgement on the agglomerating operation. This part of computation complexity is mainly on the calculation of P_{ij}^W which is $O(n^2)$. Therefore, the total computation complexity of our method is $O(n^2 \log_2^n)$. It is much less than that of WGN which is $O(n^3)$[22] and has the same order of magnitude as that of WEO which is $O(n^2 \ln n)$.[26]

4. Experiments

In this section, we apply the Q^W-based weight's agglomerative method to several networks, including a simple random weighted network with four communities to show the function of our method and a set of random weighted and un-weighted networks with known community structure to investigate the accuracy and precision of our method. Finally, several real-world networks[7,31] are investigated.

4.1. A simple weighted network

We generate a simple weighted network of 32 nodes designed specifically to show the performance of the method. In this network, edges are placed independently at random between nodes pairs with some probability p. The edges are initially un-weighted and there is no community structure to be found in this network. Then we divide the network into four groups of 8 nodes each. The edges that fall within groups are assigned fixed weight of 1 and the edges between them are assigned weights randomly from 0.0 to 1.0 with one decimal. We set κ to be 0 here. Figure 1(a) shows the result of the method that ignores edges' weights and it finds no community structure at all in the network. Figure 1(b) shows the result of the method on the basis of our proposed modularity and it can correctly divide the random weighted network into the four groups we want.



(a) (b)

Fig. 1. The division of the random weighted network of 32 nodes. The thick edges indicate that they have the fixed weights of 1 assigned to them. The community assignments are denoted by node shape and colour.

4.2. Performance on the computer-generated weighted networks

4.2.1. The analysis of accuracy

The accuracy is defined as the consistence when the community structure from algorithm is compared with the presumed communities,[8] which can be quantitatively measured by calculating the fraction of correctly identified nodes according to the presumed communities. We test the accuracy of our method to a set of computer-generated networks. The computer generated networks have 128 nodes and are divided into four groups of 32 nodes each. Edges are placed such that the average degree is fixed at 16. The test is performed for the most deteriorative situation that each node has as many connections to nodes outside its own group as it does inside. This means that there is no community structure to be found in this network if we ignore edges' weights. Here we consider two cases, the first is that we assign a fixed weight of 1 to the edges within all four groups and the second is that we assign the fixed weights of 0.25, 0.50, 0.75 and 1 to the edges within groups 1, 2, 3 and 4, respectively. To the edges falling between groups we assign random weights in the range 0–1 with the minimum interval of 0.05 for the above two cases. We increase κ from 0 to 1 to investigate the accuracy of the method and make comparisons with the two recent methods WGN and WEO. The results are shown in Fig. 2.

As figure 2 shows, the three methods perform very well in the two cases from $\kappa = 0$ all the way to around $\kappa = 0.4$ and more than 80% of all nodes can be correctly identified. When $\kappa \geq 0.4$, the classification begins to deteriorate markedly because the threshold κ is too large to provide a precise measure metric system for evaluating communities. In the first case that the fixed weight of 1 is assigned to the edges within all four groups, our method performs better than WEO and WGN when $\kappa \geq 0.6$. The reason is that in each time, WEO algorithm moves one edge with the same weights as those in the identified group to optimize the Q^W. When κ is assigned to be larger values, it

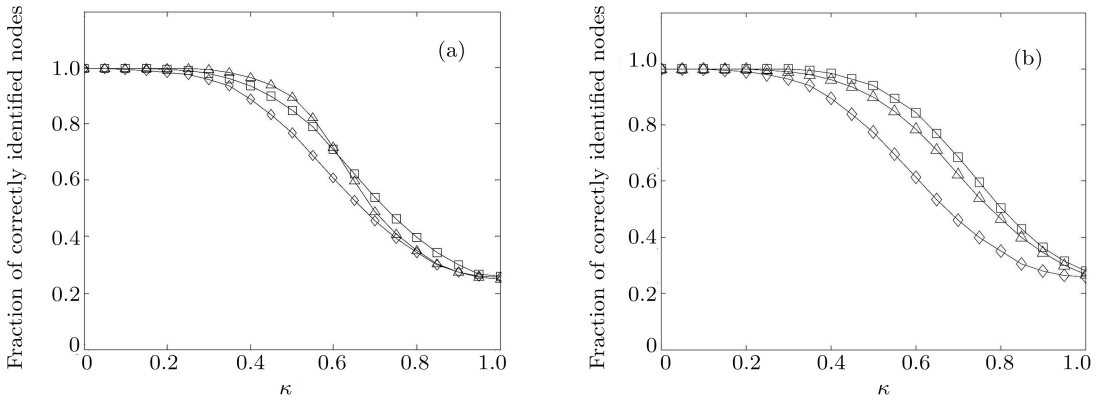


Fig. 2. The fraction of nodes of computer-generated random networks correctly identified by our method (squares), WEO algorithm (triangles) and WGN algorithm (diamonds) averaged over 300 networks, as a function of κ. Figures (a) and (b) show the results of the first and second cases respectively.

largely decreases the probability for WEO to move a 'correct' edge to optimize the Q^W. Otherwise, when κ is assigned to be small values, it increases the probability for WEO to move a 'correct' edge to optimize the Q^W, so WEO performs better than our method when $\kappa < 0.6$.

In the second case that the varied fixed weights are assigned to the edges within different groups, our method performs better than WEO and WGN all the way. This is because that in the second case, the edges in different identified groups have different weights. More edges are considered to have the same weights as those in the identified group by WEO. It largely decreases the probability for WEO to move a 'correct' edge to optimize the Q^W. When $\kappa = 1$, the communities are completely diffuse and it is impossible to find communities correctly by any algorithms and the fraction nodes are correctly identified to converge to about 0.25 by the two methods.

4.2.2. The analysis of precision

The precision is defined as the consistence among the community structures from different runs of an algorithm on the same network.[6] We use the normalized mutual information measurement (NMI)[6] which is proposed by Danon *et al.* to quantitatively measure the precision of an algorithm. The NMI is based on a confusion matrix $\boldsymbol{N}$, where the element N_{ij} is the number of nodes in the "real" community i that appears in the "found" community j. Here the "real" and "found" communities are the communities detected by two runs of an algorithm on the same network respectively. The value of $NMI(A, B)$ is the measure of the similarity between the partitions A and B. In order to effectively measure the precision of an algorithm, it needs more runs of an algorithm on the same network. We propose a matrix $\boldsymbol{NMI}$, where the element $NMI_{ij} = NMI(i, j)$ accounts for the similarity between the i-th detected community structure and the j-th detected community structure. If we run k times of an algorithm, the precision denoted by $P_{\rm pre}$ is

$$P_{\rm pre} = \frac{\sum_{i}^{k} \sum_{j(j\neq i)}^{k} NMI_{ij}}{k^2 - k}. \tag{4}$$

The denominator accounts for the summation of NMI_{ij} and the numerator accounts for the number of NMI_{ij} added in the denominator. If the community structures detected by different runs of an algorithm are identical to each other, the $P_{\rm pre}$ will take its maximum value of 1. If the communities detected by different runs of an algorithm are totally independent of each other, the $P_{\rm pre}$ will take its minimal value of 0.

For a given κ, we construct 20 networks according to the rule described in the Subsubsection 4.2.1. Then we run each algorithm 10 times to find communities in each network. The value of $P_{\rm pre}$ is an average over 20 networks and 10 runs each. Figure 3 shows the results of our proposed algorithm and WEO algorithm in terms of precision.

From Fig. 3 we can see, for small κ, there is no discrimination between the two methods. However when $\kappa > 0.5$, the precision of the WEO algorithm decreases much more quickly than our method and WGN algorithm as κ increases. The reason is that WEO algorithm moves an edge to optimize the Q^W each time.

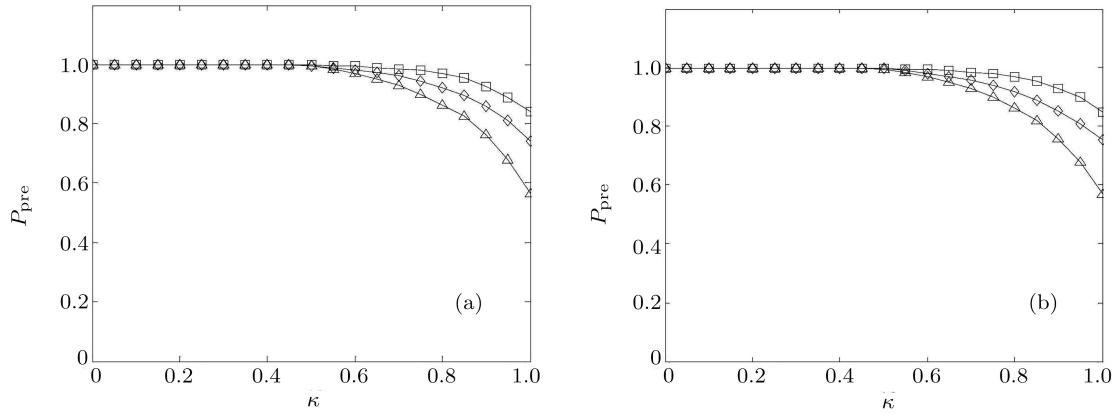


Fig. 3. The precision of our method (squares), WEO algorithm (triangles) and WGN algorithm (diamonds). Each point is an average over 20 networks and 10 runs each, as a function of κ. Figures (a) and (b) show the results of the first and second cases respectively.

With the increase of κ, more and more candidate edges can be moved to optimize the Q^W. In other words, in different runs of WEO on the same network the probability of selecting the same edge to move decreases greatly as κ increases. In the two cases, our method performs much better than the WEO and WGN algorithms. This is because the fact that our method can be considered as a global method for it agglomerates nodes in parallel. In different runs on the same network, our method divides the same edge into the same group easier than WEO. When $\kappa = 1$, the precision of our method is over 0.8, whereas the WEO algorithm is only up to 0.58 and WGN is only 0.75.

4.3. A scientific collaboration network

As a real example, we apply our method to a scientific collaboration network.[31] The network contains 126 nodes which represent research members. Two nodes are connected if they coauthored articles during 1995. Integer weights are assigned on edges equal to the total number of articles they coauthored. This way of attributing weights clearly represents the weighted interactions among the members of the department. We investigate the partition of the network into groups which represent communities according to our definition. In order to distinctly show the dividing process and results of our method, we first show the scientific collaboration network in Fig. 4(a). Then, the community structure revealed by our analysis is shown in Fig. 4(b) as a function of κ from 5 to 2.

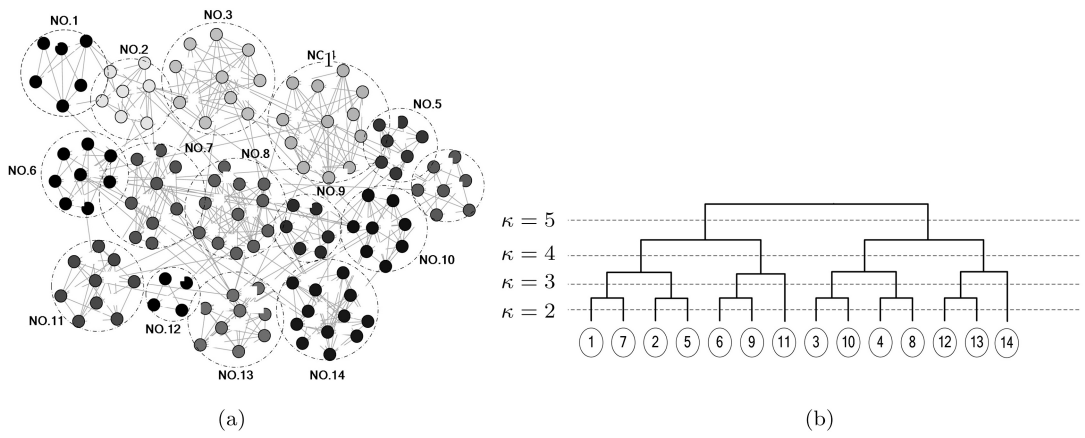


Fig. 4. (a) The collaboration network. (b) Plot of the dendrogram indicates the division of the scientific collaboration network. Each dashed line labeled with ζ indicates the dividing sequence of the network as κ decreases.

It is easy to know that the members interested in the same topics coauthored more often and uniformly than those interested in different topics. This appears to be a correlation between the topics and the divisions of the research members, which exhibits the formation of communities. In our method we use κ as a threshold to evaluate the equivalence of edges' weights, it means that the dynamical movement of κ is intrinsically connected with the hierarchical organization of communities. When κ is assigned large value, the corresponding big communities detected by the method show us a relative broad field of studying in which the researchers are interested. When ζ is assigned small value, the research members who have more steady and strong cooperative ties can be detected as one community. Here when $\kappa = 2$, the 14 communities detected by our method just correspond to the 14 research groups except several nodes are divided into other communities. When $\kappa = 1$, the collaboration network is completely shattered and divided into 41 fractions. The biggest one only has 5 nodes and the majorities have 3 or 4 nodes. Although the steadiest and strongest cooperative ties between the several members are reveled, these fractions cannot be considered as conventional communities[8] since they have too few nodes. From the above analyses, we can see that the groups found by our method distinctly reveal the hierarchical organization of collaboration relationship between members.

4.4. Other applications

For further applications and tests of our proposed method, we have applied it to two large weighted networks. One is the United States college football network.[7] The 115 nodes in the graph represent teams (identified by their college names) and 613 edges represent matches between two teams, integer weights are assigned on edges equal to the number of matches between two colleges. Since more competitions exist among the strong teams and fewer competitions exist among week teams, by optimizing Eq. (3) our method divides the more successful and less successful teams into different groups. The result efficiently reveals the various steady interactions between strong teams and weak teams. The other is the friendship networks constructed from the in-school questionnaires of the Add Health[31] friendship nomination study from the period 1994–1995, in which 75871 students in the United States were investigated. Five main communities were detected which are approximated to the race classification, i.e. white, black, Hispanic, Asian and mixed, which is in good agreement with their steady social relationships.

5. Conclusions

Differing from the conventional community detection in weighted networks where one always wants to divide the weighted network into several groups in which the edges' weights are relatively larger than those on the external between them, in this paper, we focus on dividing the weighted network into several groups within which the edges' weights distribute uniformly but at random between them. This division can effectively reveal the steady and strong cooperative ties between nodes. To achieve this goal, we propose a threshold coefficient κ to evaluate the equivalence of edges' weights and a new weighted modularity based on the weight's similarity. Then by constructing the weighted matrix and using the agglomerative mechanism, we present a weight's agglomerative method based on optimizing the modularity to detect communities. For a network with n nodes, the algorithm can detect the community structure in time $O(n^2 \log_2^n)$. It is much less than that of WGN which is $O(n^3)$[22] and has the same order of magnitude as that of WEO which is $O(n^2 \ln n)$.[26] Simulations on networks show that the algorithm has higher accuracy and lower computation complexity than the existing techniques. By changing κ, we can find a special hierarchical organization which can describe the hierarchical steady and strong connections between nodes in weighted networks. It leads to new insights into weighted networks and sheds light on understanding the relationships of groups.

References

[1] Albert R, Jeong H and Barabási A L 1999 *Nature* **401** 130

[2] Raúl P O, Julio S, Ana P and José A G 2009 *Comput. Commun.* **32** 1118

[3] Newman M E J 2001 *Proc. Natl. Acad. Sci. USA* **98** 404

[4] Watts D J and Strogatz S H 1998 *Nature* **393** 440

[5] Redner S 1998 *Euro. Phys. J.* B **4** 131
[6] Danon L, Duch J and Guilera A D 2005 *J. Stat. Mech.* P09008
[7] Girvan M and Newman M E J 2002 *Proc. Natl. Acad. Sci. USA* **99** 7821
[8] Newman M E J and Girvan M 2004 *Phys. Rev.* E **69** 026113
[9] Newman M E J 2006 *Proc. Natl. Acad. Sci. USA* **103** 8577
[10] Lancichinetti A and Fortunatol S 2009 *Phys. Rev.* E **80** 056117
[11] Bian Q X and Yao H X 2010 *Acta Phys. Sin.* **59** 3027 (in Chinese)
[12] Pu C L and Pei W J 2010 *Acta Phys. Sin.* **59** 3841 (in Chinese)
[13] Meng Q K and Zhu J Y 2009 *Chin. Phys.* B **18** 3632
[14] Liu M X and Ruan J 2009 *Chin. Phys.* B **18** 2115
[15] Estrada E 2006 *Euro. Phys. J.* B **52** 563
[16] Song Y R and Jiang G P 2009 *Acta Phys. Sin.* **58** 5911 (in Chinese)
[17] Hu K, Hu T and Tang Y 2010 *Chin. Phys.* B **19** 080206
[18] Lü L and Zhang C 2009 *Acta Phys. Sin.* **58** 1462 (in Chinese)
[19] Li Z Y, Xie Z W, Chen T and Ouyang Q 2009 *Chin. Phys.* B **18** 5544
[20] Wang X H, Jiao L C and Wu J S 2010 *Chin. Phys.* B **19** 020501
[21] Jin X Z 2010 *Chin. Phys.* B **19** 080508
[22] Bernardo M D, Garofalo F and Sorrentino F 2007 *Int. J. Bifur. Chaos* **17** 3499
[23] Wu X J and Lu H T 2010 *Chin. Phys.* B **19** 070511
[24] Onnela J P, Saramaki J, Kertesz J and Kaski K 2005 *Phys. Rev.* E **71** 065103
[25] Chen W D, Xu H and Guo Q 2010 *Acta Phys. Sin.* **59** 4514 (in Chinese)
[26] Newman M E J 2004 *Phys. Rev.* E **70** 056131
[27] Alves N A 2007 *Phys. Rev.* E **76** 036101
[28] Fan Y, Li M, Zhang P, Wu J S and Di Z R 2006 *Physica* A **370** 869
[29] Duch J and Arenas A 2005 *Phys. Rev.* E **72** 027104
[30] Grossman R 2009 *IT Professional* **11** 23
[31] Frank K A 1996 *Soc. Networks* **18** 93