Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Quantifying and identifying the overlapping community structure in networks

Hua-Wei Shen, Xue-Qi Cheng[1] and Jia-Feng Guo

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, People's Republic of China
E-mail: shenhuawei@software.ict.ac.cn, cxq@ict.ac.cn and guojiafeng@software.ict.ac.cn

Received 24 May 2009
Accepted 9 July 2009
Published 27 July 2009

Online at stacks.iop.org/JSTAT/2009/P07042
doi:10.1088/1742-5468/2009/07/P07042

Abstract. It has been shown that the communities of complex networks often overlap with each other. However, there is no effective method to quantify the overlapping community structure. In this paper, we propose a metric to address this problem. Instead of assuming that one node can only belong to one community, our metric assumes that a maximal clique only belongs to one community. In this way, the overlaps between communities are allowed. To identify the overlapping community structure, we construct a maximal clique network from the original network, and prove that the optimization of our metric on the original network is equivalent to the optimization of Newman's modularity on the maximal clique network. Thus the overlapping community structure can be identified through partitioning the maximal clique network using any modularity optimization method. The effectiveness of our metric is demonstrated by extensive tests on both artificial networks and real world networks with a known community structure. The application to the word association network also reproduces excellent results.

Keywords: network dynamics

[1] Author to whom any correspondence should be addressed.

1742-5468/09/P07042+16$30.00

Contents

1. Introduction

Many complex systems in nature and society can be described in terms of networks or graphs. The study of networks is crucial to understanding both the structure and the function of these complex systems [1, 2]. A common feature of complex networks is community structure, i.e., the existence of groups of nodes such that nodes within a group are much more connected to each other than to the rest of the network. Communities reflect the locality of the topological relationships between the elements of the target systems [3], and may shed light on the relation between the structure and the function of complex networks. Take the World Wide Web as an example, closely hyperlinked web pages form a community and they often talk about related topics [4].

The identification of community structure has attracted much attention from various scientific fields. Many methods have been proposed and applied successfully to some specific complex networks [5]–[14]. In order to quantify the community structure of networks, Newman and Girvan [6] proposed the modularity as a measure of a partition of network, in which each node only belongs to one community. The proposal of modularity has prompted the detection of community structure. However, the modularity faces several problems. For example, the modularity suffers a resolution limit problem [15, 16]. Furthermore, the modularity-based methods cannot tackle overlapping community structure, in which one node may belong to more than one community. Figure 1 shows an example network with overlapping community structure. Intuitively, overlapping community structure can be represented by a cover of the network. A cover of the network is defined as a set of clusters such that each node is assigned to one or more clusters and no cluster is a proper subset of any other cluster. As to the network in figure 1, the overlapping community structure can be represented by the cover $\{\{1, 2, 3, 4, 5, 6\}, \{3, 7, 8, 9, 10, 11, 12, 13\}, \{10, 11, 12, 14, 15, 16, 17\}, \{18, 19, 20, 21, 22, 23, 24\}\}$.

doi:10.1088/1742-5468/2009/07/P07042

J. Stat. Mech. (2009) P07042

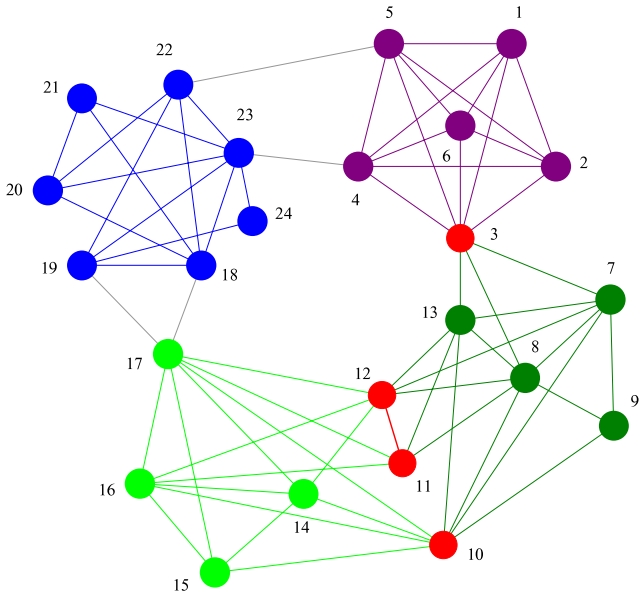

Figure 1. A schematic network with overlapping community structure. Communities are differentiated by colors and the overlapping regions are emphasized in red. The edges between communities are colored in gray.

Overlapping community structure has been widely studied [17]–[25]. In [17], the community structure is uncovered by k-clique percolation and the overlaps between communities are guaranteed by the fact that one node can participate in more than one clique. However, the k-clique method gives rise to an incomplete cover of the network, i.e., some nodes may not belong to any community. In addition, the hierarchical structure cannot be revealed for a given k. In [24], by introducing the concept of the belonging coefficients of each node to its communities, the authors proposed a general framework for extending the traditional modularity to quantify overlapping community structure. The method provides a new idea to find overlapping community structure. However, the physical meaning of the belonging coefficient lacks a clear explanation. Furthermore, the framework is hard to extend to large scale networks since it is difficult to find an efficient algorithm to search the huge solution space. Recently, Evans *et al* [25] proposed a method to identify the overlapping community structure by partitioning a line graph constructed from the original network. This method only allows the communities to overlap at nodes.

In this paper, a measure for the quality of a cover is proposed to quantify the overlapping community structure referred to as Q_c (quality of a cover). With the measure Q_c, the overlapping community structure can be identified by finding an optimal cover, i.e., the one with the maximum Q_c. The Q_c is based on a maximal clique view of the original network. A maximal clique is a clique (i.e. a complete subgraph) which is not a subset of any other clique in a graph. The maximal clique view is according to a reasonable assumption that a maximal clique cannot be shared by two communities due to it being highly connective. To find an optimal cover, we construct a maximal clique network from the original network. We then prove that the optimization of Q_c on the original network is equivalent to the optimization of the modularity on the maximal clique network. Thus the overlapping community structure can be identified through partitioning the maximal

J. Stat. Mech. (2009) P07042

clique network with an efficient modularity optimization algorithm, e.g., the fast unfolding algorithm in [14]. The effectiveness of the measure Q_c is demonstrated by extensive tests on both artificial networks and real world networks, with a known community structure, and the application to the word association network.

2. The quantifying and identifying methods

In this section, we first propose a measure Q_c to quantify the overlapping community structure of networks. Then the overlapping community structure of a network is identified by partitioning a maximal clique network constructed from the original network using a modularity optimization algorithm. Finally, some discussions about our method are given.

2.1. Quantifying the overlapping community structure

As mentioned above, the overlapping community structure can be represented as a cover of the network instead of a partition of the network. Therefore, the overlapping community structure can be quantified through a measure of a cover of the network.

As is well known, the modularity was used to measure the goodness of a partition of the network. Given an un-weighted, undirected network $G(E, V)$ and a partition P of the network G, the modularity can be formalized as

$$Q = \frac{1}{L}\sum_{c\in P}\sum_{vw}\delta_{vc}\delta_{wc}\left(A_{vw} - \frac{k_v k_w}{L}\right), \tag{1}$$

where A is the adjacency matrix of the network G, $L = \sum_{vw} A_{vw}$ is the total weight of all the edges, and $k_v = \sum_w A_{vw}$ is the degree of the vertex v.

In equation (1), δ_{vc} denotes whether the vertex v belongs to the community c. The value of δ_{vc} is 1 when the vertex v belongs to the community c and 0 otherwise. For a cover of the network, however, a vertex may belong to more than one community. Thus δ_{vc} needs to be extended to a belonging coefficient α_{vc}, which reflects how much the vertex v belongs to the community c.

With the belonging coefficient α_{vc}, the goodness of a cover C can be measured by

$$Q_c = \frac{1}{L}\sum_{c\in C}\sum_{vw}\alpha_{vc}\alpha_{wc}\left(A_{vw} - \frac{k_v k_w}{L}\right). \tag{2}$$

The idea of the belonging coefficient was proposed in [24]. Its authors also pointed out that the belonging coefficient should satisfy a normalization property. This property is formally written as

$$0 \le \alpha_{vc} \le 1, \qquad \forall v \in V, \quad \forall c \in C \tag{3}$$

and

$$\sum_{c\in C}\alpha_{vc} = 1. \tag{4}$$

Equations (3) and (4) only give the general constraints on α_{vc}, which leads to such a huge solution space that the enumeration of all the solutions is impractical. To reduce the solution space and make the problem tractable, we introduce an additivity property

for the belonging coefficient: the belonging coefficient of a vertex to a community c is the sum of the belonging coefficients of the vertex to all of c's sub-communities.

For example, we assume that $C = \{c_1, c_2, \ldots, c_{r-1}, c_r, \ldots, c_s, c_{s+1}, \ldots, c_n\}$ is a cover of the network G and $C' = \{c_1, c_2, \ldots, c_{r-1}, c_u, c_{s+1}, \ldots, c_n\}$ is another cover of G. The difference between C' and C is that the community c_u is the union of the communities $c_r, \ldots, c_s$. The additivity property of the belonging coefficient can then be formally denoted as

$$\alpha_{vc_u} = \sum_{i=r}^{s} \alpha_{vc_i}. \tag{5}$$

The belonging coefficient α_{vc} reflects how much a vertex v belongs to a community c. Intuitively, it is proportional to the total weight of the edges connecting the vertex v to the vertices in the community c, i.e.,

$$\alpha_{vc} \propto \sum_{w \in V(c)} A_{vw}, \tag{6}$$

where $V(c)$ denotes the set of vertices belonging to community c. Note that the additivity property of the belonging coefficient requires that communities are disjoint from a proper view of the network. Therefore, we introduce the maximal clique view to achieve this purpose. We define α_{vc} as the form

$$\alpha_{vc} = \frac{1}{\alpha_v} \sum_{w \in V(c)} \frac{O^c_{vw}}{O_{vw}} A_{vw}, \tag{7}$$

where O_{vw} denotes the number of maximal cliques containing the edge (v, w) in the whole network, O^c_{vw} denotes the number of maximal cliques containing the edge (v, w) in the community c, and α_v is a normalization term denoted as

$$\alpha_v = \sum_{c \in C} \sum_{w \in V(c)} \frac{O^c_{vw}}{O_{vw}} A_{vw}. \tag{8}$$

Obviously, the definition in equation (7) satisfies the normalization property. It also satisfies the additivity property if we assume that each maximal clique only belongs to one community. This assumption is reasonable since a maximal clique is highly connective such that any two communities sharing a maximal clique should be combined into a single one.

With equations (2) and (7), we obtain the detailed form of Q_c as a measure to the quality of a cover of the network. Note that when a cover degrades to a partition, Q_c becomes the modularity Q in [8] accordingly. In addition, $Q_c = 0$ when all vertices belong to the same community, and it will be shown later in section 3 that a high value of Q_c indicates a significant overlapping community structure.

2.2. Identifying the overlapping community structure

With the measure Q_c, the overlapping community structure of network can be identified by finding the optimal cover with maximum Q_c. To find the optimal cover, we construct a maximal clique network from the original network. Then the overlapping community structure can be identified through partitioning the maximal clique network.

J. Stat. Mech. (2009) P07042

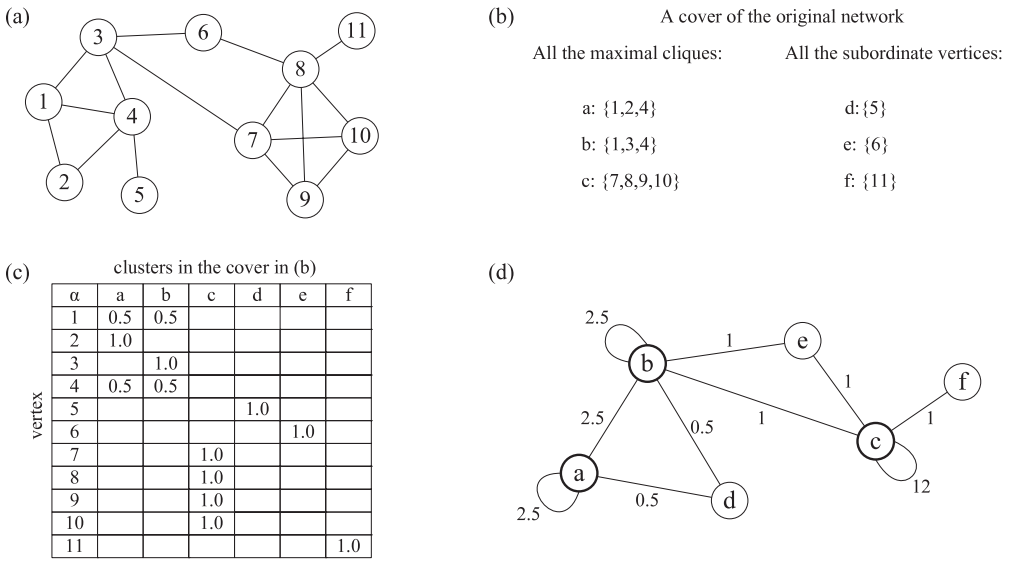

Figure 2. Illustration for the construction process of the maximal clique network. (a) The original example network. (b) A cover of the original network. In this cover, each maximal clique is a cluster and each subordinate vertex forms a cluster consisting of only one vertex. (c) The belonging coefficient of each vertex to its corresponding clusters in the cover. (d) The maximal clique network constructed from the example network. Here the parameter $k = 3$.

2.2.1. Construction of the maximal clique network. Given an un-weighted, undirected network G, a corresponding maximal clique network G' can be constructed through the following method.

The maximal clique network G' is constructed by defining its nodes and edges. We first find out all the maximal cliques in G. We can simply take all these maximal cliques as nodes of G'. In practice, however, we observe that some maximal cliques would not be so highly connective if their sizes are too small. Such a maximal clique either lies between different communities (e.g., the maximal cliques $\{4, 23\}$ and $\{5, 22\}$ in the network shown in figure 1) or connects a node to the whole network (e.g., the maximal clique $\{8, 11\}$ in the network shown in figure 2(a)). To deal with these small maximal cliques, we introduce a threshold k. Specifically, given the parameter k, we only refer to those maximal cliques with the size no smaller than k as the maximal cliques, and refer to those with the size smaller than k as subordinate maximal cliques. We then denote the vertices only belonging to subordinate maximal cliques as subordinate vertices. In this way, each maximal clique or subordinate vertex in the original network G is taken as one node of G'.

Note that all the subordinate vertices and the maximal cliques form a cover C of the original network G. For a subordinate vertex v and a cluster c in the cover C, the value of α_{vc} is defined to be 1.0 when v belongs to the cluster c and 0.0 otherwise. As to other vertices, α_{vc} can be obtained according to equation (7).

Now we can define the edge of the maximal clique network G' by defining its adjacency matrix B. Let m_x denote the set of the original network's vertices corresponding to the

xth node in G'. The element of B is defined as

$$B_{xy} = \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} A_{vw} \tag{9}$$

and the strength (degree) of the xth node

$$s_x = \sum_y B_{xy} = \sum_v \alpha_{vm_x} k_v. \tag{10}$$

For clarity, figure 2 illustrates the construction process of the maximal clique network from an example network with the parameter $k = 3$. Figure 2(b) shows the subordinate vertices and the maximal cliques. Each of them becomes a node in the resulting maximal clique network. For example, the maximal clique $\{1, 2, 4\}$ corresponds to the node a and the subordinate vertex $\{5\}$ corresponds to the node d. Each of these maximal cliques or subordinate vertices is a cluster in a cover C of the original network. Their belonging coefficients corresponding to the cover C are shown in figure 2(c). According to these belonging coefficients and equation (9), the weight of each edge of the maximal clique network is obtained. Take the edge connecting the nodes a and b as an example. As known, the node a corresponds to the maximal clique $\{1, 2, 4\}$ and the node b corresponds to the maximal clique $\{1, 3, 4\}$. Using the equation (9), the weight of this edge is $\alpha_{1a}\alpha_{3b}+\alpha_{1a}\alpha_{4b}+\alpha_{2a}\alpha_{1b}+\alpha_{2a}\alpha_{4b}+\alpha_{4a}\alpha_{1b}+\alpha_{4a}\alpha_{3b} = 0.5+0.25+0.5+0.5+0.25+0.5 = 2.5$.

The constructed maximal clique network is a weighted network though the original network is un-weighted. The total weight L' of all the edges in the maximal clique network is equal to the total weight (number) L of edges in the original network. The proof is

$$\begin{aligned} L' &= \sum_{xy} B_{xy} \\ &= \sum_{xy} \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} A_{vw} \\ &= \sum_{vw} A_{vw} \sum_x \alpha_{vm_x} \sum_y \alpha_{wm_y} \\ &= \sum_{vw} A_{vw} \\ &= L. \end{aligned} \tag{11}$$

Each vertex in the original network corresponds to more than one node in the maximal clique network. For example, in figure 2, the vertex 1 corresponds to two nodes a and b in the maximal clique network. Thus, a partition of the maximal clique network can be mapped to a cover of the original network, which holds the information about the overlapping community structure of the original network.

2.2.2. Finding the overlapping community structure. Now we investigate the overlapping community structure of the original network through partitioning its corresponding maximal clique network. To find the natural partition of a network, the optimization of modularity is the widely used technique. The partition with the maximum modularity is regarded as the optimal partition of network. We employ the algorithm proposed in [14] to partition our maximal clique network. As an example, figure 3 shows the partition of

J. Stat. Mech. (2009) P07042

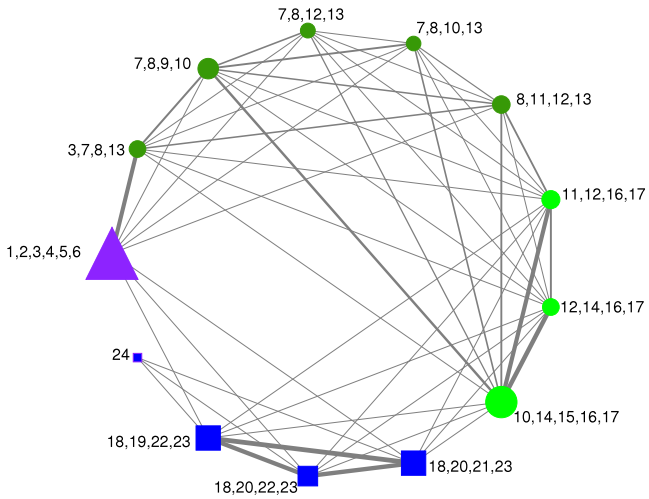

Figure 3. The maximal clique network constructed from the schematic network in figure 1. The label near each node shows its corresponding vertices in the original network. The width of line indicates the weight of the corresponding edge. The self-loop edge of each node is omitted and its width is reflected by the volume of the associated circles, squares or triangles. In addition, the optimal partition of the maximal clique network is also depicted. The communities in this partition are differentiated by shapes. Furthermore, the circle-coded community can be partitioned into two sub-communities. The four communities are shown in different colors, which are identical to the communities depicted in figure 1. Here k is 4.

a maximal clique network. Different parts of the partition are differentiated by shapes or colors.

As mentioned above, each partition of the maximal clique network corresponds to a cover of the original network and the cover tells us the overlapping community structure. The key problem lies in that whether the optimal partition of the maximal clique network corresponds to the optimal cover of the original network. To answer this question, we analyze the relation between the modularity of the maximal clique network and the Q_c of the original network.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_l\}$ be a partition of the maximal clique network and $\mathcal{C} = \{c_1, c_2, \ldots, c_l\}$ be the corresponding cover of the original network. Here, l is the size of $\mathcal{P}$ or $\mathcal{C}$, i.e., the number of communities. Using modularity, the quality of the partition $\mathcal{P}$ can be measured by

$$Q = \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \left(B_{xy} - \frac{s_x s_y}{L'} \right). \tag{12}$$

Using equations (9) and (10), we have

$$\begin{aligned} Q &= \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \left(\sum_{vw} \alpha_{vm_x} \alpha_{wm_y} A_{vw} - \frac{1}{L'} \sum_v \alpha_{vm_x} k_v \sum_w \alpha_{wm_y} k_w \right) \\ &= \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} \left(A_{vw} - \frac{k_v k_w}{L'} \right) \end{aligned}$$

$$= \frac{1}{L}\sum_{i}\sum_{vw}\alpha_{vc_i}\alpha_{wc_i}\left(A_{vw}-\frac{k_v k_w}{L}\right)$$
$$= Q_{\mathrm{c}}. \tag{13}$$

Equation (13) tells us that the optimization of the Q_{c} on the original network is equivalent to the optimization of the modularity on the maximal clique network. Thus we can find the optimal cover of the original network by finding the optimal partition of the corresponding maximal clique network. The optimal cover reflects the overlapping community structure of the original network.

2.3. Discussions

As to our method, it is important to select an appropriate parameter k. On one hand, the parameter k affects the constituent of the overlapping regions between communities. According to the definition of subordinate vertices, they are excluded from the overlapping regions. Thus the larger the parameter k, the less the number of vertices which can occur in the overlapping regions. When $k \to \infty$, the maximal clique network is identical to the original network and no overlap is identified. On the other hand, since the subordinate maximal cliques are not so highly connective, the parameter k should not be too small in practice. The choice of the parameter k depends on the specific networks. Observed from many real world networks, the typical value of k is often between 3 and 6. Additionally, as to networks where larger cliques are rare, our method is close to the traditional modularity-based partition methods. In this case, rare overlaps will be found.

Both the traditional modularity and the Q_{c} are based on the significance of link density in communities compared to a null-model reference network, e.g., the configuration model network. However, differently from the traditional modularity which requires that each node can only belong to one community, Q_{c} requires that each maximal clique can only belong to one community. In this way, Q_{c} takes advantage of both the local topological structure (i.e., the maximal clique) and the global statistical significance of link density. Similarly to the traditional modularity, however, the measure Q_{c} also suffers the resolution limit problem [15], especially when applied to large scale complex networks. Recently, some methods [26] have been proposed to address the resolution limit problem of modularity. These methods are also appropriate to the measure Q_{c}.

Now we turn to the efficiency of our method. It is difficult to give an analytical form of the computational complexity of our method. Here we only discuss what influences the efficiency of our method. Our method consists of three stages, finding out the maximal cliques, constructing the maximal clique network and partitioning the maximal clique network. As to the first stage, we need to find out all the maximal cliques in the network. It is widely believed to be a non-polynomial problem. However, for real world networks, finding all the maximal cliques is easy due to the sparseness of these networks. The computational complexity of the second stage depends on the number of edges in the original networks. Finally, the partition stage rests with the number of the maximal cliques and subordinate vertices. Taken together, our method is very efficient on real world networks.

In addition, as mentioned above, the overlapping community structure can be identified by the optimization of Q_{c}. Similarly, iteratively applying this method to each

community, we can investigate the sub-communities correspondingly. In this way, a rigid hierarchical relation of overlapping communities can be identified from the whole network.

3. Results

In this section, we extensively test our method on artificial networks and real world networks with a known community structure. Then we apply our method to a large real world complex network, which has been shown to possess an overlapping community structure.

3.1. Tests on artificial networks

To test our method, we utilize the benchmark proposed in [27]. It provides benchmark networks with heterogeneous distributions of node degree and community size. In addition, it allows for the overlaps between communities. This benchmark poses a much more severe test to community detection algorithms than Newman's standard benchmark [6]. There are many parameters to control the generated networks in this benchmark, the number of nodes N, the average node degree $\langle k\rangle$, the maximum node degree $\max k$, the mixing ratio μ, the exponent of the power-law node degree distribution $t1$, the exponent of the power-law distribution of community size $t2$, the minimum community size $\min c$, the maximum community size $\max c$, the number of overlapped nodes on, and the number of memberships of each overlapped node om. In our tests, we use the default parameter configuration where $N = 1000$, $\langle k\rangle = 15$, $\max k = 50$, $t1 = 2$, $t2 = 1$, $\min c = 20$, $\max c = 50$, $on = 50$ and $om = 2$. By tuning the parameter μ, we test the effectiveness of our method on the networks with different fuzziness of communities. The larger the parameter μ, the fuzzier the community structure of the generated networks is.

To evaluate the effectiveness of an algorithm for the identification of overlapping community structure, a measure is needed to compare the cover found by the algorithm with the ground truth. In [23], a measure is proposed to compare two covers, which is an extension form of *variation of information*. The more similar two covers are, the higher the value of the measure is. In this paper, we adopt it to compare the overlapping community structure found by our method and the known overlapping community structure in the benchmark networks.

Figure 4 shows the results of our method with $k = 4, 5, 6$ on the benchmark networks. Our method gives rather good results when the μ is smaller than 0.5. All of the values of the variation of information are above 0.8. Note that in these cases, communities are defined in the strong sense [28], i.e., each node has more neighbors in its own community than in the others. We also test other settings of k which are larger than 6, and find similar results.

3.2. Tests on real world networks

Our first real world network for test is Zachary's karate club network [29], which is widely used as a benchmark for the methods of community identification. This network characterizes the social interactions between the individuals in a karate club at an American university. A dispute arose between the club's administrator and its principal

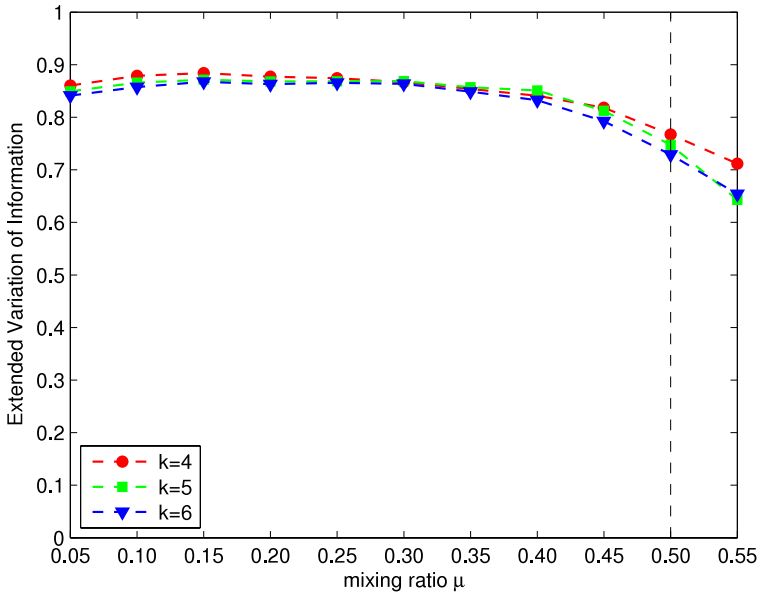

Figure 4. Test of our method on the benchmark networks. The parameter k in the legend corresponds to the parameter k in our method. The threshold $\mu = 0.5$ (dashed vertical line in the figure) marks the border beyond which communities are no longer defined in the strong sense [28], i.e., such that each node has more neighbors in its own community than in the others. Each point corresponds to an average over 100 graph realizations.

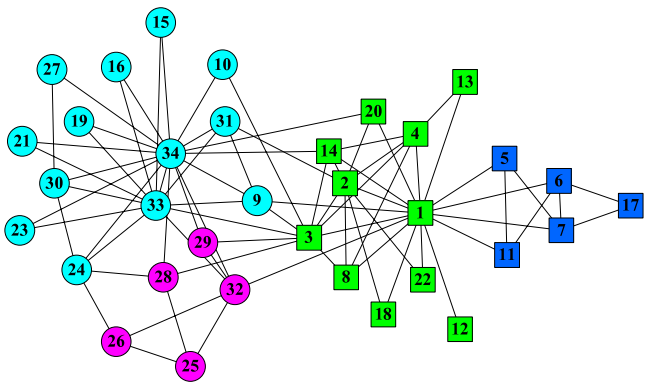

Figure 5. The network of the karate club studied by Zachary [29]. The real social fission of this network is represented by two different shapes, circle and square. The different colors show the partition obtained by our method with the parameter $k = 4$.

karate teacher, and as a result the club eventually split into two smaller clubs, centered around the administrator and the teacher respectively. The network and its fission is depicted in figure 5. The administrator and the teacher are represented by nodes 1 and 33 respectively.

Feeding this network into our method with the parameter $k = 4$, we obtain the result shown in figure 5. Similar to many existing community detection methods, our

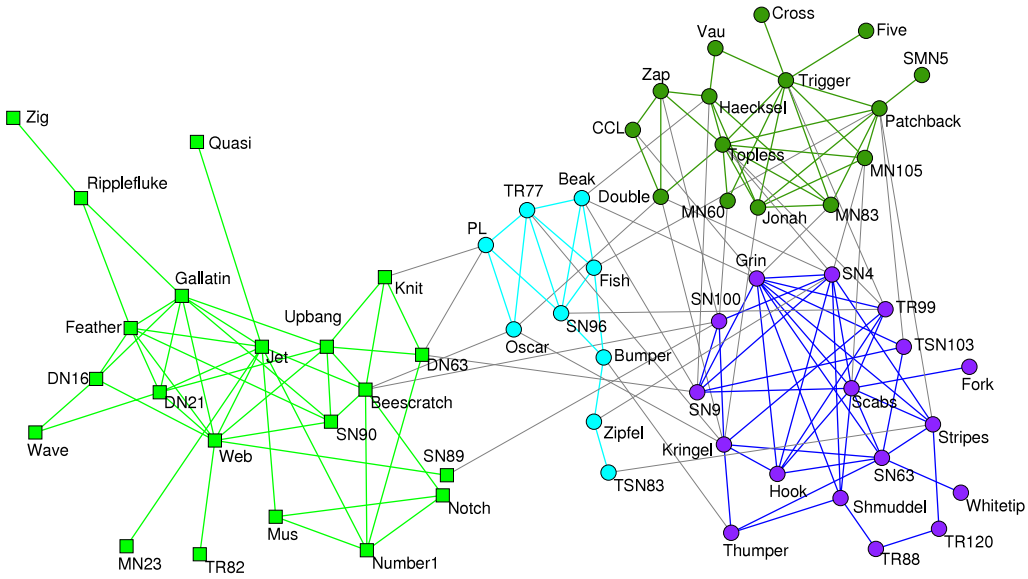

Figure 6. The community structure identified by our method from the network of the bottlenose dolphins of Doubtful Sound. The primary split of the network is represented by different shapes, square and circle. The different colors show the partition obtained by our method with the parameter $k = 4$.

method partitions the network into four communities. This partition corresponds to the modularity with the value 0.417, while the real partition into two sub-networks has a modularity 0.371. Actually, no vertex is misclassified by our method. The real split of the network can be obtained exactly by pair-wise merging of the four communities found by our method.

We also note that no overlaps are found when $k = 4$. Actually, no overlaps can be found when k is no smaller than 4 for this network. Overlaps between communities emerge when the parameter k is set to 3. The value of Q_c corresponding to the resulting cover is 0.385 and in total three overlapped communities are found by our method. They are {1, 5, 6, 7, 11, 17}, {1, 2, 3, 4, 8, 9, 12, 13, 14, 18, 20, 22} and {3, 9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34}. The overlapping regions consist of three vertices, being 1, 3 and 9. Each of them is shared by two communities. Such vertices are often misclassified by traditional partition-based community detection methods. Except the vertices occurring in the overlapping regions, other vertices reflects the real split of the network.

We also test our method on another real world network, a social network of 62 bottlenose dolphins living in Doubtful Sound, New Zealand. The network was constructed by Lusseau [30], with ties between dolphin pairs being established by observation of statistically significant frequent association. The network splits naturally into two groups, represented by the squares and circles in figure 6.

By applying our method with $k = 4$ to this network, four communities are obtained, denoted by different colors in figure 6. The green community is connected loosely to the other three ones. Regarding the three circle-denoted communities as a sole community, it and the green community correspond to the known division observed by Lusseau [30].

Furthermore, the three circle-denoted communities also correspond to a real division among these dolphins. The further division appears to have some correlation with the gender of these animals. The blue one consists mainly of females and the other two almost entirely of males.

Similarly to the Zarchay's karate network, the overlaps between communities cannot be detected when the parameter k is not less than 4. When $k = 3$, overlaps between the circle-denoted communities emerge while the green community keeps almost intact. The Q_c is 0.490 as to the resulting cover. The vertices occurring in overlapping regions are Beak, Kringel, MN105, Oscar, PL, SN4, SN9 and TR99 among which the vertices *Beak* and *Kringel* are shared by all the three circle-denoted communities. Again these overlapping vertices are often misclassified by traditional partition-based methods.

3.3. Application to the word association network

Now we apply our method to a large real world complex network, namely the word association network.

The data set for the word association network is from the demo of the software *CFinder* [31]. This network consists of 7207 vertices and 31784 edges, and has been shown to possess overlapping community structure [17]. It is constructed from the South Florida Free Association norms list [32]. Initially, the network is a directed, weighted network. The weight of a directed edge from one word to another indicates the frequency that the people in the survey associated the end point of the edge with its start point. These directed edges were replaced by undirected ones with a weight equal to the sum of the weights of the corresponding two oppositely directed edges. Furthermore, the edges with the weight less than 0.025 were deleted. In this way, an un-weighted, undirected network is obtained, and it is this network we deal with.

Applying our method to the word association network, we obtain in total 20 communities which overlap with each other. The value of the corresponding Q_c is as high as 0.503, indicating a strong overlapping community structure. The size of these found communities are very large so that there is no specific semantic meaning for each community. To investigate what is correlated to the overlapping community structure, we apply our method to these communities iteratively and a hierarchy of overlapping communities is obtained. We find that the sub-communities have a certain correlation with semantic meaning of words. As an example, table 1 shows us the communities around the word *play*. The five overlapping communities represent different meanings of the word *play*, respectively related to *theater*, *musical instruments*, *children*, *sports* and *toys*. Except the common-shared word *play*, four other words are shared by some of these communities. They are *fun*, *game*, *toy* and *toys*. The overlap between these communities characterizes the direct, local relationship between them through sharing members. However, the extent of closeness between communities is sometimes reflected by the indirect, global relationship between them. One of this kind of relationship is the 'genealogical' relationship between communities, which can be illustrated by the hierarchy of overlapping communities. Figure 7 is an example for hierarchy of communities. As shown in figure 7, the communities 1 and 2 are in the same branch of the hierarchy, indicating that the meanings represented by them are closer. This can be validated by examining the words contained in these two communities. Similarly, the communities 4

J. Stat. Mech. (2009) P07042

Table 1. The overlapping communities around the word *play*. For each community, a short description is also given. The overlapped words are emphasized in italic text.

No.	Description	Words in each community
1	Theater	act actor actress bow character cinema curtsey dance director do drama entertain entertainment film guide involve juggler lead movie participate perform performance *play* portray producer production program scene screen show sing stage television theater
2	Musical instrument	alto band banjo bass beep blues brass bugle cello clarinet clef compose concert conductor country drum faddle fiddle flute guitar harp honk horn instrument jazz keyboard loud music oboe orchestra piano *play* rock saxophone symphony tenor treble trombone trumpet tuba tune viola violin woodwind
3	Children	adults balls children family friends *fun* grown-ups guardians kids love mischief nursery parents *play* playground play_dough prank putty *toy toys* tricycle
4	Sports	active arena athlete athletic baseball basketball black_and_white field football *fun game* illustrated inactive jock pigskin *play* recreation referee soccer sports stadium umpire
5	Toys	board boardwalk checkers chess *fun game* games monopoly nintendo *play* plaything strategy *toy toys* vcr video winning yo_yo

and 5 are also closely related. However, the distance between the communities 3 and 5 is larger although they share as many as 4 words. The overlaps between communities and the hierarchy of these communities provide us a more complete understanding to the relationship between communities.

4. Conclusions

This paper focuses on the problem of quantifying and identifying the overlapping community structure of networks. There are two main contributions. Firstly, a measure Q_c for the quality of a cover of network is proposed to quantify the overlapping community structure. The effectiveness of the measure Q_c is demonstrated by the experimental results that networks with significant overlapping community structure have a cover with a high Q_c. Secondly, a maximal clique network is constructed from the original network, and then the overlapping community structure can be identified using any modularity optimization method on the maximal clique network.

The Q_c is an extension of traditional modularity with the consideration that the maximal clique instead of a single node can only belong to one community. In this way, Q_c takes advantage of both the local topological structure (i.e., the maximal clique) and the global statistical significance of link density compared with a null-model reference network. In addition, Q_c can be naturally used to simultaneously identify the overlapping

J. Stat. Mech. (2009) P07042

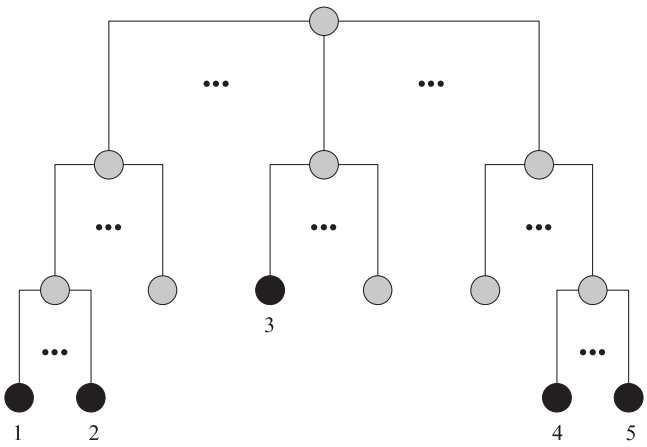

Figure 7. Part of the hierarchy of communities extracted from the word association network. The dark-filled circles correspond to the five communities shown in table 1.

and hierarchical community structure of networks. Such a method is helpful to more completely understand the functional and structural properties of networks.

As further work, we will consider the generalization to weighted and/or directed networks. There is also an interesting problem about the selection of the parameter k in our method. We will further investigate how to determine an appropriate k for a given network in the future.

Acknowledgments

This work was funded by the National Natural Science Foundation of China under grant number 60873245, the National High-Tech R&D Program of China (the 863 program) under grant number 2006AA01Z452, and the National Basic Research Program of China (the 973 program) under grant number 2004CB318109. The authors gratefully acknowledge S Fortunato and A Lancichinetti for providing the test benchmark and useful discussions on it. The authors thank Mao-Bin Hu for helpful suggestions.

References

[1] Albert R and Barabási A-L, 2002 *Rev. Mod. Phys.* **74** 47
[2] Newman M E J, 2003 *SIAM Rev.* **45** 167
[3] Cheng X Q, Ren F X, Zhou S and Hu M B, 2009 *New J. Phys.* **11** 033019
[4] Flake G W, Lawrence S, Giles C L and Coetzee F M, 2002 *IEEE Comput.* **35** 66
[5] Girvan M and Newman M E J, 2002 *Proc. Nat. Acad. Sci.* **99** 7821
[6] Newman M E J and Girvan M, 2004 *Phys. Rev.* E **69** 026113
[7] Newman M E J, 2004 *Phys. Rev.* E **69** 066133
[8] Clauset A, Newman M E J and Moore C, 2004 *Phys. Rev.* E **70** 066111
[9] Guimerà R and Amaral L A N, 2005 *Nature* **433** 895
[10] Duch J and Arenas A, 2005 *Phys. Rev.* E **72** 027104
[11] Newman M E J, 2006 *Proc. Nat. Acad. Sci.* **103** 8577
[12] Raghavan U N, Albert R and Kumara S, 2007 *Phys. Rev.* E **76** 036106
[13] Sales-Pardo M, Guimerà R, Moreira A A and Amaral L A N, 2007 *Proc. Nat. Acad. Sci.* **104** 15224
[14] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E, 2008 *J. Stat. Mech.* P10008
[15] Fortunato S and Barthélemy M, 2007 *Proc. Nat. Acad. Sci.* **104** 36
[16] Kumpula J M, Saramaki J, Kaski K and Kertesz J, 2007 *Eur. Phys. J.* B **56** 41

J. Stat. Mech. (2009) P07042

[17] Palla G, Derényi I, Farkas I and Vicsek T, 2005 *Nature* **435** 814
[18] Baumes J, Goldberg M K, Krishnamoorthy M S, Magdon-Ismail M and Preston N, 2005 *Proc. IADIS Applied Computing* p 97
[19] Zhang S, Wang R S and Zhang X S, 2007 *Physica* A **374** 483
[20] Palla G, Farkas I J, Pollner P, Derényi I and Vicsek T, 2007 *New J. Phys.* **9** 186
[21] Farkas I J, Ábel D, Palla G and Vicsek T, 2007 *New J. Phys.* **9** 180
[22] Shen H W, Cheng X Q, Cai K and Hu M B, 2009 *Physica* A **388** 1706
[23] Lancichinetti A, Fortunato S and Kertész J, 2009 *New J. Phys.* **11** 033015
[24] Nicosia V, Mangioni G, Carchiolo V and Malgeri M, 2009 *J. Stat. Mech.* P03024
[25] Evans T S and Lambiotte R, 2009 *Phys. Rev.* E **80** 016105
[26] Arenas A, Fernández A and Gómez S, 2008 *New J. Phys.* **10** 053039
[27] Lancichinetti A and Fortunato S, 2009 arXiv:0904.3940
[28] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D, 2004 *Proc. Nat. Acad. Sci.* **101** 2658
[29] Zachary W W, 1977 *J. Anthropol.* **33** 452
[30] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M, 2003 *Behav. Ecol. Sociobiol.* **54** 396
[31] Adamcsek B, Palla G, Farkas I J, Derényi I and Vicsek T, 2006 *Bioinformatics* **22** 1021
[32] Nelson D L, McEvoy C L and Schreiber T A, 1998 http://w3.usf.edu/FreeAssociation/

J. Stat. Mech. (2009) P07042