Physica A 397 (2014) 129–143

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Closed walks for community detection

Yang Yang [a], Peng Gang Sun [b], Xia Hu [c], Zhou Jun Li [a,*]

[a] State Key Laboratory of Software Development Environment, Beihang University, 100191, China
[b] School of Computer Science and Technology, Xidian University, 710071, China
[c] Department of Computer Science and Engineering, Arizona State University, United States

HIGHLIGHTS

- Our method performs well on computer generated networks and real-world networks.
- Closed walks of small orders are basic elements in constructing community structure.
- Our method is a better tradeoff of accuracy and runtime.
- Our method is a novel way to solve the double peak structure problem.
- Our work can provide us with a new perspective for understanding community structure in complex networks.

ARTICLE INFO

Article history:
Received 21 August 2012
Received in revised form 1 November 2013
Available online 7 December 2013

Keywords:
Complex networks
Community structure
Edge clustering coefficient
Closed walks

ABSTRACT

In this paper, we propose a novel measure that integrates both the concept of closed walks and clustering coefficients to replace the edge betweenness in the well-known divisive hierarchical clustering algorithm, the Girvan and Newman method (GN). The edges with the lowest value are removed iteratively until the network is degenerated into isolated nodes. The experimental results on computer generated networks and real-world networks showed that our method makes a better tradeoff of accuracy and runtime. Based on the analysis of the results, we observe that the nontrivial closed walks of order three and four can be considered as the basic elements in constructing community structures. Meanwhile, we discover that those nontrivial closed walks outperform trivial closed walks in the task of analyzing the structure of networks. The double peak structure problem is mentioned in the last part of the article. We find that our proposed method is a novel way to solve the double peak structure problem. Our work can provide us with a new perspective for understanding community structure in complex networks.

1. Introduction

In recent years, some efforts [1–3] have been made to show that community structures are frequently observed in most complex systems. The community structure is a set of nodes with more internal links than external. The task of community detection is to find these community structures. Community detection is of great importance, because it can help us to understand the organization and function of the systems, the dynamics and evolution of the network, and so on. Furthermore, the algorithms for community detection are widely used in many fields, such as the Internet and the World Wide Web [4–6], epidemiology networks [7–10], biological networks (PPI networks [11], metabolic networks [12,13], ecological webs [14,15]), social networks(political blogs [16,17], and co-authorship networks [18]).

* Correspondence to: Beihang University, Xueyuan Road No. 37, Haidian District, 100191, Beijing, China. Tel.: +86 13910520973.
E-mail addresses: yangyangfuture@gmail.com (Y. Yang), psun@mail.xidian.edu.cn (P.G. Sun), huxia001@gmail.com (X. Hu), lizj@buaa.edu.cn (Z.J. Li).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2013.11.034

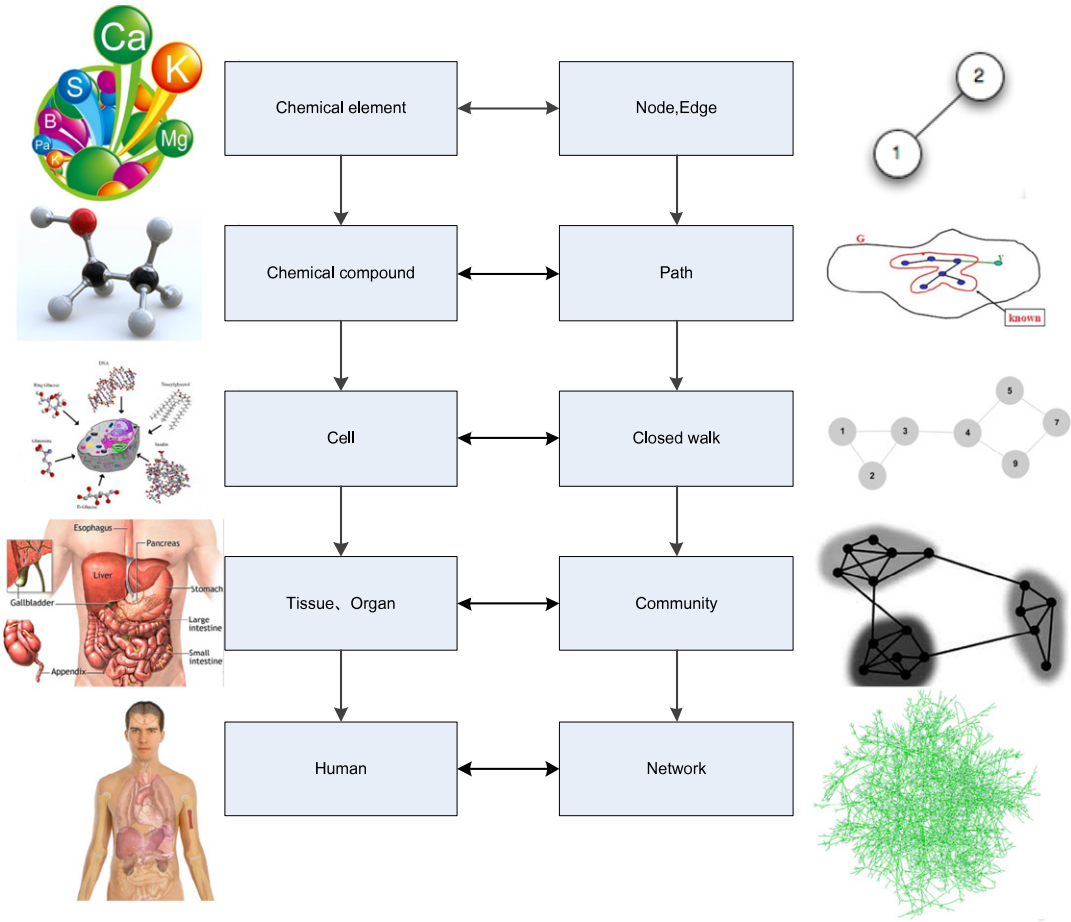


Fig. 1. Illustration of community structure by human beings.

Over the years, many algorithms have been proposed based on the analysis of two basic elements of networks: nodes and edges. For example, Breiger et al. [19] proposed the agglomerative method for community detection based on the node similarity. By using the dissimilarity index between the nearest-neighboring nodes, a divisive method for community identification is given by Zhou [20]. Analogously, Wu et al. [21] presented the core attachment method, which considered the inherent organization within protein complexes [21]. Frey et al. [22] proposed an approach to discover communities by passing messages between data points [22]. In Ref. [23], the communities were reinvented as groups of links and could be detected by analyzing the attributes of edges in complex networks. Duch and Arenas proposed a community detection algorithm by using extremal optimization [24], while R. Guimerà et al. discussed the modularity from fluctuations in random graphs and complex networks [25]. In addition, some algorithms focused on the analysis of paths to detect communities, such as the betweenness-based method [1] and the information centrality-based method [3].

We illustrate the community structures of real world networks in Fig. 1. At different levels, human beings can be viewed as tissues/organs, cells, chemical compounds and chemical elements. Human beings are composed of tissues and organs. Tissues and organs are constructed from cells. Cells consist of many chemical compounds, while chemical compounds are composed of numerous chemical elements. Likewise, complex networks can be viewed at different levels as well. Many algorithms focus on the analysis of the nodes and paths to detect communities in complex networks; however, few methods focus on special subgraphs, such as closed walks, which are another type of element for community formation.

In this paper, we propose a novel measure to detect communities. By repeatedly calculating the number of closed walks with different orders for edges, then removing the edges with the lowest value, the network will be broken into smaller groups. According to the experimental results and our analysis, we find that nontrivial closed walks of order 3 and 4 can be considered as basic elements in constructing community structures. Nontrivial closed walks outperform trivial closed walks for analyzing the structure of networks. Furthermore, our method is a novel way to solve the double peak problem. In short, our work can provide us with a new perspective for understanding community structure in complex networks.

The rest of this paper is organized as follows. In Section 2, we discuss some basic concepts and the rationale of our method. In Section 3, we evaluate our method on an analog network. In Section 4, we test our method on several computer generated

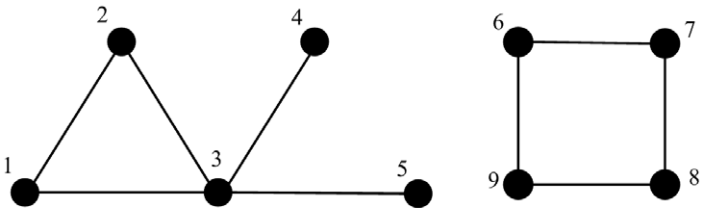


Fig. 2. Illustration of graphs.

Table 1
The types of walks.

Type		Walks	
Open walks		$4 \to 3 \to 5$	
Closed walks	Trivial closed walks	$1 \to 2 \to 1 \to 2 \to 1$	$3 \to 4 \to 3 \to 5 \to 3$
	Nontrivial closed walks	$1 \to 2 \to 3 \to 1$	$6 \to 7 \to 8 \to 9 \to 6$

networks. In Section 5, we apply our method on real world networks and analyze the experimental results. The double peak structure problem is also solved in this part. Finally, we briefly conclude the paper in Section 6.

2. Our method for finding community structures

The network we analyzed can be represented as a connected, simple undirected graph G which contains n nodes and m edges. $A = (a_{ij})_{n\times n}$ is the adjacency matrix of graph G. If two nodes i and j are connected by an edge, then two nodes are adjacent and a_{ij} is equal to 1, otherwise a_{ij} is equal to 0. The entry a_{ii} on the main diagonal is set to 0. In the following subsections, we first introduce some definitions and concepts, and then our method.

2.1. Closed walks

A closed walk, which is directly related to the subgraph of the network [26,27], is a kind of walk which starts and ends at the same node. For instance, in Fig. 2, $\{1, 2, 3\}$ forms a closed walk: $1 \to 2 \to 3 \to 1$, while $\{4, 3, 5\}$ forms an open walk: $4 \to 3 \to 5$. As the types of walks illustrated in Table 1, we divide closed walks into two types: trivial and nontrivial [26,28,29]. If all nodes are distinct in a closed walk, except for the start node and end node, then the closed walk is categorized as a nontrivial closed walk. Otherwise, it belongs to the trivial closed walks. It is noteworthy that all closed walks of order 3 are nontrivial.

2.2. Edge clustering coefficient

The edge clustering coefficient [2] is similar to the concept of the node clustering coefficient. It equals the number of cyclic structures to which a given edge belongs, divided by the number of cyclic structures that might potentially include it. If there is an edge between node i and j, the definition of the edge clustering coefficient is

$$C_{i,j} = \frac{z_{i,j}^{(g)}}{\min[(k_i - 1), (k_j - 1)]} \tag{1}$$

where $z_{i,j}^{(g)}$ counts the times that the edge belongs to cyclic structures of order g and k_i and k_j are respectively equal to the degree of nodes i and j. $\min[(k_i - 1), (k_j - 1)]$ is the maximal possible number of cyclic structures. g, the order of cyclic structures, is from 3 to infinity.

The purpose of the formula is that the edges that connect communities are likely to join few or no cyclic structures. Consequently, these edges will have small values of $C_{i,j}$. When the number of cyclic structures is zero, $C_{i,j} = 0$. To avoid this problem, $B_{i,j}^{(g)}$ is equal to $z_{i,j}^{(g)} + 1$.

The formula is as follows:

$$C_{i,j} = \frac{B_{i,j}^{(g)}}{\min[(k_i - 1), (k_j - 1)]}. \tag{2}$$

Based on the concept of edge clustering coefficient, Radicchi et al. [2] checked the accuracy of their method by comparing its performance with the Girvan and Newman method. It turns out that their method's performance is not always as well as Girvan and Newman method's performance in some cases.

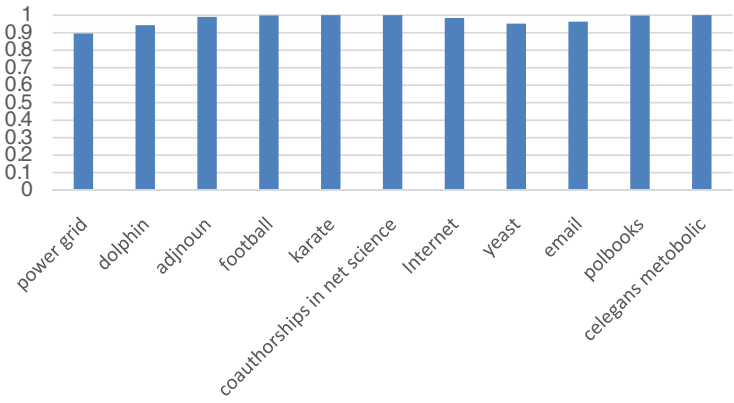


Fig. 3. The proportion of edges in the order 3 and 4 closed walks.

2.3. Modularity

We know that a good division is obtained if most of the edges fall into the same community, with relatively few edges connecting different communities. In order to evaluate how good a division of G is, a modularity function Q is proposed by Girvan and Newman [1,30]. To calculate the *modularity*, a symmetric matrix **e** is introduced. If we divide the network into n communities, we can get an $n \times n$ symmetric matrix **e**. The element e_{ij} of matrix **e** is the fraction of the number of edges over total edges that connect community i and community j [1]. The trace of the matrix $Tr\ e = \sum_i e_{ii}$ shows the number of edges that fall into the same community. The row sums $a_i = \sum_j e_{ij}$ represent the number of edges that connect two vertices in community i. In a network in which edges fall between vertices without regard to the communities they belong to, we would have $e_{ij} = a_i a_j$ [1]. In this case, the *modularity* is defined as

$$Q = \sum_i (e_{ii} - a_i^2) = Tr\ e - \left\| e^2 \right\| \tag{3}$$

where $\left\| e^2 \right\|$ represents the sum of the elements of the matrix e^2. This quantity then measures the degree of correlation between the probability of having an edge joining two sites and the fact that the sites belong to the same community [3]. In the beginning, if we do not divide the single community into subsets, the value of Q equals 0. Afterwards, we repeatedly remove the edge with the lowest (or highest) score. The value of Q gets higher as we divide the community into several communities. We will get a strong community structure if the value of Q approaches a peak. Actually, values of Q for networks typically fall in the range from about 0.05 to 0.7, according to our algorithm. To understand the process, we plot the Q corresponding to the communities after each removal of edges in Fig. 7. The plot is detailed in Section 4.

2.4. Our method for community detection

In this subsection, we explain why our method takes nontrivial closed walks of order 3 and 4 into account. To evaluate the edges, it is unreasonable to take only one type of closed walk into consideration. As for Fig. 5, a common structure in Section 3, we cannot identify the right community structure by using closed walks of order 3 or 4. It has been observed that closed walks of other types are significant in real networks. Consequently, in order to detect community structure correctly, we should take into account closed walks of different types together.

Theoretically, the community structure of the whole network is composed of closed walks of different orders. Closed walks of order 3 and 4 are frequently observed in complex networks. Although the number of closed walks of order 5 is huge, statistical data shows that almost all these closed walks consist of closed walks of order 3 and 4. As shown in Fig. 3, the proportion of edges that participated in closed walks of order 3 and 4 reaches 90%, even in sparse networks (power grid, yeast network). In other networks, the proportion is at least 94.34%. This means that we can utilize closed walks of order 3 and 4 to evaluate almost all the edges. Thus we only need to consider closed walks of order 3 and 4. This is in accordance with the short circle property (SCP) [31]. In that paper, short circle means a closed walk of small order. Agarwal et al. constructed a correlated keyword graph and detected emerging topics by using the SCP.

Actually, the reason that we neglect closed walks of order 5 and choose closed walks of order 3 and 4 has sociological significance. Nicholas A. Christakis [32] wrote that "we are connected to everyone by 6 degrees and influence those up to 3 degrees". Moreover, he further explained the reasons why the influence dissipates after 3 degrees: (1) Intrinsic Decay: corruption of information. (2) Network Instability: ties become unstable at 4+ degrees of separation. (3) Evolutionary Purpose: we evolved in small groups where everyone was connected by 3 degrees or less. In other words, in Fig. 4(a), node

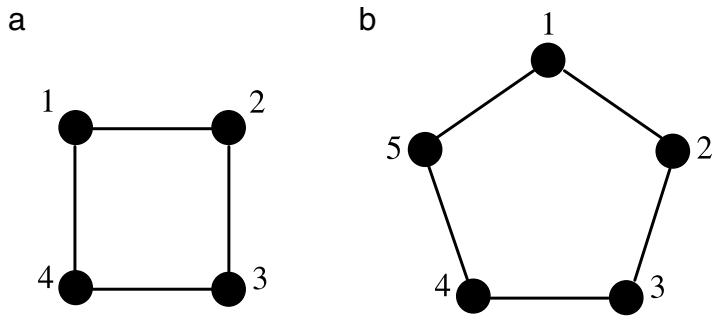


Fig. 4. Closed walks of order 4 and 5.

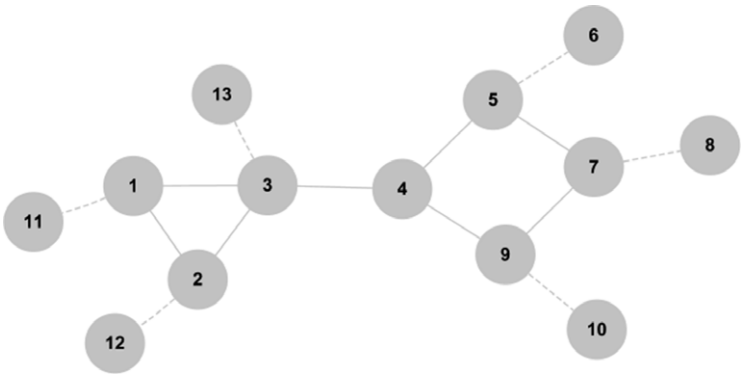


Fig. 5. An analog network composed of 13 nodes and 14 edges. Apparently, we can divide this network into two communities: {1, 2, 3, 11, 12, 13} and {4, 5, 6, 7, 8, 9, 10}.

1 can directly influence node 2 and indirectly influence node 2 by the path $1 \to 4 \to 3 \to 2$. However, if the order of a closed walk is larger than 4, for example the pentagon in Fig. 4(b), the indirect influence from node 1 to node 2 decreases sharply and this closed walk becomes instable. In this case, we merely take into account nontrivial closed walks of order 3 and 4.

In addition, information usually propagates along paths without repeated nodes. For instance, in Fig. 4(a), node 1 propagates information to node 3 by the path $1 \to 4 \to 3$. It is meaningless that node 1 propagates information to node 3 by the path $1 \to 4 \to 1 \to 4 \to 3$. Thus, nontrivial closed walks without repeated nodes are more useful.

Experimentally, we take closed walks of different orders (order 3, 4, 5) into consideration. We also take open walks, trivial and nontrivial closed walks into account. The formula is as follows (α, β, $\gamma = 0$ or 1):

$$C_{i,j} = \alpha \frac{z_{i,j}^{(3)} + 1}{\min[(k_i - 1), (k_j - 1)]} + \beta \frac{z_{i,j}^{(4)} + 1}{\min[(k_i - 1), (k_j - 1)]} + \gamma \frac{z_{i,j}^{(5)} + 1}{\min[(k_i - 1), (k_j - 1)]}. \tag{4}$$

Thus, we can get several combinations. However, our method takes nontrivial closed walks of order 3 and 4 into consideration. Thus, $\alpha = 1, \beta = 1, \gamma = 0$, $z_{i,j}^{(3)}$ and $z_{i,j}^{(4)}$ are respectively equal to the number of closed walks of order 3 and nontrivial closed walks of order 4 that an edge participated in.

The formula of our method is as follows:

$$C_{i,j} = \frac{z_{i,j}^{(3)} + 1}{\min[(k_i - 1), (k_j - 1)]} + \frac{z_{i,j}^{(4)} + 1}{\min[(k_i - 1), (k_j - 1)]}. \tag{5}$$

We use formula (5) to evaluate the edges and assume that the edges that lie inside the communities have the highest values, while the edges that lie between communities are those with the lowest values. Our algorithm for finding communities is a divisive method, by consistently removing edges with the lowest $C_{i,j}$ until the network breaks up into components.

Table 2
The community structure of the network in Fig. 5 processed by our method, the Girvan and Newman method, the method based on edge clustering coefficient (order $g = 3$, $g = 4$) and the information centrality method.

Method	Q	Community structure
Girvan and Newman method	0.4260	{1, 2, 3, 11, 12, 13}, {4, 5, 6, 7, 8, 9, 10}
Edge clustering coefficient $g = 3$	0.3661	{1, 2, 3, 11, 12, 13}, {4}, {5, 6}, {7, 8}, {9, 10}
Edge clustering coefficient $g = 4$	0.3546	{1, 11}, {2, 12}, {3, 13}, {4, 5, 6, 7, 8, 9, 10}
Edge information centrality	0.4260	{1, 2, 3, 11, 12, 13}, {4, 5, 6, 7, 8, 9, 10}
Our method	0.4260	{1, 2, 3, 11, 12, 13}, {4, 5, 6, 7, 8, 9, 10}

The detail of the algorithm is as follows:

Algorithm 1 Closed walks for community detection algorithm

1: **Input**: Graph G, edge set E, node set N.
2: **Output**: Label of community i and the member set E_i.
3: **Process**:
4: Read the data file of the graph G and initialize the deleted edge set D.
5: **while** $D \neq E$ **do**
6: **for** each edge $\notin D$ **do**
7: Calculate the closed walks of different orders that an edge participated in.
8: Calculate the value for the edges $\notin D$.
9: **end for**
10: Remove the edge(s) with the lowest value and add the edge(s) to set D.
11: Calculate and record the modularity Q of the new network, the label of community i and the member set E_i.
12: **end while**
13: **return** label of community i and its member set E_i.

The main step requires a time of the order of m, which is the total number of edges in the network. As in the Girvan and Newman algorithm [1], it is important to recalculate the scores every time after an edge has been removed. Furthermore, this operation has to be repeated for all remaining edges, which does not scale with m. So we can evaluate the total time as $O(a\,m + b\,m^2)$, while the Girvan and Newman method takes a time $O(mn)$ [1]. The information centrality method takes a time $O(m^3 n)$ [3,33]. In conclusion, the Girvan and Newman method runs faster than our method, however our method outperforms the Girvan and Newman method according to the experimental results.

3. Testing the method on an analog network

In this section, our purpose is to show that order 3 and 4 closed walks are equally important in identifying communities. Consequently, we constructed an analog network with 13 nodes and 14 edges in Fig. 5. There exists a nontrivial closed walk of order three {1, 2, 3} and a nontrivial closed walk of order four {4, 5, 7, 9} in the network. Apparently, the network can be divided into two communities by removing the edge between node 3 and node 4. We apply our algorithm together with other algorithms to this analog network and use the modularity function to identify what the community structure is. The results are presented in Table 2.

According to the results in Table 2, the edge clustering coefficient method $g = 3$ or 4 [2] cannot identify the right community structure. The edge clustering coefficient method ($g = 4$) splits the network into four communities: {1, 11}, {2, 12}, {3, 13}, {4, 5, 6, 7, 8, 9, 10}. Obviously, the community {1, 2, 3, 11, 12, 13} is wrongly classified. The edge clustering coefficient method ($g = 3$) splits the network into five communities: {1, 2, 3, 11, 12, 13}, {4}, {5, 6}, {7, 8}, {9, 10}. Obviously, the community {4, 5, 6, 7, 8, 9, 10} is wrongly classified. Since our method takes closed walks of order 3 and nontrivial closed walks of order 4 into consideration, we can successfully get a correct split {1, 2, 3, 11, 12, 13}, {4, 5, 6, 7, 8, 9, 10}. The Girvan and Newman method and the information centrality method can also get a correct split.

The structure of this analog network is not a special one, however, it is commonly observed in complex networks. The leaf nodes in Fig. 5 are not all necessary. If we delete all leaves in Fig. 5 and add an arbitrary number of leaves to node 1 and node 2, the edge clustering coefficient $g = 4$ method will wrongly detect the community. In addition, if we add an arbitrary number of leaves to node 5, node 6 and node 7, the edge clustering coefficient $g = 3$ method cannot get the correct division. The reason that we add leaf nodes 6, 8, 10, 11, 12, 13 is that we try to construct a network where the edge clustering coefficient method $g = 3$ and the edge clustering coefficient method $g = 4$ do not work at the same time. In this case, we can interpret the difference among several methods in only one figure.

4. Testing the method on computer generated networks

We test our algorithm on computer generated networks which are well defined 128-node random networks. The random walk networks are generated as follows: we generated a large number of graphs with $n = 128$ vertices and divided them into

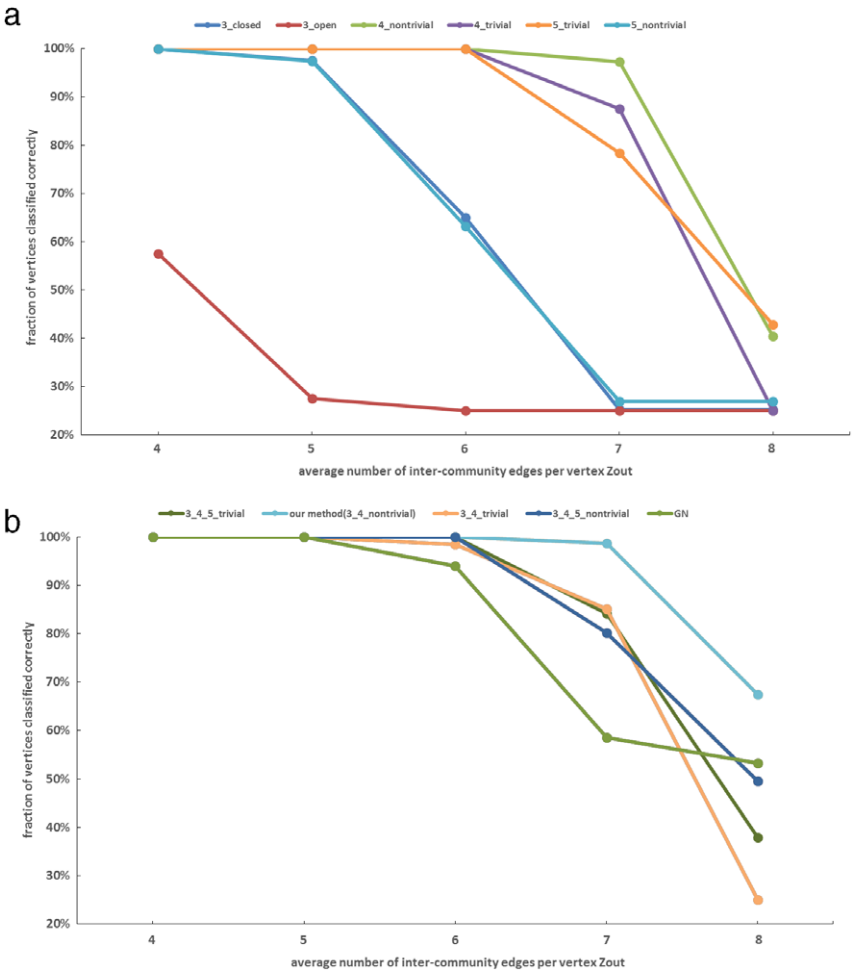


Fig. 6. The methods applied to 128-node random networks.

four communities of 32 vertices each, which are the groups 1–32, 33–64, 65–96 and 97–128. Edges are placed independently at random between vertex pairs with probability P_{in} for an edge to fall between vertices in the same community and P_{out} to fall between vertices in different communities. The values of P_{in} and P_{out} are chosen to make the expected degree of each vertex equal to 16. We have on average Z_{in} edges connecting two nodes that are in the same communities and Z_{out} edges connecting two nodes in different communities. The relationship between Z_{in} and Z_{out} is $Z_{in} + Z_{out} = 16$.

We apply community detection methods on the 128-node random walk networks. The experimental results are represented in Fig. 6. Each type of line in the figure depicts the accuracy as a function of the average number Z_{out} of edges from each vertex to vertices in other communities. In Fig. 6(a), according to the formula (5), the graphic symbol 3_open means that $\alpha = 1$, $\beta = 0$, $\gamma = 0$, and $z_{i,j}^{(3)}$ is equal to the number of order 3 open walks that an edge participated in. The graphic symbol 3_4_5_nontrivial in Fig. 6(b) means $\alpha = 1$, $\beta = 1$, $\gamma = 1$. $z_{i,j}^{(3)}$, $z_{i,j}^{(4)}$, $z_{i,j}^{(5)}$ are respectively equal to the number of closed walks of order 3, the number of nontrivial closed walks of order 4 and the number of closed walks of order 5 that an edge participated in. Obviously, according to the results in Fig. 5, we find that the combination ($\alpha = 1$, $\beta = 1$, $\gamma = 0$, $z_{i,j}^{(3)}$ and $z_{i,j}^{(4)}$ are respectively equal to the number of order 3 closed walks and nontrivial closed walks of order 4) outperforms the others. The experimental results agree with our analysis. We analyze artificial networks for various Z_{out}, ranging from 4 to 8, with a step of 1. As many algorithms can always find the correct classes when $0 \leq Z_{out} \leq 4$, we do not analyze them. For each value of Z_{out}, 100 samples are produced. We analyze the accuracy of this algorithm by comparing its performance with the Girvan and Newman method and the edge-clustering coefficient method. The information centrality method is not considered, as the method takes a time $O(m^3 n)$. According to the results in Fig. 6, we find that our algorithm and the Girvan and Newman method perform equally well in the horizontal axis sector [4,5]. Our algorithm leads from the horizontal axis sector [5,8]. Edge clustering coefficient and our algorithm are equally good in the horizontal axis sector [4,6]. In the horizontal axis sector [6,8], where community structures are very hard to detect, our algorithm clearly performs better, while the Girvan and Newman method inevitably starts to fail in detecting communities.

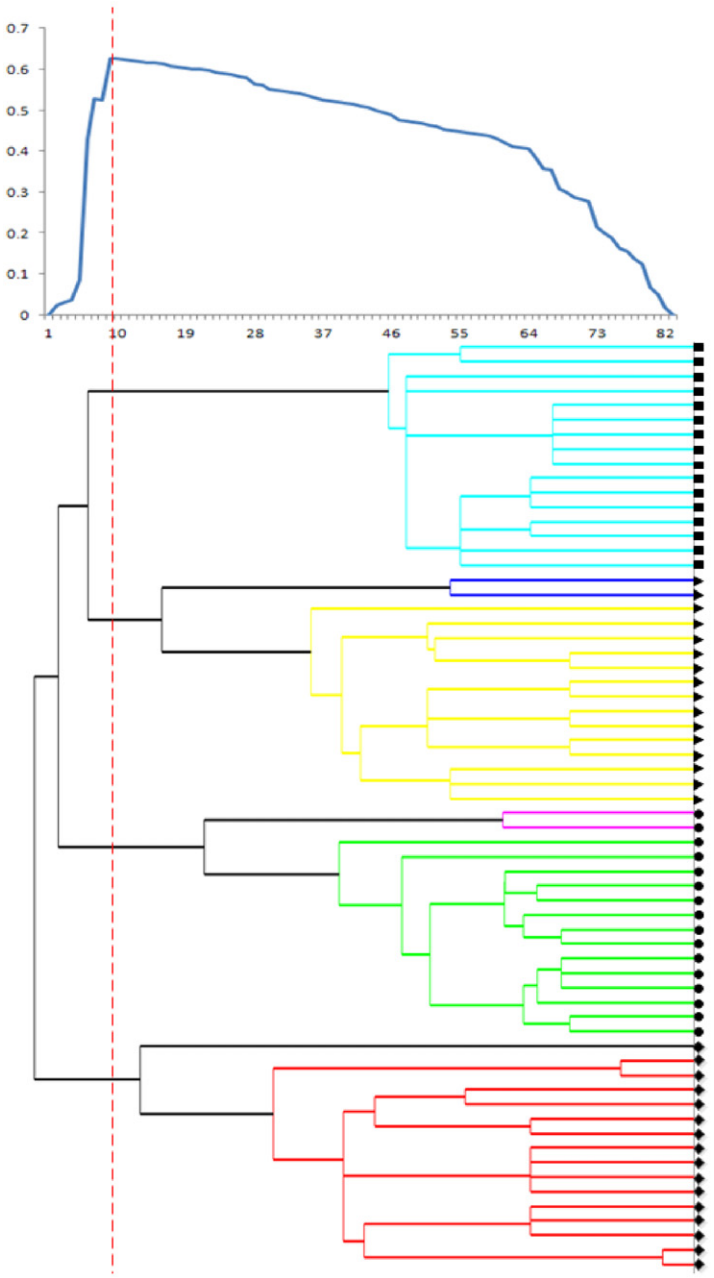


Fig. 7. Dendrogram of the communities in a 64-node random network. This network has been obtained by dividing the nodes into 4 groups of 16 nodes each and considering $Z_{in} = 7$ and $Z_{out} = 1$. When the modularity reaches its peak, we get a good partition of the network.

To illustrate our purpose clearly, we take a 64-node random network with 8 edges per node instead of a 128-node random network. We show an example in Fig. 7, which is a dendrogram of the communities found by applying our algorithm. Here, $Z_{in} = 7$ and $Z_{out} = 1$, i.e. the network is strongly clustered. We analyze the hierarchical tree to figure out which of the divisions is a proper split for the 64-node random network by using the measure of the cohesiveness of the communities. In Fig. 7, the x-coordinate represents the number of steps of the algorithm. As the structure of network changes, the value of Q (y-coordinate) changes, otherwise it keeps its value. We can see that the modularity has a single clear peak at a point in the plot, which indicates that the network is divided into four groups.

To analyze the differences among the Girvan and Newman method, the information centrality method and our method, we show experimental results by drawing four scatter plots in Fig. 8. The scatter plots (a) and (b) respectively show the correlation between edge information centrality and our method on 128-node random networks ($Z_{out} = 4$ and $Z_{out} = 7$). The scatter plots (c) and (d) respectively show correlation between edge betweenness centrality and our method on

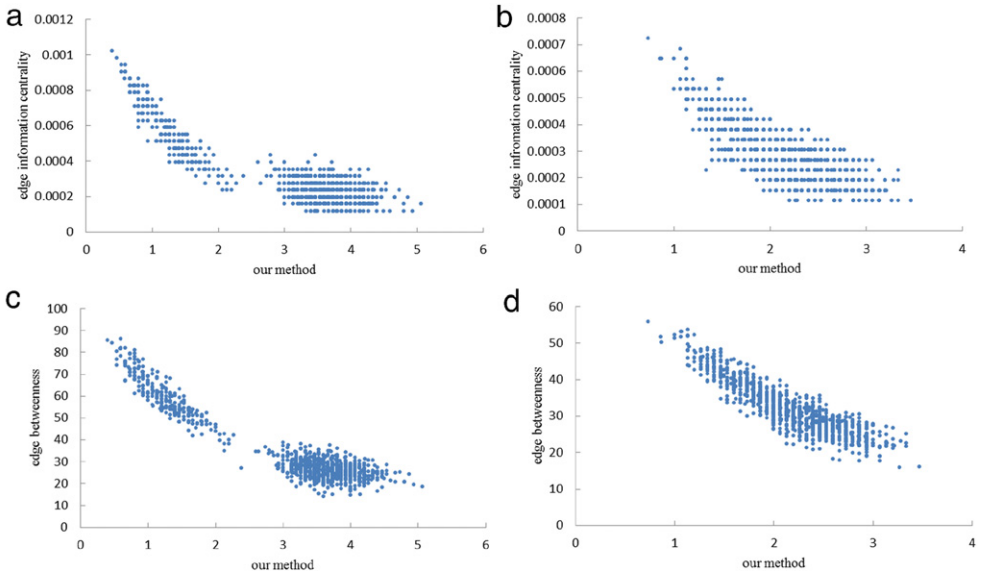


Fig. 8. Correlation among edge betweenness, edge information centrality and our method. Each point in the scatter plots refers to an edge of an artificially generated network with 128 nodes $Z_{out} = 4$ and $Z_{out} = 7$. The two networks respectively show a case in which the communities are distinctly separated and a case in which the communities are hard to detect.

Table 3
The ratio of correct classification on the American football teams network.

Method	The ratio of correct classification	No. of groups
Girvan and Newman method	80.87% (93/115)	9/11
Edge information centrality	79.13% (91/115)	10/11
Our method	**89.57**% (**103**/**115**)	**11**/**11**

128-node random networks ($Z_{out} = 4$ and $Z_{out} = 7$). The results show that the three methods are related, though there are some notable differences. The edges with higher information and betweenness are not always the edges with the lower values processed by our method. For instance, in the case $Z_{out} = 4$, the edge that will be removed by our algorithm is not the one with the largest betweenness. According to the distribution of nodes in Fig. 8, we see that plots (a) and (b) scatter more broadly than plots (c) and (d). It means that the correlation between the Girvan and Newman method and our method is more interrelated than the correlation between the information centrality and our method.

5. Applications to real networks

After applying our algorithm to artificial networks, we try to prove that our algorithm works well for real world networks. We present six networks here, although we analyzed more. The networks are American college football teams [18], dolphins [34–38], co-appearance network of characters in the novel Les Miserables [33], the Western States Power Grid of the United States [39], the karate club network and the primate network. In the last part of this section, we deduce that our method can solve the double peak problem. They have been studied by other researchers. In this case, we can easily understand the similarities and differences between different approaches [28,40–43].

5.1. The network of American college football teams

The network, which represents the schedule of games between American college football teams in a season, is divided into well known "conferences". There are 11 conferences plus a few other teams which do not belong to any conference. The team played more games with the teams in the same conference than with the teams in different conferences. Fig. 9 shows the community structures we have derived with our method. In Table 3, we find that our method can identify 11 groups, which are in accordance with the 11 conferences, while the Girvan and Newman method and the information centrality method respectively find 9 and 10 conferences. The teams which are labeled as Sunbelt are not recognized correctly; this group is misclassified as well in the analysis of the Girvan and Newman method and the information centrality method. The reason is that the Sunbelt teams played basically the same number of games against Western Athletic teams. All of the Sunbelt teams are represented as squares in Fig. 9. The teams which are represented as triangles are wrongly labeled as Sunbelt. However, our method can pick out four correct Sunbelt teams, while the Girvan and Newman method and the information centrality method find less than four correct Sunbelt teams. We compare the ratio of correct classification of our

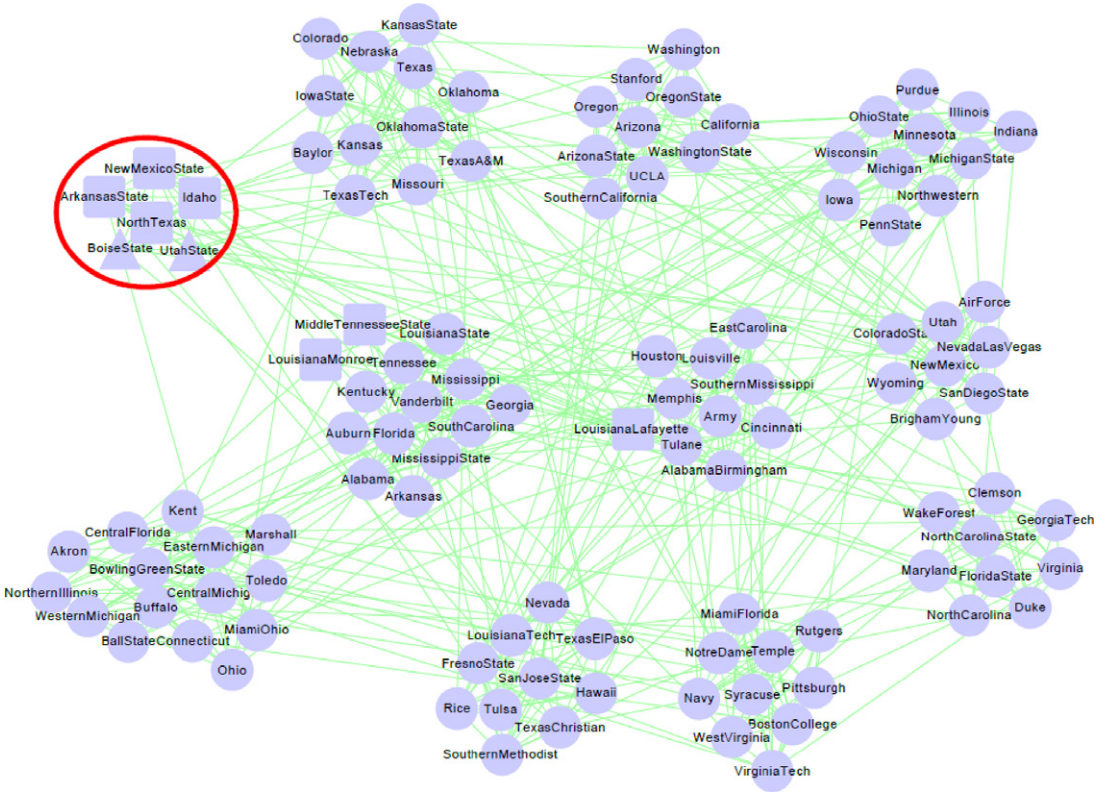


Fig. 9. Community structure found by our algorithm in the college football network. The network is divided into eleven main groups.

method with other methods according to the baseline. In Table 3, we can see that the accuracy of our method is 89.57%, while the accuracy of the other methods is about 80%. The numbers "93/115" in the brackets mean that the method can correctly identify 93 teams in all 115 teams. The numbers "9/11" mean that the method can identify 9 communities, whereas in fact there are 11 groups.

5.2. The network of bottlenose dolphins

We present the social network of 62 bottlenose dolphins in Fig. 10. The dolphins lived in Doubtful Sound, New Zealand. Lusseau [35,36] collected the data of dolphins according to his field studies of dolphins for two years. The ties between dolphin pairs are established by observation of the statistically significant frequent association. In Fig. 10, the network is divided into two groups, which correspond to the known division of the dolphin community. The squares and circles respectively label the members of the two groups in the figure. The *modularity* for the split into two groups is $Q = 0.395996$. When Q reaches 0.511953, the network is divided into smaller subgroups. The larger of the two groups is divided into four subgroups, which are depicted by different colors in Fig. 10.

The split of the network agrees with the fact that the dolphins are separated into two groups along the lines, just because individuals on the boundary between the groups disappeared. The two subgroups of the network joined together when the individuals later appeared again. So Lusseau reported that the developments of dolphin networks, like human social networks, is closely tied to the evolution of the community. The main fact for the dolphins is that the largest subgroup (circles in Fig. 10) consists of females and the others (squares in Fig. 10) almost entirely of males. This situation matches with the split processed by our method. The results of our method are in accordance with the results of other methods.

5.3. The co-appearance network of characters in the novel Les Miserables

In Fig. 11, we depict the interactions between major characters in Victor Hugo's novel Les Miserables [33]. We test our method on the co-appearance network of characters in the novel Les Miserables.

The vertices in this network represent characters and the edges between two vertices denote co-appearance of the characters in one or more scenes. We get a good division for this network, with $Q = 0.484606$, as shown in Fig. 11.

The community structures found by our method clearly reflect the actual structure of characters in the book. Central characters [44], such as Jean Valjean and Javert, are correctly picked out. The communities clearly reflect the subplot

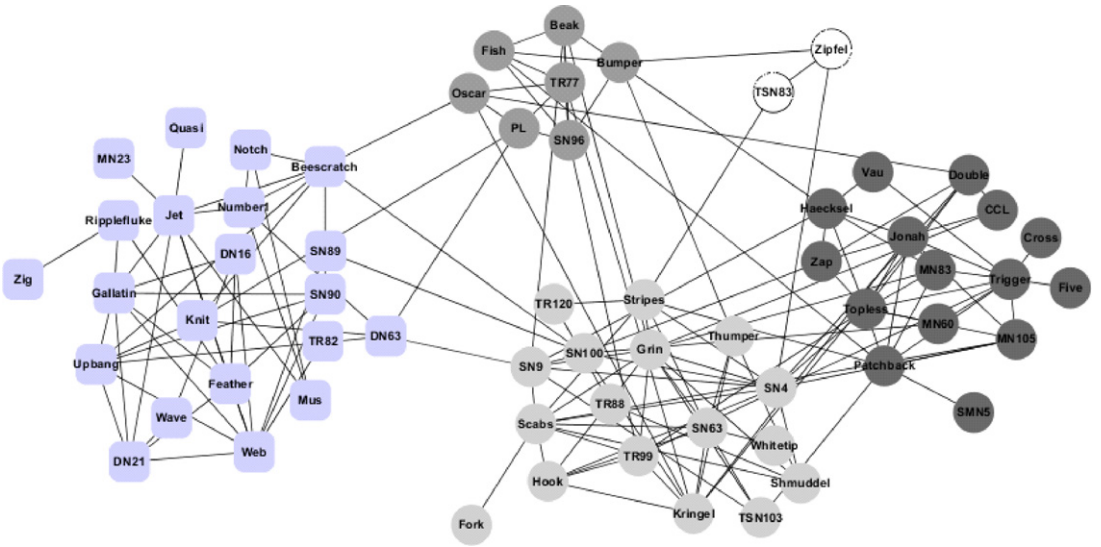


Fig. 10. Community structure found by our algorithm in the bottlenose dolphins of Doubtful Sound [29,31]. The squares and circles present the primary division of the network. Then the circle group is divided into four smaller groups, which are marked by different colors respectively.

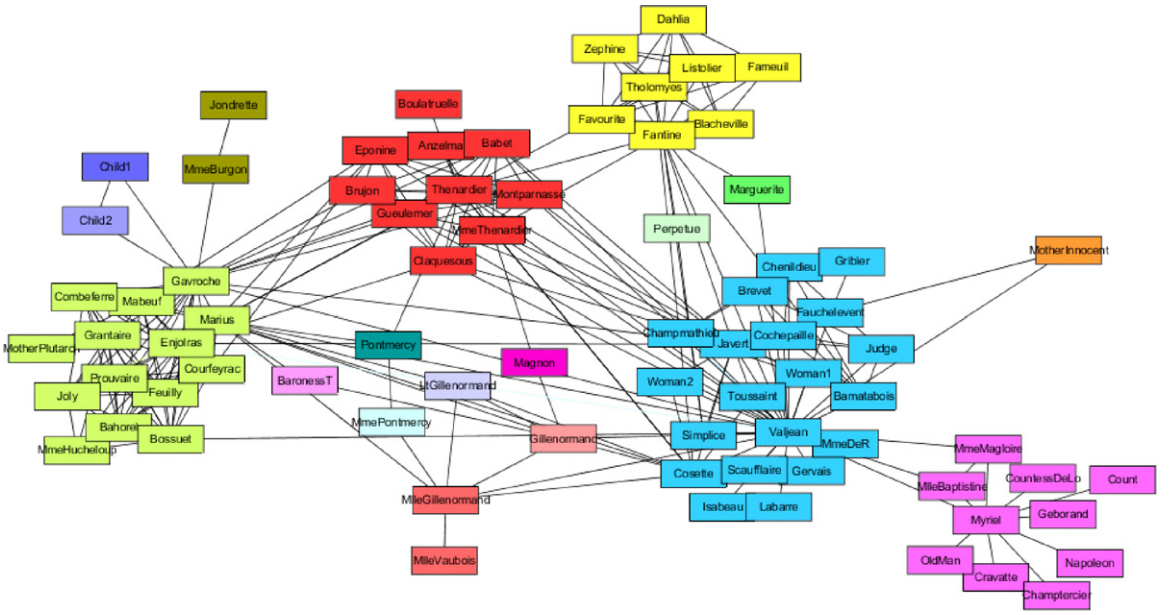


Fig. 11. Community structure detected by our algorithm in the co-appearance network of characters in the novel Les Miserables. The communities are marked by different colors.

structure of the book. Some subplots centered on Marius, Cosette, Fantine and the bishop Myriel are also picked out. The Girvan and Newman method splits many special characters (Perpetue, Marguerite, MmePontmercy, Magnon et al.) into some communities. Different from the Girvan and Newman method, our method can pick out these characters in the network. Our method believes that Perpetue, Marguerite, MmePontmercy, Magnon, Gillenormand, BaronessT and LtGillenormand are the weak ties connecting different communities.

5.4. The Western States Power Grid of the United States

In this part, we apply our method to a network of the Western States Power Grid of the United States [45]. The network includes 4941 nodes and 6594 edges. In order to verify the community structure processed by our method, we depict the community size distributions in Fig. 12. From the figure, we can see that the distributions feature a power law behavior

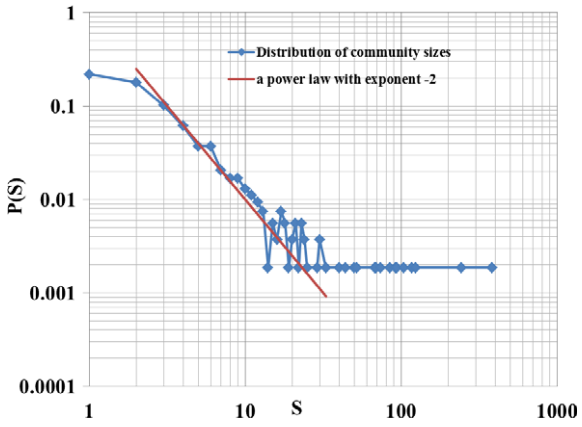


Fig. 12. Distribution of community size of the power grid network. The behavior is well reproduced by a power law with exponent −2.

$P(S) \approx S^{-\tau}$ with $\tau \approx 2$. The result we observed coincides with the principle that the community size distributions are power laws, with exponent $-\tau$ [46].

5.5. Double peak structure examples

In this subsection, we apply three community detection methods on two double peak structure networks, i.e., the karate club network and the primate network. Classification accuracy is used as the evaluation metric in the experiments.

According to the above experimental results and Refs. [1–3], the largest modularity Q always corresponds to a good community structure. The plot of modularity Q sometimes shows a double peak structure [3], as shown in Figs. 13 and 15(b)(c). However, the smaller modularity Q of the GN method and the information centrality method corresponds to a good community structure on some double peak structure networks. Thus, when the plot of modularity Q has two or more peaks, we are not able to figure out which peak corresponds to a good community structure, especially when the network is huge or the community structure is unknown. People who want to detect communities intend to find the right community structure with the largest peak (the largest modularity Q) instead of identifying communities by themselves.

Take the karate club [47] as an example. The karate club analyzed by Zachary consists of 34 persons. If person A is a friend of person B, there is an edge between A and B. There are 78 edges in this club. The friendship relationships are investigated over two years. During two years of follow-up, Zachary et al. found that the club split into two clubs, due to a conflict between an administrator and a coach in the club. The structure of karate club is shown in Fig. 14.

We apply our proposed method, the GN method and the information centrality method on the karate club network. The modularity is used to identify the community structure. Fig. 13 shows the community structure found by our method. As shown in Fig. 13, the Q-plot represents a double peak structure and the first peak is very minor. The double peak structure also exists in the GN and the information centrality method. For the karate club, the smaller peak of GN corresponds to a good community structure. It is the same with the information centrality method, which is also based on "shortest path". Differently, the larger peak of our method corresponds to a better community structure. While the modularity of our method is equal to 0.26, we have four components, two small groups and two large groups. When the modularity reaches its maximum, our method misclassified nodes 3, 4, 8, 13, 20, 25, while the GN method misclassified nodes 3, 5, 6, 7, 10, 11, 17, 25, 28, 29 and the information centrality method misclassified nodes 5, 6, 7, 10, 11, 12, 17, 27, 29. In Table 4, we can see that the accuracy of our method is 82.4%, while the ratios of correct classification for the other methods are lower than 75%. The numbers "28/34" in the brackets mean that the method can correctly identify 28 members in all 34 members. The numbers "4/2" mean that the method can identify four communities, while in fact there are only two communities. The reason why our method works is that almost all edges participate in small order closed walks. In this case, we can evaluate edges by the number of closed walks that they participate in. The edges that participate in a few closed walks are "loose" edges. Our method focuses on deleting the "loose" edges first, while the GN method focuses on the edges with large betweenness. Nevertheless, the edges with large betweenness can participate in many closed walks and have significant correlations with the surrounding nodes. This means that the "loose" edges are more likely than the large betweenness edges to lie between different communities, especially when two communities have overlapping nodes.

Take the primate network [47] as an example. Linda Wolfe [48] collected a data set which recorded three months of interactions amongst a group of 20 monkeys. The interactions were defined as their joint presence at a river. The data set labels the sex and age of each monkey. We further apply the methods on the primate network and use modularity Q to identify the community structures. The plots of modularity of GN and the information centrality method show a double peak structure in Fig. 15(b)(c). Differently, the plot of modularity of our method shows only one peak in Fig. 15(a). The sole peak corresponds to a good community structure. In Table 5 we compare the ratio of correct classification of our method

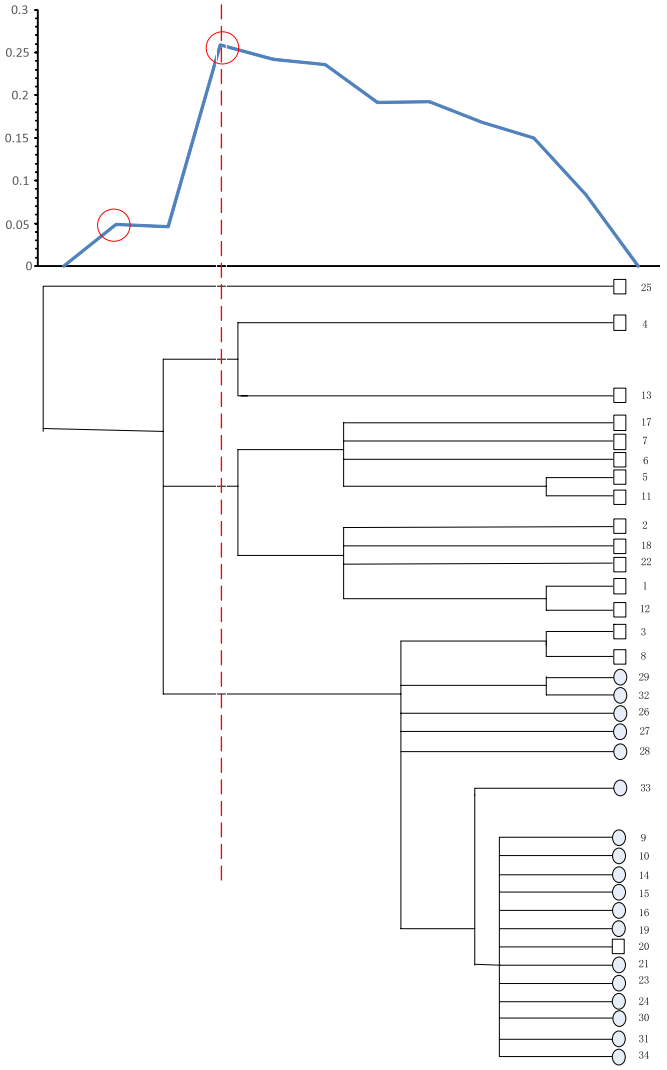


Fig. 13. Community structure in the karate club network.

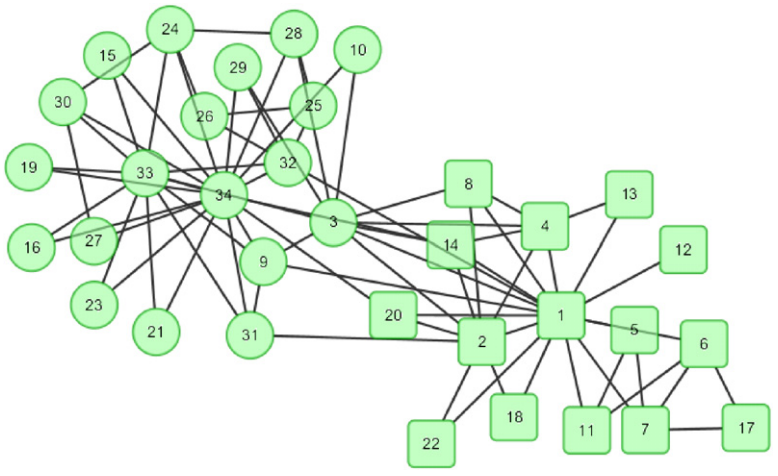


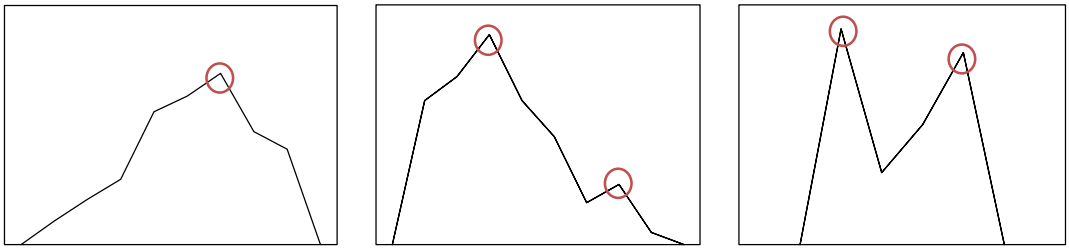
Fig. 14. The karate club network.

Table 4
The ratio of correct classification on the karate club network.

Method	The ratio of correct classification	No. of groups
Our method	**82.4**% (**28**/**34**)	**4**/**2**
Girvan and Newman method	70.6% (24/34)	5/2
Edge information centrality	73.5% (25/34)	7/2

Table 5
The ratio of correct classification on the primate network.

Method	The ratio of correct classification	No. of groups
Our method	**75**% (**15**/**20**)	**7**/**2**
Girvan and Newman method	75% (15/20)	7/2
Edge information centrality	60% (12/20)	11/2



(a) Modularity of our method. (b) Modularity of GN. (c) Modularity of edge information centrality.

Fig. 15. Plots of modularity on the primate network.

with other methods according to the "sex" baseline. The accuracies of our method and the GN method are both 75%, while the accuracy of the information centrality method is 60%.

Actually, it is an advantage for community detection that the largest (only) peak corresponds to a good partition. The reason is as follows: When we detect community structures, we commonly do not know what the network's community structure is. Consequently, if the plot of modularity has two or more peaks, people are not able to figure out which peak corresponds to a better community structure. Our method, which focuses on firstly deleting the "loose" edges, can solve the problem. It is a rather convenient property that the largest modularity corresponds to a good graph partition.

6. Conclusions and future work

This paper presents a new measure based on the concept of closed walks and edge clustering coefficient. The measure evaluates edges by counting the number of closed walks that an edge participates in. Edges with the lowest value always lie between communities. Therefore we remove the edges with the lowest value and recalculate the value of the remaining edges until all edges are removed. We utilize the modularity Q to discover a good division. The time complexity of our method is $O(a\,m + b\,m^2)$, where m is the number of edges, and a and b are constants. Our method runs faster than the information centrality method. Our method also outperforms the GN method when the community structure is hard to detect. We also tested our method on an analog network, computer generated networks and real world networks. The results clearly show that our algorithm performs well.

In this paper, we find that small order closed walks are basic elements in constructing community structures. In addition, information will generally propagate through paths without repeated nodes. In this case, nontrivial closed walks without repeated nodes are more practical and efficient to propagate information than trivial closed walks. The double peak structure is mentioned in this paper. We find that our method is a good way to identify communities on the double peak structure networks.

Actually, closed walks concentrate on local connectivity. It is quite different from the edge betweenness, which is a global metric. Thus, our future work is to improve the complexity of our method. Furthermore, we can analyze the characteristics of complex networks from the perspective of closed walks. For instance, we have found that some robust networks have more closed walks. Consequently, closed walks can provide us with a novel way to improve the robustness of networks. To sum up, analyzing small order closed walks in networks gives us a new insight into the structure of networks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61202175, 61170189, 61202239 and 61370126), the Research Fund for the Doctoral Program of Higher Education (20120203120015, 20111102130003), the Fund of the State Key Laboratory of Software Development Environment (SKLSDE-2013ZX-19).

References

[1] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E. 69 (2004) 026113.
[2] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (2004) 2658–2663.
[3] S. Fortunato, V. Latora, M. Marchiori, Method to find community structures based on information centrality, Phys. Rev. E. 70 (2004) 056104.
[4] M. Faloutsos, P. Faloutsos, C. Faloutsos, On powerlaw relationships of the Internet topology, Comput. Commun. Rev. 29 (1999) 251–262.
[5] R. Albert, H. Jeong, A.-L. Barabási, Internet: diameter of the World-Wide Web, Nature 401 (1999) 130–131.
[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Netw. 33 (2000) 309–320.
[7] A. Kleczkowski, B.T. Grenfell, Mean-field-type equations for spread of epidemics: the 'small world' model, Physica A 274 (1999) 355–360.
[8] C. Moore, M.E.J. Newman, Epidemics and percolation in small-world networks, Phys. Rev. E. 61 (2000) 5678–5682.
[9] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) 3200–3203.
[10] R.M. May, A.L. Lloyd, Infection dynamics on scale-free networks, Phys. Rev. E 64 (2001) 066112.
[11] Roded Sharan, Igor Ulitsky, Ron Shamir, Network-based prediction of protein function, Nat. Mol. Syst. Biol. 3 (2007) 88.
[12] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabási, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.
[13] A. Wagner, D. Fell, The small world inside large metabolic networks, Proc. R. Soc. Lond. Ser. B. 268 (2001) 1803–1810.
[14] J.A. Dunne, R.J. Williams, N.D. Martinez, Food-web structure and network theory: the role of connectance and size, Proc. Natl. Acad. Sci. 99 (2002) 12917–12922.
[15] J. Camacho, R. Guimer'a, L.A.N. Amaral, Robust patterns in food web structure, Phys. Rev. Lett. 88 (2002) 228102.
[16] Lada A. Adamic, Natalie Glance, The political blogosphere and the 2004 US election: divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery, vol. 407, 2005, pp. 360–43.
[17] S. Redner, How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J. B 4 (1998) 131–134.
[18] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. 98 (2001) 404–409.
[19] R.L. Breiger, S.A. Boorman, P. Arabie, An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling, J. Math. Psychol. 12 (1975) 328–383.
[20] H. Zhou, Distance, dissimilarity index, and network community structure, Phys. Rev. E. 67 (2003) 061901.
[21] M. Wu, X. Li, C.-K. Kwoh, S.-K. Ng, A core-attachment based method to detect protein complexes in PPI networks, BMC Bioinform. 10 (2009) 169.
[22] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[23] Y.-Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (2010) 761–764.
[24] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E. 72 (2) (2005) 027104.
[25] R. Guimerà, M. Sales-Pardo, L.A.N. Amaral, Modularity from fluctuations in random graphs and complex networks, Phys. Rev. E. 70 (2) (2004) 025101.
[26] E. Estrada, J.A. Rodríguez-Velázquez, Subgraph centrality in complex networks, Phys. Rev. E. 71 (2005) 056103.
[27] D.J. Jacobs, M.F. Thorpe, Generic rigidity percolation in two dimensions, Phys. Rev. E. 53 (1996) 3682–3693.
[28] E. Estrada, J.A. Rodríguez-Velázquez, Subgraph centrality and clustering in complex hyper-networks, Physica A 364 (2006) 581–594.
[29] B. Karrer, M.E.J. Newman, Random graphs containing arbitrary distributions of subgraphs, Phys. Rev. E. 82 (2010) 066118.
[30] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E. 78 (2008) 046110.
[31] Manoj K. Agarwal, Krithi Ramamritham, Manish Bhide, Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments, PVLDB 5 (10) (2012) 980–991.
[32] Nicholas A. Christakis, James H. Fowler, Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives, first ed., Little, Brown and Company, 2009.
[33] D.E. Knuth, Stanford GraphBase: A Platform for Combinatorial Computing, The Addison-Wesley Professional, 2009.
[34] V. Latora, M. Marchiori, Economic small-world behavior in weighted networks, Eur. Phys. J. B 32 (2003) 249–263.
[35] D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. Lond. Ser. B 270 (2003) S186–S188.
[36] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (2003) 396–405.
[37] Vito Latora, Massimo Marchiori, Efficient behavior of small-world networks, Phys. Rev. Lett. 87 (19) (2001) 198701.
[38] Z.-Z. Zhang, S.-G. Zhou, T. Zou, Self-similarity, small-world, scale-free scaling, disassortativity, and robustness in hierarchical lattices, Eur. Phys. J. B 56 (2007) 259–271.
[39] M.L. Sachtjen, B.A. Carreras, V.E. Lynch, Disturbances in a power transmission system, Phys. Rev. E. 61 (5) (2000) 4877–4882.
[40] M.E.J. Newman, Detecting community structure in networks, Eur. Phys. J. B 38 (2004) 321–330.
[41] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[42] J.M. Pujol, J. Béjar, J. Delgado, Clustering algorithm for determining community structure in large networks, Phys. Rev. E. 74 (2006) 016107.
[43] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (2006) 8577–8582.
[44] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (2012) 1777–1787.
[45] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E. 74 (2006) 036104.
[46] V. Latora, M. Marchiori, A measure of centrality based on network efficiency, New J. Phys. 9 (2007) 188.
[47] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropological Res. 33 (1977) 452–473.
[48] M.G. Everett, S.P. Borgatti, The centrality of groups and classes, J. Math. Sociol. 23 (1999) 181–201.