Physica A 391 (2012) 1788–1796

Contents lists available at SciVerse ScienceDirect

Physica A

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

Efficient algorithm based on neighborhood overlap for community identification in complex networks

Kun Li [a,b,*], Xiaofeng Gong [a,b], Shuguang Guan [c,d], C.-H. Lai [e,b]

[a] Temasek Laboratories, National University of Singapore, Singapore

[b] Beijing-Hong Kong-Singapore Joint Center of Nonlinear and Complex systems (Singapore), National University of Singapore, Kent Ridge, 119260, Singapore

[c] Institute of Theoretical Physics, East China Normal University, Shanghai, 200062, PR China

[d] Department of Physics, East China Normal University, Shanghai, 200062, PR China

[e] Department of Physics, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 14 April 2011
Received in revised form 14 September 2011
Available online 8 October 2011

Keywords:
Complex networks
Community identification
Weak ties

ABSTRACT

Community structure is an important feature in many real-world networks. Many methods and algorithms for identifying communities have been proposed and have attracted great attention in recent years. In this paper, we present a new approach for discovering the community structure in networks. The novelty is that the algorithm uses the strength of the ties for sorting out nodes into communities. More specifically, we use the principle of weak ties hypothesis to determine to what community the node belongs. The advantages of this method are its simplicity, accuracy, and low computational cost. We demonstrate the effectiveness and efficiency of our algorithm both on real-world networks and on benchmark graphs. We also show that the distribution of link strength can give a general view of the basic structure information of graphs.

1. Introduction

The study of complex networks has revealed many interesting characteristics of natural and artificial systems composed of large numbers of interconnected components. For instance, the scale-free degree distributions observed in many real-world networks have significant implications for various dynamical processes on such networks. Another common feature of complex networks is the community structure, or clustering. Communities, also called clusters or modules, are groups of nodes which probably share common properties and/or play similar functional roles within the graph. Communities are observed in many networked systems. For example, communities in a social network might represent real social groupings, perhaps by interest or background [1,2]; communities in a citation network might represent related papers on a specific topic [3]; communities in a metabolic network might represent cycles and other functional groupings [4,5]; communities on the web might represent pages on related topics [6]. Undoubtedly, fast and accurate identification of these communities could help us obtain better understanding and visualization of the structure of networks.

Although there is no unified mathematical definition of a community structure, a community in a network displays common structural properties. Basically, a community is a group of nodes that are more highly connected to each other than to nodes in other groups. Nodes within the group are more topologically similar to each other than to the rest of the network. Due to this characteristic, usually networks with community structure are highly heterogeneous in terms of link strengths. It is shown that complex networks often organize themselves according to the global efficiency principle, meaning

* Correspondence to: Temasek Laboratories, National University of Singapore, 119260, Singapore. Tel.: +65 65167634.
E-mail address: tsllk@nus.edu.sg (K. Li).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.09.027

that the link strengths are optimized to maximize the overall flow in the network [7,8]. In this case the weight of a link should be correlated with its betweenness centrality, which is proportional to the number of shortest paths between all pairs of nodes passing through it [1,7,9]. Another possibility is that the strength of a particular link depends only on the nature of the relationship between two individuals, and is thus independent of the network surrounding the link. Furthermore, the much studied strength of weak ties hypothesis [10–12] states that the strength of a tie between two nodes A and B increases with the overlap of their "friendship circles", resulting in the importance of weak ties in connecting communities. This leads to high betweenness centrality for weak links, which can be seen as the mirror image of the global efficiency principle [13].

Recently, there has been a lot of efforts in defining, detecting, and identifying communities [1–6,14–17,21,22]. The goal of a community identification algorithm is to discover possible natural divisions of networks, by only using the information encoded in the network topology. So far, many algorithms have been proposed which are based on betweenness measures [14], random walk [15], spectral [16], and flow maximization [17], etc. These methods have been successfully implemented to identify communities in many complex networks. However, they all have some disadvantages. For example, some methods require setting the partition number in advance; other algorithms, especially those based on the global properties of a network, are actually not computationally efficient to deal with very large size of real-world networks. In fact, for realistic large complex networks, e.g., networks with millions of nodes, it is desirable that we have a fast algorithm even for a preliminary identification of the possible community structure of the network. Our algorithm is motivated by the observation that two neighboring nodes within a community usually have more common neighbors, i.e., the topological overlap, than two neighboring nodes in different communities do. We find that the relative topological overlap of two neighboring nodes can be defined as the strength of the link connecting them. According to this measure, the nodes with strong links can be naturally grouped together as in a community. By applying this method to several artificial and real networks, we show that our algorithm has the following advantages. First and foremost, it is effective and can provide accurate community identification. Secondly, it is a localized community detection algorithm, and thus can be used to deal with realistically large networks with relatively low computational cost. Thirdly, it does not need *a priori* information on the number and sizes of the communities that are supposed to be determined at the end of the algorithm. In addition, it is very easy to implement.

The rest of this paper is organized as follows. In Section 2, we describe in detail the algorithm and the implementation of our method. In Section 3 we provide a number of examples to test our algorithm. Conclusions are drawn in Section 4.

2. Community detection using topological overlap

Let us start with a simple network G, i.e., undirected and with no loops or multiple edges, on a finite vertex set $V = 1, 2, \ldots, n$ and edge set E, represented by the adjacency matrix $A(G)$. The symmetric $n \times n$ adjacency matrix takes values $A(G)_{ij} = 1$, if there is an edge connecting vertices (i, j) and 0 otherwise.

To identify communities in the network, our method is motivated by the strength of weak ties hypothesis. The strength of weak ties hypothesis states that the strength of a tie between two nodes A and B increases with the overlap of their "friendship" circles, resulting in the importance of weak ties in connecting communities. A general feature of community structure network is that the nodes within a group are much more connected to each other than to the rest of the network. Thus, consistent with the strength of weak ties hypothesis, the majority of the strong links should be found within the clusters, while most links connecting different communities should be significantly weaker than the links within the communities. In order to quantify the strength of a link, we measure the relative topological overlap as follows:

$$Q_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \tag{1}$$

where n_{ij} is the number of common neighbors of i and j, and $k_i(k_j)$ denotes the degree of node $i(j)$.

The topological overlap of two nodes has been found useful in biological networks [18–20]. The definition of TOM can be found in the Supplementary material of [18] as following: $w_{ij} = J_{ij}/\min(k_i, k_j)$, where J_{ij} is same as n_{ij} (plus 1 if there is a direct link between i and j). Here, our relative topological overlap measure (RTOM) reflects the fraction of their common neighbors of the two nodes i and j. Actually, the link strength can be understood as the edge-clustering coefficient, which is an extension of the usual node-clustering coefficient. Here, n_{ij} is the number of triangles containing node i and j. The edge-clustering coefficient is the fraction of the number of triangles the edge belongs to, to all possible triangles include the edge. A basic observation in complex network is that many triangles exist within clusters, while edges connecting in different communities involve few or no triangles. By computing the strength for each link in the network, we complement the adjacency matrix $A(G)$ with the relative friends overlap matrix Q. The adjacent matrix $A(G)$ describes the connections among vertices, while the matrix Q further presents the strength of each connection. If i and j have no common acquaintances we have $Q_{ij} = 0$, and this link represents potential bridges between two different communities. If i and j are part of the same neighborhood of friends, $Q_{ij} = 1$ (Fig. 1). The stronger the link between two nodes, the more their neighbors overlap. As a matter of fact, this quantity gives a better description of the local similarity among the nodes.

We assume that each node in the network chooses to join a community where it has strong average aggregated link strength. Consequently, the average link strength for a node staying in its community is always greater than that going out to a different community. Starting with an arbitrary node, we can extract its associated community by exploiting

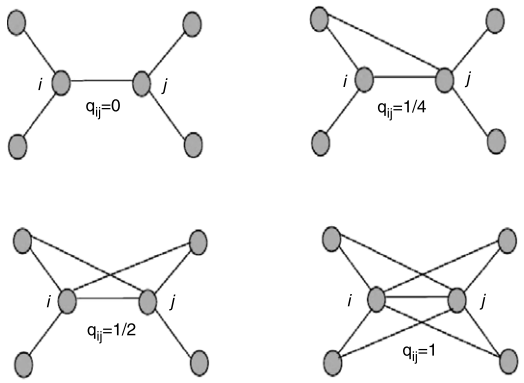

Fig. 1. Illustration of the relative neighborhood overlap between two nodes i and j.

its connectivity with other nodes in the network according to the link strength. Thus, initially the network is tentatively separated into two groups of nodes. A node is assigned based on the relative link strength. After going over all the nodes in the network, we have partitioned the network into two subgroups. By repeating this procedure, we can further divide the subgroups until there is no weak link within each of them. Our algorithm can be summarized in the following steps.

1. For a given network, compute link strength Q_{ij} for all edges.
2. Start with an arbitrary node A and then extract its associated community. Initially, the network is tentatively divided into two groups, where group 1 is with node A and nodes $j(Q_{Aj} > 0)(g1(A, j(Q_{Aj} > 0)))$ and other nodes in group 2 $(g2(others))$.
3. At each step, a random node B is placed in both groups to compare the link strengths

$$w_1 = \left\langle \sum_{j \in g1} Q_{Bj} \right\rangle, \qquad w_2 = \left\langle \sum_{j \in g2} Q_{Bj} \right\rangle. \tag{2}$$

   If $w_1 > w_2$, node B is placed in group 1; if $w_1 < w_2$, then node B is placed in group 2; if $w_1 = w_2$, move node B in a fuzzy node set temporarily.
4. For those nodes with zero strength links to both groups, move them into the fuzzy node set.
5. Iterate the cycle from step 3 until all nodes are covered.
6. Analyze the fuzzy node f. If the degree k_f of node f is 1, f is in the same group as its connected node. If $k_f >= 2, f$ is in the same group as its neighbor with the highest degree. Then, all nodes have been divided into two subgroups, g1 and g2. Empty the fuzzy node set.
7. Repeat from step 2 and further divide g1 and g2 until $\min(Q_{g1,g2}) > q$. Here, q is a control parameter which is defined as $1.5 * \min(Q_{ij})$ in our numerical simulations.

In community networks, the weak links act as bridges that connect different communities, whereas the strong links are predominant within the communities. Thus, when the minimum link strength of a network is larger than a certain value, the network can be regarded as one single community and cannot be further divided. q is a relative value determined by the density of the inter- and intra-communities connections. In our simulation, we found that when $\min(Q_{g1,g2}) > 1.5 * \min(Q_{ij})$, the result is acceptable. For the single link node, $k = 1$, the link strength is zero. It can be naturally grouped with its connecting node. For the bridge node with $k >= 2$ that connects different communities, it is reasonable that such a node is classified in the same group as its neighbor with the highest degree, since the node with a high degree usually has more attraction.

Note that our community detection algorithm is based on the relative link strength. The measure of the link strength is not determined by the global network topology, but by the local network structure in the link's nearest neighborhood. As the algorithm only performs a local exploration of the network, computationally it is very efficient. The complexity of the algorithm consists of link strength measure and iterative node moving. The computation of Q can be done in $O(nm)$ for a sparse network with bounded node degree as is always the case in real world applications. The partition process is of order $O(\log(z)m)$, z is the number of communities. Thus the total complexity of our algorithm is $O(nm)$. Furthermore, the number of communities and their sizes are supposed to be unknown, which will be determined by the algorithm itself automatically.

3. Evaluation community identification

In this section we present a few applications of our algorithm to particular problems, including three real-world networks and benchmark networks.

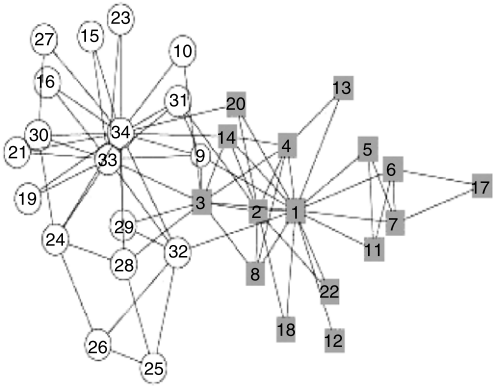

Fig. 2. Zachary's karate club network. Square nodes and circle nodes represents the two split communities [14]. The often misclassified nodes are 3, 9 and 10.

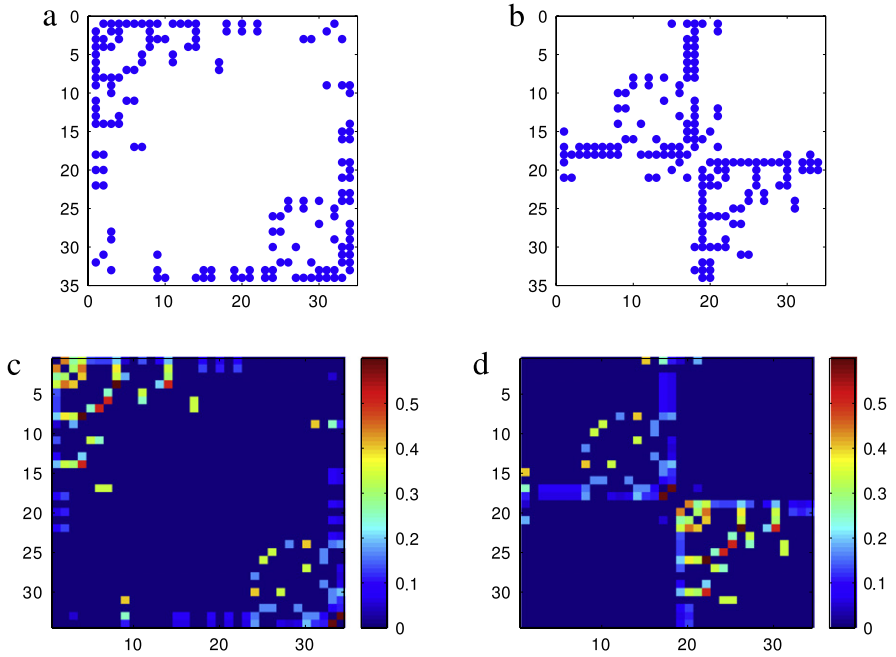

Fig. 3. Illustration of the Karate club connections and link strengths. The adjacency matrix of (a) the Karate club network, (b) the output communities. The link strength distribution of (c) the observed network and (d) the output communities.

3.1. Zachary's karate club network

The first one is Zachary's Karate club network [23] which has become a benchmark for all methods of community detection. This is a network of friendship among 34 members of a karate club as nodes and 78 edges representing friendship between members. Due to a leadership issue, the club split into two distinct groups. Our analysis of Zachary's network shows in Figs. 2 and 3. The two communities are identified exactly as the observed splitting of the network. The majority of the strong links are found within each clusters and the links connecting the two clusters are visible weaker than the links within the cluster. Moreover, we can clearly detect several vertices lying between the two main groups, like 3, 9 and 10; such vertices are often misclassified by other community detection methods. In Table 1, the corresponding link strengths of these nodes are listed. We can see that nodes 3 and 9 can be unambiguously clustered together with their strongly connected nodes by our algorithm. For node 10, there is no common friend among his connections. It is reasonable to place it in the same community as node 34, since larger degree nodes normally have more attraction.

3.2. American college football teams network

The second network we investigated is the U.S. college football network that consists of 115 college teams represented as nodes and has 613 edges between teams that present games played between two teams during the regular season in the

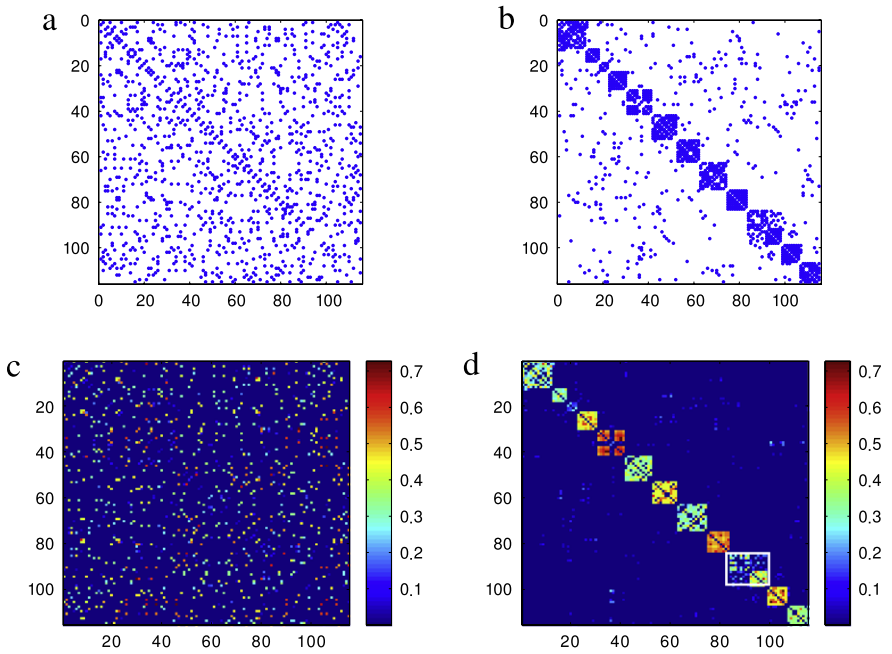

Fig. 4. Illustration of American college football teams network connections and link strengths. The adjacency matrix of (a) the football teams network, (b) the output communities. The link strength distribution of (c) the observed network and (d) the output communities. The white square in the right down corner of (d) is team Mid American which is divided into two groups in our algorithm.

Table 1
Analysis of some often misclassified nodes in Zachary's Karate club network.

Node 3 (in the same group of node 4)										
j	1	2	4	8	9	10	14	28	29	33
Q_{3j}	0.2632	0.3077	0.4000	0.3333	0.1818	0	0.3000	0	0	0.0526
Node 9 (in the same group of node 31)										
j	1	3	31	33	34					
Q_{9j}	0.0556	0.1818	0.4000	0.2500	0.1111					
Node 10 (in the same group of node 34)										
j	3 ($k = 10$)	34 ($k = 17$)								
Q_{10j}	0	0								

year 2000 [1]. The teams are divided into eleven conferences (communities) and each team plays more games within its own conference than interconference games. We measure the link strengths of this network as shown in Fig. 4 and identify thirteen groups. The conference labeled as Mid American is further divided into two groups by our algorithm. This result is well supported by the link strengths between this conference as highlighted in Fig. 4(d). The team labeled as Sunbelt is broken into two groups and grouped with members of the West Athletic Conference. This is coincident with the analysis of Girvan and Newman and the reason is explained in [1].

3.3. Dolphin social network

The Dolphin social network is a community of bottlenose dolphins living off Doubtful Sound, New Zealand [24]. There are 62 dolphins and edges are set between network members that are seen together more often than expected by chance. The network splits naturally into two large groups and the larger one also splits into three smaller subgroups [2]. Our analysis shows that this is indeed the case. The results are shown in Figs. 5 and 6. Only one node, SN89, is misclassified by our method. SN89 has no friends circle. The edges connected with SN89 run as a bridge between the two large groups. As a fuzzy node in our method, it is grouped with the higher degree node it connects.

3.4. Tests on LFR benchmark graphs

We now apply our method on LFR benchmark graphs [25]. The benchmark networks are built with the input parameters: network size is 1000, the maximum degree is 50, the exponent of the degree distribution is -2, and that of the community

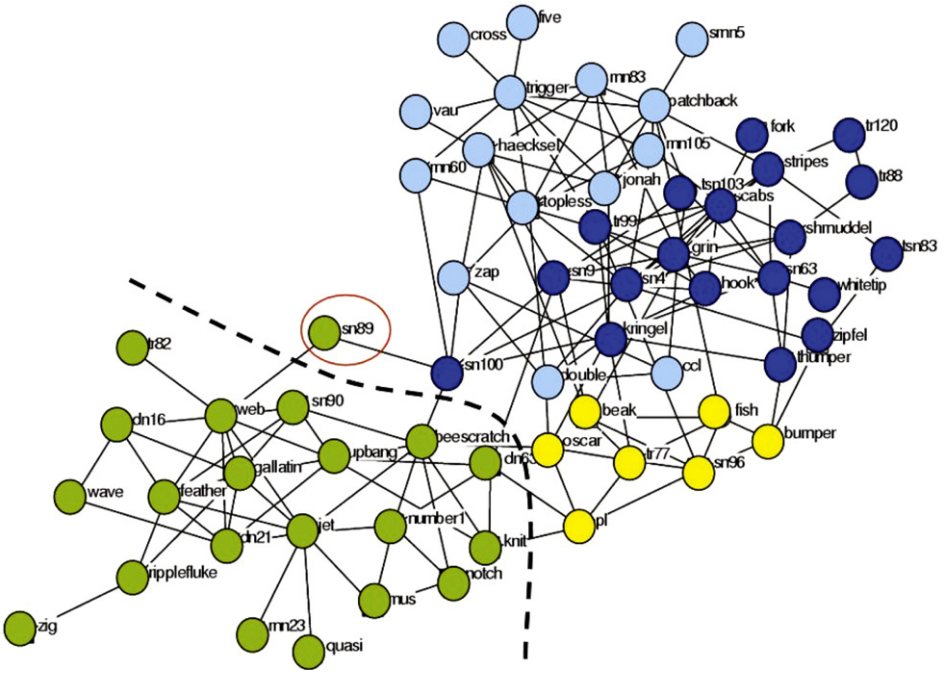

Fig. 5. Dolphin social network. The colors correspond to the partitions given by our method. The dash line denotes the primary split of the network into two groups.

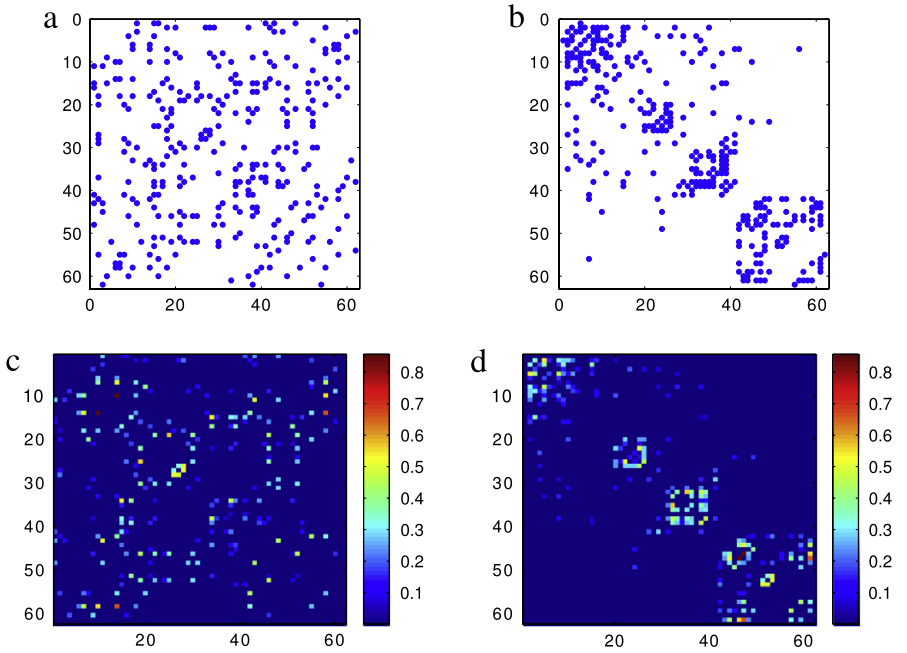

Fig. 6. Illustration of dolphin network connections and link strengths. The adjacency matrix of (a) the dolphin network, (b) the output communities. The link strength distribution of (c) the observed network and (d) the output communities.

size distribution is -1, the communities have between 20 to 50 nodes. To quantify our algorithm, we use the normalized mutual information measure proposed in [26]. In Fig. 7 we plot the performance of our algorithm as a function of the mixing parameter μ. The compare result with the infomap by Rosvall and Bergstrom [27], which is a very efficient method [26], is shown in Fig. 8. The general shape of the mutual information curve is similar to that of the best performance methods, although the mutual information values are somehow lower for low values of the mixing parameters. We examine the communities in detail and find that our method splits or merges communities so the community structure is not exactly correct.

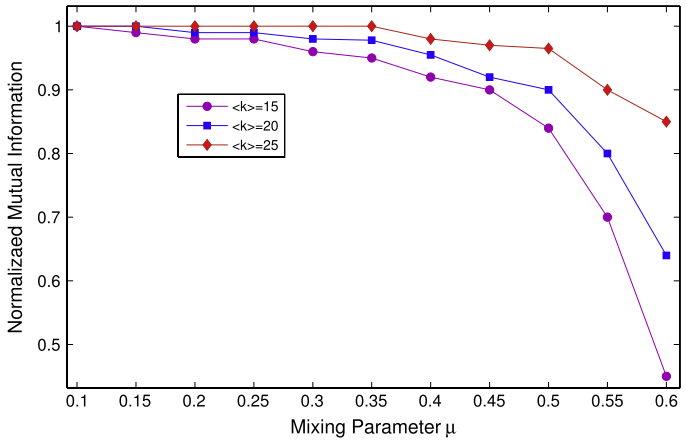

Fig. 7. Tests of the algorithm on the LFR benchmark. The value of normalized mutual information is the average value over 100 realization for each value of the mixing parameter.

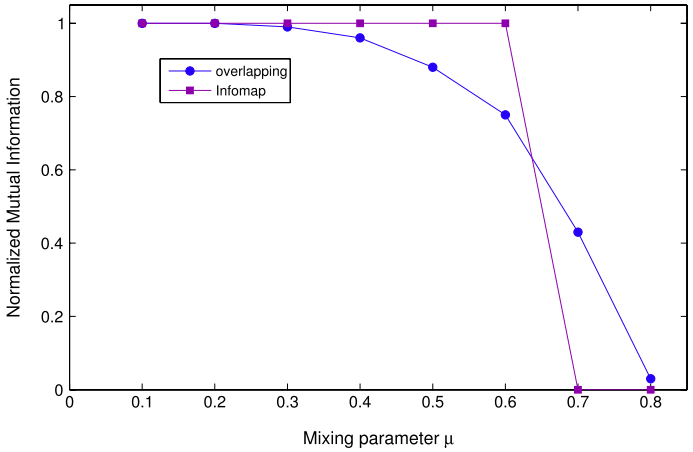

Fig. 8. Tests of the algorithms on LFR benchmark. Parameters used were the same as in Ref.[26].

As we have described it, the relative friends overlap matrix Q not only describes the connections among nodes, but also presents the strength of each connection. The more their neighbors overlap, the stronger the link between two nodes. In strong community graphs, the value of average link strength is large and these values are within a wide range because most of the links stay in community (small μ). The average value becomes small and the range becomes narrow when more links are attached to different community (large μ). In Fig. 9, we show how the overlap distribution scales with the mixing parameter μ. We have known that the performance of most algorithms is worse after the threshold $\mu_c = 0.5$, the border beyond which communities are no longer defined in the strong sense [26]. Our explanation is: it is the narrow link strength distribution that makes the difference between the average link strength for a node staying in its community and going out to a different community is small and thus hard to detect.

In random graphs, the linking probabilities of the nodes are independent of each other. In this way there will be inhomogeneity in the density of the links on the graphs. Consequently, there is less friends overlap and the overall link strength is weak. The value of Q in random network is small and does not spread widely. Now we study the link strength distribution of random networks. The networks are scale-free random graphs, network size is 1000, the maximum degree is 100, the exponent of the degree distribution is -2. In Fig. 10 we show the link strength distribution as a function of the average degree of the graph. The small average value and narrow scope is consistent with the feature of random networks.

From all of the above, it is clear that for a given network, we should analysis the link strength distribution before start to implement community identification algorithm. The relative friends overlap matrix Q gives the basic structure information of graphs, a strong community network or a random like graph.

4. Conclusion and discussion

We have presented a new algorithm for identifying community structures in various networks. Unlike previous methods that focus on the global network properties such as edge betweenness, our approach uses the local information of topological

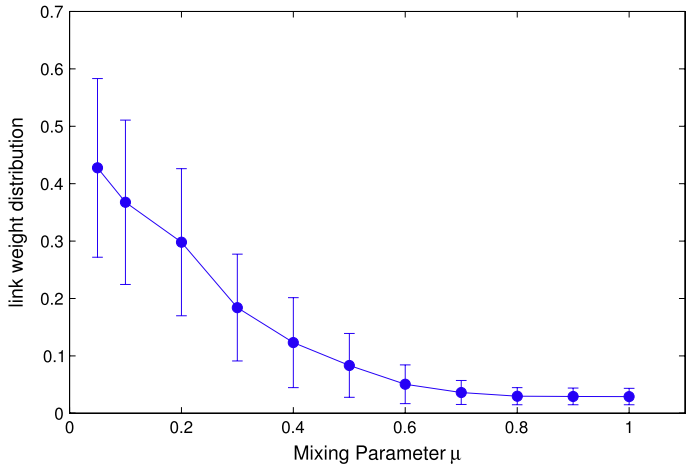

Fig. 9. The distribution of link strength Q scales with the mixing parameter μ. Parameters used are the same as performance tests.

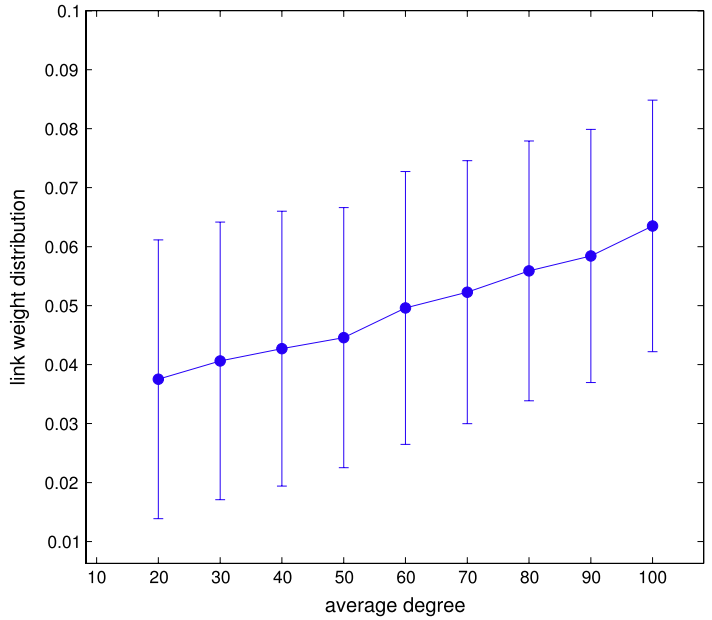

Fig. 10. The distribution of link strength Q scales with the average degree of random networks.

overlap of neighboring nodes to detect communities. Nodes can be grouped into communities according to the strength of links which depends only on the topological relationship between two connecting nodes. Nodes are assigned the same community where they have stronger link strengths.

Our method requires no prior knowledge on the community structure. Since it only performs a nearest neighborhood exploration of the network, computationally it is very efficient, and make it possible to analyze large size networks with relatively low computational cost. The application of our method to a number of real networks and benchmark graphs has verified its effectiveness and accuracy for identifying communities in complex networks. On the other hand, the analysis of link strength matrix Q can give a general view of the basic structure information of graphs before implement any community identification algorithm.

We would like to emphasize that our method can be naturally extended to find overlapping communities. Overlapping nodes in a network can belong to more than one community. From the topological structure point of view, overlapping nodes have equivalent topological effect on the communities they may belong to. In our method, overlapping nodes are part of those fuzzy nodes which have the same link strength with the communities of which they could be a member. Here, we place them in the same group as the highest degree node they connect to.

Finally, we remark that the community structure in a network could be refined by our method. This is because that our stopping criteria q is a relative value that depends on the densities of inter- and intra-communities connections. For a specific

q value, a community may be further divided into smaller clusters. This is the case shown in the examples of the football team network.

References

[1] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA 99 (2002) 7821.
[2] D. Lusseau, M.E.J. Newman, Proc. R. Soc. Lond. B(Suppl.) 271 (2004) S477–S481.
[3] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 98 (2001) 404–409.
[4] R. Guimera, L.A.N. Amaral, Nature 895 (2005) 433.
[5] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Nature 814 (2005) 435.
[6] G.W. Flake, S. Lawrence, C. Lee Giles, F.M. Coetzee, IEEE Computer 35 (3) (2002) 66.
[7] K-I. Goh, B Kahng, D. Kim, Phys. Rev. Lett. 87 (2001) 278701.
[8] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, JR. Banavar, Science 272 (1996) 984–986.
[9] LC. Freeman, Sociometry 40 (1977) 35–41.
[10] M. Granovetter, Am. J. Sociol. 78 (1973) 1360–1380.
[11] M. Granovetter, Getting a Job: A Study of Contacts and Careers, 2nd ed., University of Chicago Press, Chicago, 1995.
[12] P. Csermely, Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks, 1st ed., Springer, Berlin, 2006.
[13] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabasi, Proc. Natl. Acad. Sci. USA 104 (2007) 7332–7336.
[14] M.E.J. Newman, M. Girvan, Phys. Rev. E 69 (2004) 026113.
[15] H. Zhou, Phys. Rev. E 67 (2003) 041908–061901.
[16] M.E.J. Newman, Phys. Rev. E 74 (2006) 036104.
[17] F. Wu, B.A. Huberman, Eur. Phys. J. B 38 (2004) 331.
[18] E. Ravasz, A.L. Somera, D.A. Mongru, Z. Oltvai, A.-L Barabsi, Science 297 (2002) 1551–1555.
[19] B. Zhang, S. Horvath, Stat. Appl. Gen. Mol. Biol 4 (2005) 17.
[20] A.M. Yip, S. Horvath, BMC Bioinformatics 8 (2007) 22.
[21] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl. Acad. Sci. USA 101 (2004) 2658–2663.
[22] Santo Fortunato, Phys. Rep. 486 (2010) 75–174.
[23] W.W. Zachary, J. Anthropol. Res. 33 (1997) 452.
[24] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, Behav. Ecology Sociobiology 54 (2003) 396.
[25] Andrea Lancichinetti, Santo Fortunato, Filippo Radicchi, Phys. Rev. E 78 (2008) 046110.
[26] Andrea Lancichinetti, Santo Fortunato, Phys. Rev. E 80 (2009) 056117.
[27] M. Rosvall, C.T. Bergstrom, Proc. Natl. Acad. Sci. USA 105 (2008) 1118.