J Comb Optim (2012) 23:425–442
DOI 10.1007/s10878-010-9356-0

A combinatorial model and algorithm for globally searching community structure in complex networks

**Xiang-Sun Zhang · Zhenping Li ·
Rui-Sheng Wang · Yong Wang**

Published online: 15 October 2010

Abstract Community structure is one of the important characteristics of complex networks. In the recent decade, many models and algorithms have been designed to identify communities in a given network, among which there is a class of methods that globally search the best community structure by optimizing some modularity criteria. However, it has been recently revealed that these methods may either fail to find known qualified communities (a phenomenon called resolution limit) or even yield false communities (the misidentification phenomenon) in some networks. In this paper, we propose a new model which is immune to the above phenomena. The model is constructed by restating community identification as a combinatorial optimization problem. It aims to partition a network into as many qualified communities as possible. This model is formulated as a linear integer programming problem and its NP-completeness is proved. A qualified min-cut based bisecting algorithm is designed to solve this model. Numerical experiments on both artificial networks and real-life complex networks show that the combinatorial model/algorithm has promising performance and can overcome the limitations in existing algorithms.

Keywords Complex network · Community structure · Heuristic algorithm · Linear integer programming · NP-hard

X.-S. Zhang · Y. Wang (✉)
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China
e-mail: ywang@amss.ac.cn

X.-S. Zhang
e-mail: zxs@amt.ac.cn

Z. Li
School of Information, Beijing Wuzi University, Beijing 101149, China

R.-S. Wang
Department of Physics, Pennsylvania State University, University Park, PA 16802, USA

1 Introduction

In recent years complex network research becomes a hot topic since many systems in the real world can be represented as a network, in which nodes denote the objects of interest and edges that connect nodes describe the relationships between them. Examples range from social networks, technological networks to biological networks such as the scientific collaboration networks, email communication networks and protein interaction networks. These different types of complex networks have been revealed to have many common topological features such as scale-free and small-world (Albert and Barabási 2002). Importantly, many complex networks have community or modular structure (Girvan and Newman 2002; Fortunato 2010), i.e., networks consist of specific and relatively separate dense subnetworks. In a widely used qualitative definition, a community is a sub-network whose nodes are connected tightly inside and sparsely to the outside (Ravasz et al. 2002; Radicchi et al. 2004; Newman 2006; Wang et al. 2008; Zhang et al. 2007). Uncovering such community structure not only helps us understand the topological structure of large-scale networks, but also reveals the functionality of each component. This is, for example, confirmed by the modular organization of biological networks (Ravasz et al. 2002), where the communities are sets of components with similar functions and the modular structure is the result of evolutionary constraints.

A large number of methods have been developed for detecting communities in complex networks. As summarized in Zhang et al. (2009), existing models and algorithms for community identification can be generally categorized into local and global methods. Local methods identify a subset of nodes as a community according to certain local connection conditions, independently from the structure of the rest of the network. Methods in this category include clique overlap-based hierarchical clustering (Everett and Borgatti 1998), clique percolation method (Palla et al. 2005), and subgraph fitness method (Lancichinetti et al. 2009). Global methods usually optimize certain quantitative functions encoding the quality of the overall partition of the network, such as Potts model (Reichardt and Bornholdt 2004), information theoretical method (Rosvall and Bergstrom 2007), random walk methods (Rosvall and Bergstrom 2008), and optimization of modularity measures (Newman and Girvan 2004; Li et al. 2008). One popular measure used in the global models for community detection is the modularity function Q developed by Newman and Girvan (2004). A large number of methods have been devised for community detection based on optimizing Q (Newman 2006; Guimerà and Amaral 2005; Zhu et al. 2007; Schuetz and Caflisch 2008; Agarwal and Kempe 2008). In a recent study, Li et al. proposed another quantitative measure D called modularity density to improve community detection (Li et al. 2008). In addition to local methods and global methods, there is another important class, namely, multiresolution methods (Arenas et al. 2008; Ronhovde and Nussinov 2009; Cafieri et al. 2010). The common motivation underlying them is that community structure of complex networks is complicated and has multiple scales and that methods which allow analyzing different scales of communities or a hierarchical structure of communities are more practical.

Besides the community detection methods mentioned above, some studies explore quantitative definitions for community to make the community identification problem more quantitatively solved. Radicchi et al. gave a quantitative community definition in a weak sense (Radicchi et al. 2004), where the weak community definition states that the sum of degrees within a candidate subnetwork should exceed the sum of all degrees towards the rest of the network. Hu et al., introduced a definition entitled as "the most weak" (Hu et al. 2008), where the sum of degrees within a candidate subnetwork only needs to exceed the sum of degrees towards any one of other candidate subnetworks.

Although much research has been done on community detection in the global sense, existing models and algorithms suffer from various limitations. For example, optimization of modularity function Q has been exposed to suffer from a so-called resolution limit problem, i.e., communities in some special networks may not be resolved by optimization of Q even in an extreme case where the network consists of complete graphs connected by single bridges (Fortunato and Barthelemy 2007). In other words, optimization of Q fails to zoom in some small qualified communities. Modularity density D alleviated the resolution limit to some extent (Li et al. 2008). In addition to the resolution limit phenomenon, there is another serious limitation in optimization of Q and D, that is, the misidentification phenomenon (Zhang et al. 2009), which means that some derived communities do not satisfy the weak community definition or even the most weak community definition. In other words, these communities have sparser connection within them than between them which disobeys the basic intuitive sense for a subnetwork to be a community. For the resolution limit of Q, an excellent improvement is the multiresolution algorithm developed in Arenas et al. (2008), which introduces a parameter into the modularity formula. This parameter does not affect the structural properties of the graph in most cases and allows analyzing different scales of community structure from microscale to macroscale. In this study, we seek to overcome the limitations of modularity optimization from another perspective.

In this paper, we take community identification as a combinatorial optimization problem. We build a model to partition a network into as many qualified communities as possible, formulate it into a linear integer programming problem, and prove the NP-completeness of this problem. Then a qualified min-cut based bisecting heuristic algorithm is designed to solve this model. Experimental results in both artificial networks and real-life complex networks are presented to show the effectiveness of our proposed model and algorithm.

2 A combinatorial model for community identification

Given a network $G=(V,E)$ and its adjacency matrix $A=[a_{ij}]$, we denote V_s as a subset of V and $\overline{V}_s = V\backslash V_s$ as the complementary set of V_s, then V_s induces a community in a weak sense (Radicchi et al. 2004) if

$$L(V_s, V_s) > L(V_s, \overline{V_s}), \tag{1}$$

where $L(V_s, V_s) = \sum_{i\in V_s}\sum_{j\in V_s} a_{ij}$, $L(V_s, \overline{V_s}) = \sum_{i\in V_s}\sum_{j\in \overline{V_s}} a_{ij}$. A looser community definition called "the most weak community definition" has been proposed as follows (Hu et al. 2008): V_s induces a community if

$$L(V_s, V_s) > \max_{t:t \neq s} L(V_s, V_t). \tag{2}$$

Given a partition $P_K = (G_1, G_2, \dots, G_K) = ((V_1, E_1), \dots, (V_K, E_K))$ of the network G, where K is the number of communities induced by the partition, the modularity function Q (Newman and Girvan 2004) is defined as

$$Q = \sum_{s=1}^{K} \left[\frac{L(V_s, V_s)}{2L} - \left(\frac{L(V_s, V_s) + L(V_s, \overline{V_s})}{2L} \right)^2 \right] \equiv \sum_{s=1}^{K} Q_s, \tag{3}$$

where $L = L(V, V)/2$ is the total number of links in the network. This measure compares the number of edges inside a given subnetwork with the expected value in a randomized network of the same size and the same degree distribution. A large number of methods have been devised for community detection by optimizing Q (Newman 2006; Guimerà and Amaral 2005; Zhu et al. 2007; Schuetz and Caflisch 2008; Agarwal and Kempe 2008).

However, Q has been exposed to suffer from a so-called resolution limit problem, i.e., communities in some special network structures may not be resolved by optimization of Q even in an extreme case where the network consists of cliques connected by single links (Fortunato and Barthelemy 2007). This is illustrated via a simple example in Fig. 1(a), where optimization of Q will group several cliques into a community. Li et al. proposed another quantitative measure D called modularity density to evaluate candidate community structures of a given network (Li et al. 2008):

$$D = \sum_{s=1}^{K} \frac{L(V_s, V_s) - L(V_s, \overline{V_s})}{|V_s|} \equiv \sum_{s=1}^{K} D_s, \tag{4}$$

where $|V_s|$ is the number of nodes in community s. This measure, based on the concept of graph density, incorporates node information in a community and improves the resolution limit of Q. However, in Zhang et al. (2009), the authors point out that D is not completely free from the resolution limit through a discrete convex/concave programming analysis.

In addition to the resolution limit phenomena of Q and D revealed in Fortunato and Barthelemy (2007) and Zhang et al. (2009), there is another serious limitation in optimization of Q and D, that is, misidentification. It means that some derived communities do not satisfy the weak community definition or even the most weak community definition, and thus have sparser connection within them than between them. This misidentification problem disobeys the basic intuitiveness for a subnetwork to be a community. An example of misidentification is illustrated in Fig. 1(b), where, for example, let $m = 5$, then optimization of Q partitions the network into three communities (C_n, C_{n-1}, C_m) when $n \geq 13$, and optimization of D partitions the network into three communities (C_n, C_{n-1}, C_m) when $n \geq 21$ (Zhang et al. 2009). However, C_m is a subnetwork violating the most weak community definition.

The above discussion states that modularity optimization of Q or D is only an approximate procedure. Then it is necessary to set up a complete model to accurately

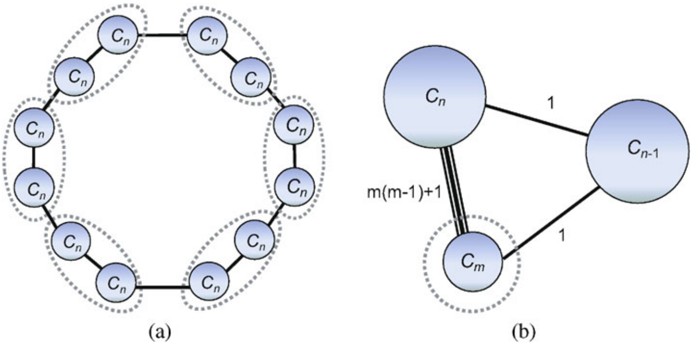

Fig. 1 Illustration of the resolution limit and misidentification phenomena. (**a**) The network consists of a ring of n cliques ($n \geq 3$), connected through single links. Assuming that there are 2^k cliques. The network has a clear modular structure where each community corresponds to a single clique. But optimizing Q combines two neighboring cliques as one community (Fortunato and Barthelemy 2007) as shown in the subfigure (**a**) and fails to obtain the correct partition. (**b**) The network consists of three cliques C_n, C_{n-1}, C_m with different sizes. When $n \gg m$, the clique C_m is not a qualified community, however, both Q and D will identify C_m as a community in some range of n

characterize the solution. In this paper, we reconsider the community identification problem from the view of combinatorial optimization. In essence, it is a problem to partition a network into subnetworks such that each subnetwork satisfies a given community definition which could be one ranging from the weak definition to the most weak definition. There are many partitions of a given network satisfying the requirement. Among them a partition with finest structure, that is, any subnetwork cannot be further split into two smaller communities, is the desired solution.

Problem definition Given a network, the community identification problem is to partition the network into as many non-overlapping subnetworks as possible such that each subnetwork satisfies a given community definition.

In the following of this paper, we use the weak community definition (1) to set up a mathematical model. For other community definitions there would be a similar framework. Let n be the number of nodes in the network and it is also the maximum number of possible communities, and L be the number of edges in the network. z_{lk} denotes whether the edge e_l belongs to the k-th community, $l = 1, 2, \ldots, L, k = 1, 2, \ldots, n$, where $e_l = (v_i, v_j)$ represents the edge connecting the nodes v_i and v_j. Let x_{ik} be a binary variable indicating whether the node v_i belongs to community k. Then the relations between z_{lk} and x_{ik}, x_{jk} can be described as:

$$z_{lk} \leq x_{ik} \quad \text{and} \quad z_{lk} \leq x_{jk}$$

which indicate that if one adjacent node of an edge is not in community k, then this edge definitely does not belong to the community. We use

$$x_{ik} + x_{jk} - 1 \leq z_{lk}$$

to indicate that if both v_i and v_j are in community k, then the edge connecting these two nodes must be in the community. Let y_k be a binary variable denoting whether the k-th community is empty. $y_k = 0$ if and only if the k-th community has no nodes, so

$$y_k \leq \sum_{i=1}^{n} x_{ik} \leq n y_k.$$

The weak community definition condition when $y_k = 1$ can be formulated as:

$$2\sum_{l=1}^{L} z_{lk} > \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk}. \tag{5}$$

To incorporate the case when $y_k = 0$, we restate the inequality (5) as follows,

$$2\sum_{l=1}^{L} z_{lk} \geq \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk} + y_k. \tag{6}$$

Therefore, the community identification model can be formulated as a linear integer programming as follows:

$$\begin{aligned}
\max \quad & \sum_{k=1}^{n} y_k \\
\text{s.t.} \quad & \sum_{k=1}^{n} x_{ik} = 1 \\
& z_{lk} \leq x_{ik} \\
& z_{lk} \leq x_{jk} \\
& x_{ik} + x_{jk} - 1 \leq z_{l,k} \\
& \sum_{i=1}^{n} x_{ik} \geq y_k \\
& \sum_{i=1}^{n} x_{ik} \leq n y_k \\
& 2\sum_{l=1}^{L} z_{lk} \geq \sum_{j=1}^{n}\sum_{i=1}^{n} x_{ik} a_{ij} - 2\sum_{l=1}^{L} z_{lk} + y_k \\
& x_{ik} \in \{0, 1\}, y_k \in \{0, 1\}, z_{lk} \in \{0, 1\} \\
& i = 1, 2, \ldots, n, k = 1, 2, \ldots, n, l = 1, 2, \ldots, L
\end{aligned} \tag{7}$$

In general, the linear integer programming is NP-hard and directly solving the above model is not realistic for large networks. However, this linear integer programming model is important in designing approximate algorithms with approximation factor guarantees, through widely used primal-dual methods. In the subsequent section, we will prove the complexity of the community identification problem and give a qualified min-cut based bisecting heuristic algorithm.

3 NP-completeness of the community identification problem

Let $G=(V,E)$ be an undirected graph, and $d_G(v)=|\{u\in V|(u,v)\in E\}|$ denote the degree of the node v in G. A graph G is a cubic graph if $d_G(v)=3$ for every $v\in V$. Any subset of vertices $S\subseteq V$ creates a cut of G, which is denoted by $C(S,\bar{S})=\{(u,v)|u\in S, v\in V\setminus S\}$. The size of $C(S,\bar{S})$ is defined as $L_G(S,\bar{S})=|C(S,\bar{S})|$ and $d_G(S)=\sum_{v\in S} d_G(v)=L_G(S,S)+L_G(S,\bar{S})$ denotes the sum of degrees of the nodes in the subset $S\subseteq V$. To prove the NP-completeness of the problem, we first prove the NP-completeness of a simplest case, that is, partition a network into two subnetworks such that each subnetwork satisfies the weak community definition, which we call as the qualified cut problem. The corresponding decision version of this problem can be formulated as follows:

The Qualified Cut Problem
Instance: An undirected graph $G=(V,E)$.
Question: Is there a subset $S\subset V$ such that $L_G(S,S)>L_G(S,\bar{S})$ and $L_G(\bar{S},\bar{S})>L_G(\bar{S},S)$

Then we show that any instance of the maximum cut problem for cubic graph, which has been proved to be NP-hard (Alimonti and Kann 2000), can be transformed into a qualified cut problem in polynomial time, and the solution of the maximum cut problem for cubic graph exactly corresponds to that of the qualified cut problem. We note that the qualified cut problem is a special case of the decision version for the conductance problem, which has been often stated to be an NP-complete problem in the literature (Šíma and Schaeffer 2006). Thus we can borrow some ideas from the NP-completeness proof from Šíma and Schaeffer (2006) and Shi and Malik (2000). The detailed NP-completeness proof of our qualified cut problem is as follows.

Maximum Cut for Cubic Graph (Max Cut-3)
Instance: A cubic graph $G=(V,E)$ and a positive integer a.
Question: Is there a cut $(B,\bar{B})$, such that $L_G(B,\bar{B})>a$?

Theorem 1 *The Qualified Cut Problem is NP-complete.*

Proof The Qualified Cut Problem belongs to NP since a nondeterministic algorithm can guess a subset $S\subset V$ and verify $L_G(S,S)>L_G(S,\bar{S})$ and $L_G(\bar{S},\bar{S})>L_G(\bar{S},S)$ in polynomial time. The NP-completeness can be proved by reducing the maximum cut problem on cubic graphs to the Qualified Cut Problem in polynomial time.

Given a Max Cut-3 instance, that is, a cubic graph $G=(V,E)$ with $n=|V|$ vertices, and a positive integer $a=n$, a corresponding undirected graph $G'=(V',E')$ for the Qualified Cut Problem is constructed to be composed of two fully connected copies of the complement of G, that is $V' = V_1 \cup V_2$ where $V_i = \{v^i | v \in V\}$ for $i = 1, 2$, and $E' = E_1 \cup E_2 \cup E_3$ where $E_i = \{(u^i, v^i) | u, v \in V, u \neq v, (u, v) \notin E\}$ for $i = 1, 2$, and $E_3 = \{(u^1, v^2) | u, v \in V\}$. The number of vertices in G' is $|V'| = 2n$ and the number of edges is $|E'| = (2n - 4)n$ since $d_{G'}(v) = 2n - 4$ for every $v \in V'$ due to G being a cubic graph. It follows that G' can be constructed in polynomial time. For a subset $\emptyset \neq S \subset V'$ in G' with $k = |S| < 2n$ vertices, $S_i = \{v \in V | v^i \in S\}$ for $i = 1, 2$ are projections of S to V_1 and V_2, respectively.

Now we verify the correctness of the reduction by proving that the Max Cut-3 instance has a solution if and only if the corresponding Qualified Cut instance is solvable. Firstly we assume that a cut $(B, \bar{B})$ exists in G whose size satisfies $L_G(B, \bar{B}) > n$. Let $S^B = \{v^1 \in V_1 | v \in B\} \cup \{v^2 \in V_2 | v \in \bar{B}\} \subseteq V'$ be the subset in G' whose projections to V_1 and V_2 are $S_1^B = B$ and $S_2^B = \bar{B}$ respectively. Since $|S^B| = n$, $L_G(B, \bar{B}) = L_G(\bar{B}, B)$, $d_{G'}(S^B) = d_{G'}(\overline{S^B})$, $d_{G'}(S^B) = n(2n - 4)$, and

$$\begin{aligned} L_{G'}(S^B, \overline{S^B}) &= n(2n - n) - L_G(S_1^B, \overline{S_1^B}) - L_G(S_2^B, \overline{S_2^B}) \\ &= n^2 - L_G(B, \bar{B}) - L_G(\bar{B}, B) \\ &< n^2 - 2n, \end{aligned}$$

then $d_{G'}(S^B) = d_{G'}(\overline{S^B}) > 2L_{G'}(S^B, \overline{S^B})$. So, $L_{G'}(S^B, S^B) > L_{G'}(S^B, \overline{S^B})$ and $L_{G'}(\overline{S^B}, \overline{S^B}) > L_{G'}(S^B, \overline{S^B})$, which means that the two communities induced by S^B and $\overline{S^B}$ satisfy the weak definition.

For the converse, assume that the subset $\emptyset \neq S \subset V'$ in $G' = (V', E')$ satisfies the weak definition, i.e., $L_{G'}(S, S) > L_{G'}(S, \bar{S})$ and $L_{G'}(\bar{S}, \bar{S}) > L_{G'}(\bar{S}, S)$. Without loss of generality, assume $|S| = k \leq n$. It is easy to verify that $k \geq 2$ according to the weak definition. So we have $2 \leq k \leq n$. Let $S_i = \{v \in V | v^i \in S\}$ for $i = 1, 2$. Since $d_{G'}(S) = L_{G'}(S, S) + L_{G'}(S, \bar{S}) = k(2n - 4)$, $L_{G'}(S, \bar{S}) = k(2n - k) - L_G(S_1, \bar{S_1}) - L_G(S_2, \bar{S_2})$, from $L_{G'}(S, S) > L_{G'}(S, \bar{S})$, we have $k(2n - 4) > 2(k(2n - k) - L_G(S_1, \bar{S_1}) - L_G(S_2, \bar{S_2}))$, so $2 \leq k \leq n$, $L_G(S_1, \bar{S_1}) + L_G(S_2, \bar{S_2}) > k(n - k + 2) \geq 2n$. Let $B \subseteq V$, $(B, \bar{B})$ be a maximum cut in G, then $L_G(B, \bar{B}) \geq L_G(S_i, \bar{S_i})$ for $i = 1, 2$. So $2L_G(B, \bar{B}) \geq L_G(S_1, \bar{S_1}) + L_G(S_2, \bar{S_2}) > 2n$, then $L_G(B, \bar{B}) > n$. That is, $(B, \bar{B})$ is a solution of the Max Cut-3 instance. □

4 A qualified min-cut algorithm

Let $\widetilde{P}$ be the set of all partitions of the network G. A partition $P = (N_1, N_2, \ldots, N_K)$ is called *feasible* if all its resulting subnetworks satisfy the weak community definition. A partition

$$P' = (N_1, \ldots, N_{s-1}, N', N'', N_{s+1}, \ldots, N_K)$$

is a son-partition of $P = (N_1, N_2, \ldots, N_K)$ via a *simple-split operation* if $N' \cup N'' = N_s$ and N', N'' are nonempty connected subnetworks satisfying the weak community definition. It is easy to check that a son-partition of a feasible partition P via the simple-split operation is feasible.

A heuristic principle here is to obtain a feasible partition with the largest number of communities. A simple-split operation should produce two subnetworks as dense as possible or produce two subnetworks with minimal links between them. It is easy to bethink of the min-cut problem and its corresponding polynomial algorithms. Through finding the min-cut of a given graph, the network can be partitioned into two subnetworks N', N'' by removing the minimal cut between them. We call this process as a *min-cut operation*. Although we can use the min-cut operation step by step to partition the graph into several subgraphs, we cannot ensure that the derived subgraphs satisfy the weak definition. In the following, we will define a *qualified min-cut operation* and propose a heuristic algorithm based on qualified min-cut operation.

Given a graph $G = (V, E)$, we suppose $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$. A cut $C(V_1, V_2)$ is qualified if the subgraphs induced by V_1 and V_2 satisfy the weak definition in G. A qualified min-cut is a qualified cut with a minimum number of edges. The community identification problem can be solved based on a series of qualified min-cut operations.

The Qualified Min-Cut (QMC) algorithm

- Input a graph $G = (V, E)$. Set $\tilde{N} = \{G\}$, $\tilde{P} = \emptyset$.
- **Step 1**: Take one element N' in $\tilde{N}$, let $\tilde{N} = \tilde{N} \setminus \{N'\}$. Let $G_1 = N'$, $V_1 = V(G_1)$, $E_1 = E(G_1)$, $m = |V_1|$, $k = 1$, $C = \emptyset$, $T = \emptyset$, $W = \infty$.
- **Step 2**: Randomly select a vertex $v \in V_1$. $T = T \cup \{v\}$. Add the vertex that is most tightly connected with T into T step by step until $T = V_1$. The two vertices added last are denoted by s_k and t_k. Find the minimum s_k-t_k cut $C_k(s_k$-$t_k)$ of graph G_1.
- **Step 3**: Check whether the minimum s_k-t_k cut $C_k(s_k$-$t_k)$ is qualified. If C_k is qualified and $|C_k| < W$, then $C = C_k$ and $W = |C_k|$.
- **Step 4**: Revise graph G_1 to a new one by merging s_k and t_k to one node v_{st}. That is, $V_1 \leftarrow (V_1 \setminus \{s_k, t_k\}) \cup \{v_{st}\}$, $E_1 \leftarrow (E_1 \setminus \{(u, v)|u \in V_1, v = s_k, t_k\}) \cup \{(u, v_{st})|(u, s_k) \in E_1$ or $(u, t_k) \in E_1\}$. If $k = m - 1$, go to **Step 5**; otherwise, let k=k+1, $T = \emptyset$, go to **Step 2**.
- **Step 5**: If $C = \emptyset$, then there is no qualified cut in G_1, which means that N' can not be divided into two communities. Let $\tilde{P} = \tilde{P} \cup \{G_1\}$. Otherwise, C is a qualified cut for N'. Delete all the edges of C from graph N' and obtain two subgraphs N'_1 and N'_2. Let $\tilde{N} = \tilde{N} \cup \{N'_1, N'_2\}$.
- **Step 6**: Repeat **Step 1**—**Step 5** until $\widetilde{N} = \emptyset$. $\tilde{P}$ is an approximate optimal partition of the given graph G.

From the definition of QMC, we can obtain the following proposition.

Proposition 1 *If there is no qualified min-cut in a given network, then the network cannot be partitioned into communities satisfying the weak definition.*

Corollary 1 *If there is a potential community structure in a subnetwork N', the qualified min-cut operation will certainly divide N' into two qualified communities.*

The proposition and corollary described above guarantee that the resulting partition by the proposed QMC algorithm has no resolution limit and misidentification problems. The running time of the QMC algorithm mainly depends on the running time of Step 2. Step 2 is the algorithm of finding qualified min-cuts of a graph, which is based on the min-cut algorithm in Stoer and Wagner (1997) by adding a procedure checking whether a cut is qualified. The running time of the checking procedure is at most $O(|E|)$, and the running time of the min-cut algorithm is $O(|V||E| + |V|^2 \log|V|)$. So the running time of Step 2 is at most $O(|V||E| + |V|^2 \log|V|)$. Suppose there are at most k communities in the network. By the weak community definition, we know that $k < \min\{|V|/2, |E|/2\}$. The network can be divided into a union of communities by implementing Step 2 at most k times. So the total running time of the QMC algorithm is $O(k|V||E|+k|V|^2 \log|V|)$, which is less than $O(|V|^2|E|+|V|^3 \log|V|)$ of the algorithm in Girvan and Newman (2002).

5 Experimental results

In this section, we apply our QMC algorithm to a class of widely used artificial networks and to several well studied real-world networks, and compare the results of the QMC algorithm with that of optimization of Q and D in terms of classification accuracy and ability of detecting meaningful communities. The QMC algorithm is coded using the MATLAB version 7.6 and available upon request. The simulated annealing (SA) algorithm designed in Guimerà and Amaral (2005) is used to optimize Q and D.

5.1 Rings of cliques

We first test our algorithm on a type of widely used exemplar networks, that is, rings of cliques. It has been pointed out that Q suffers from the resolution limit in this type of networks (Fortunato and Barthelemy 2007). The detailed modularity optimization of Q and D on these networks has been discussed in Zhang et al. (2009). We first give a qualitative discussion on the ring of cliques shown in Fig. 1(a). In this case, any min-cut including two edges must be qualified. So we can divide the ring into two subnetworks satisfying the weak community definition in the first operation. Then for every subnetwork, the min-cut including only one edge must be qualified, and we can further divide it into smaller subnetworks and so on. Finally, we divide the network into a union of cliques. Since any min-cut of a clique is not qualified, so the clique cannot be further divided into smaller communities and the QMC algorithm stops with each clique as a community.

Then we generate a series of rings of cliques with different parameters to test whether our QMC method has the resolution limit in real computation. In the first series of rings of cliques, each ring has 50 6-cliques. The number of edges between two adjacent cliques (denoted by l_{bw}) varies from 1 to 30. Figure 2 shows the computational results on this series of rings of cliques by the QMC algorithm, optimization of

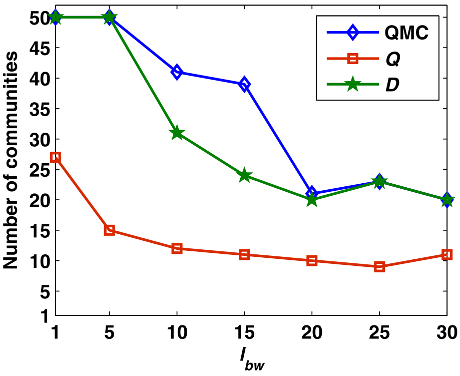

Fig. 2 Computational results on rings of cliques with different number of between-edges l_{bw}. Here the x-axis denotes the parameter l_{bw} and the y-axis denotes the number of identified communities

Q, and optimization of D, respectively. The result indicates that the number of communities detected by the three algorithms decreases when the number of between-edges l_{bw} increases. When $l_{bw} > 15$, the single cliques are not qualified communities anymore, so it is reasonable for the three algorithms to group several cliques into one community. However, when $l_{bw} < 15$, those single cliques remain qualified communities, but optimization of Q cannot recognize these cliques as communities. Especially, even when $l_{bw} = 1$, that is, the network consists of cliques connected by single edges, Q still cannot resolve the cliques as small communities, which is consistent with the fact that Q suffers from serious resolution limits. We also can see that the resolution limit of D is much slighter than Q. Our QMC algorithm can detect all cliques as communities when $l_{bw} < 10$. When $l_{bw} \geq 10$, our algorithm is able to detect more communities than both Q and D.

In the second series of rings of cliques, the number of between-edges l_{bw} takes 5 and 8 respectively, which ensure that all the cliques are qualified communities. The number of cliques (denoted by parameter L) varies from 3 to 21. The computational results by the three community detection methods are summarized in Fig. 3. When l_{bw} takes 5, the results are shown in Fig. 3(a), where the solution by optimization of D is identical to that of the QMC algorithm. We can see that both QMC and D can find all cliques correctly no matter how many cliques the networks have. However, the resolution limit of Q appears when the number of cliques L is larger than 7. From the results when l_{bw} takes 8, as shown in Fig. 3(b), both Q and D suffers from the resolution limit even when L is small. On the other hand, the QMC algorithm has much better performance than both Q and D.

5.2 Random ad-hoc networks

Ad-hoc networks are another type of widely used exemplar networks for theoretically analyzing modularity measures Q and D (Zhang et al. 2009). Once again we first give an analytic discussion on this type of networks. Assume that an ad hoc network consists of n lumps, each with m nodes. A lump is not necessary to be a clique but a dense graph that cannot be divided into two qualified communities. Let d_{in} be the degree of every vertex within a lump and d_{bw} be the number of edges between any pair of lumps.

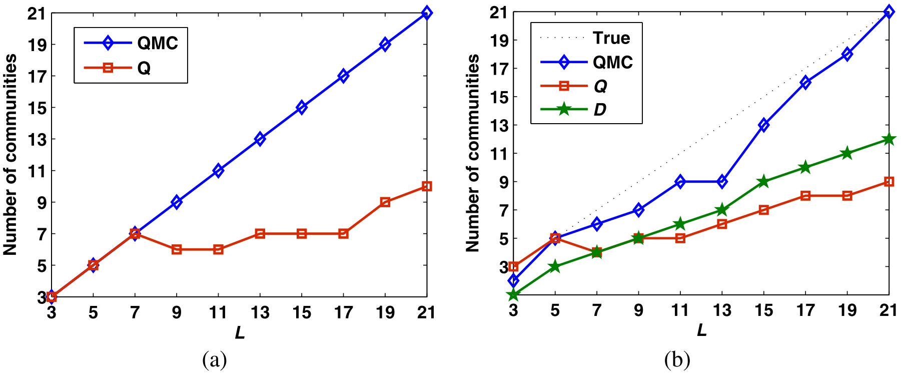

Fig. 3 Computational results on ring of cliques with the number of cliques L. Here the x-axis denotes the parameter L and the y-axis denotes the number of identified communities. (**a**) l_{bw} is fixed to be 5; (**b**) l_{bw} is fixed to be 8

Assume that there is a min-cut of the ad-hoc network, whose removal divides the network two subnetworks, with one subnetwork including k lumps. Then the other includes $n-k$ lumps, and the cut has $(n-k)kd_{bw}$ edges. The minimal value of $(n-k)kd_{bw}$ is $(n-1)d_{bw}$ which can be obtained when $k=1$ or $k=n-1$.

Case 1. The min-cut is qualified. In this case, each lump satisfies the weak community definition, i.e. $md_{in} > (n-1)d_{bw}$, then the ad-hoc network can be partitioned into two communities. One consists of a lump while the other consists of $n-1$ lumps. In the subsequent step, for the larger community of $n-1$ lumps, the min-cut includes $(n-2)d_{bw}$ edges, which is also qualified. So we can divide the larger community into two smaller communities: one is a single lump, and the other includes $n-2$ lumps. In the same way, we can divide the larger one step by step. Finally we obtain n communities.

Case 2. The min-cut is not qualified. Suppose that there exists a qualified min-cut (not the min-cut) which divides the ad-hoc network into two communities: one includes k lumps (where $k < n/2$), and the other includes $n-k$ lumps. The number of edges in the cut is $(n-k)kd_{bw}$, which increases when k becomes large. So the smaller community obtained by deleting the qualified min-cut must be a minimal community, i.e. k lumps compose a community satisfying the weak community definition but $k-1$ or less lumps cannot form a community satisfying the weak community definition. Suppose that there are qualified min-cuts in the ad-hoc network, then the QMC algorithm selects the new merged node as the start node in each operation of Step 2, so we can find the qualified min-cut and the ad-hoc network can be correctly partitioned.

Now we discuss the condition with which the qualified min-cut exists. The inside edges of k lumps (where $k < n/2$) is $kmd_{in} + k(k-1)d_{bw}$, the number of edges between the two parts is $k(n-k)d_{bw}$. The necessary condition for k lumps to be a community is $kmd_{in} + k(k-1)d_{bw} > k(n-k)d_{bw}$ which means $md_{in} > (n-2k+1)d_{bw}$.

Case 2.1. If n is odd, then $2k \leq n-1$, so $md_{in} > 2d_{bw}$.

Case 2.2. If n is even, then $2k \leq n$, so $md_{in} > d_{bw}$.

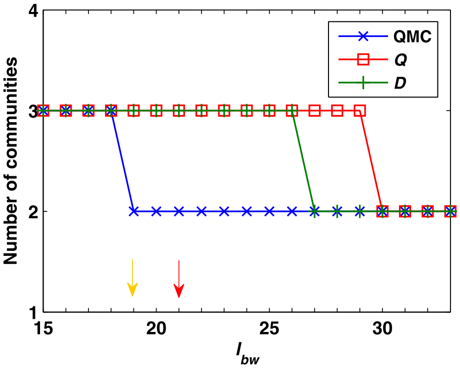

Fig. 4 (Color online) Computational results on uneven ad-hoc networks. The *yellow arrow* ($l_{bw} = 19$) denotes the critical value for K_5 to be a weak community, and the *red arrow* ($l_{bw} = 21$) denotes the critical value for K_5 to be a most weak community

In other words, when n is odd and $md_{in} \leq 2d_{bw}$, the ad-hoc network cannot be partitioned into two or more communities; When n is even and $md_{in} \leq d_{bw}$, the ad-hoc network cannot be partitioned into two or more communities. Therefore, the condition of no qualified min-cut existing in the ad-hoc network is as follows.

(1) When n is even, $md_{in} \leq d_{bw}$.
(2) When n is odd, $md_{in} \leq 2d_{bw}$.

In the following, we turn to do numerical experiments on some uneven ad-hoc networks, where there are three cliques in these networks, K_{30}, K_{20} and K_5. The number of edges between K_5 and either K_{30} or K_{20} is 1, while the number of edges between K_{30} and K_{20} (denoted by l_{bw}) changes from 15 to 33. The result, shown in Fig. 4, indicates that the QMC algorithm can partition the networks into three communities when $l_{bw} \leq 18$ and into two communities when $l_{bw} \geq 19$, with each community satisfying the weak community definition. Optimization of D detects three communities for $l_{bw} \geq 26$, and optimization of Q detects three communities for $l_{bw} \geq 29$. When $l_{bw} \geq 19$ (the yellow arrow in Fig. 4), the clique K_{20} is not a qualified community any more. When $l_{bw} \geq 21$ (the red arrow in Fig. 4), the clique K_{20} even does not satisfy the most weak definition.

5.3 The karate club network

Now we test our methods on real networks. The first example is the famous karate club network analyzed by Zachary (1977). It consists of 34 members of a karate club as nodes and 78 edges representing friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and the club's instructor, the club split into two small ones. In the past few years, many researchers have investigated this network and tried to partition it into several communities. Most researchers think it is reasonable to partition it into four communities, though the real partition is a two-community division. Our method partitions the network into four communities which cannot be divided further according to the weak definition after three division operations. The result is quite close to those obtained by other methods including optimization of Q ($Q = 0.420$) and D ($D = 7.85$) (Agarwal and Kempe 2008; Cafieri et al. 2010;

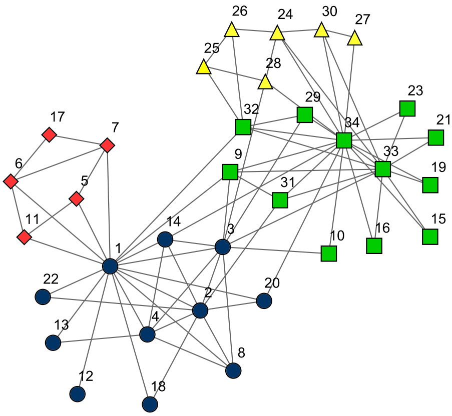

Fig. 5 The community structure in the karate club network detected by our method

Li et al. 2008; Girvan and Newman 2002; Newman and Girvan 2004). This partition is shown in Fig. 5. The first division occurs at a qualified min-cut of 4 edges, which partitions the network into two subnetworks, and one is an indivisible community $C_1 = \{5, 6, 7, 11, 17\}$. The larger subnetwork can be partitioned into two subnetworks by the second qualified min-cut of size 9, and one subnetwork is an indivisible community $C_2 = \{24, 25, 26, 27, 28, 30\}$. The other one can be further partitioned into the following two indivisible communities by the third qualified min-cut of size 10, $C_3 = \{1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22\}$, $C_4 = \{9, 10, 15, 16, 19, 21, 23, 29, 31, 32, 33, 34\}$. Although our result is not exactly identical to those obtained by other methods (Agarwal and Kempe 2008; Cafieri et al. 2010; Li et al. 2008; Girvan and Newman 2002; Newman and Girvan 2004), our partition has comparable number of edges between different communities to other results, which means that it is a reasonable result. In addition, merging two pairs of the communities, we can have a partition exactly corresponding to the real partition of the network, reflecting community structure at another scale.

5.4 The football team network

The United States college football team network has been widely used as a benchmark test example in network science due to its natural community structure (Li et al. 2008; Zhang et al. 2007; Wang et al. 2008). It is a network representation of the schedule of Division I games for the 2000 season: The nodes in the network represent the 115 teams, while the edges represent 613 games played in the course of the year. The teams are divided into 13 conferences containing around 8–12 teams each. Games are generally more frequent between members of the same conferences than between members of different conferences, with teams playing an average of about seven intraconference games and four interconference games in the 2000 season. Interconference play is not uniformly distributed; teams that are geographically close to one another but belong to different conferences are more likely to play one another than teams separated by large geographic distances. We apply our QMC algorithm to the football team network, and find that our algorithm partitions the network into 13 communities, which is shown in Fig. 6. The correct rate of our method is more

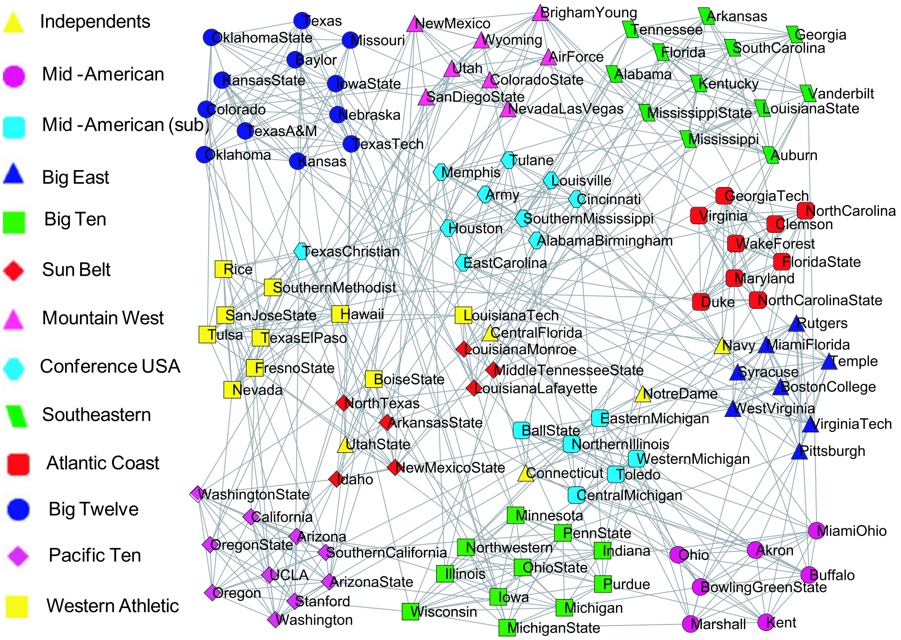

Fig. 6 (Color online) Computational results of QMC on the football team network. The nodes with the same shapes and colors are teams in the same conference, and the dense subgraphs in the layout are communities detected by the QMC algorithm

than 91%, which means that the detected community structure is in a high agreement with the true community structure. Optimization of Q on the football team network leads to a 10-community partition with $Q = 0.604$ and optimization of D leads to a 11-community partition with $D = 44.388$, which indicates that the resolution limit in Q and D prevents them from finding the 13-community partition. A widely used normalized mutual information index I_{NMI} (Danon et al. 2005) is used to evaluate if two partitions are similar or not. We found that $I_{\mathrm{NMI}}(P_{\mathrm{true}}, P_Q) = 0.878$, $I_{\mathrm{NMI}}(P_{\mathrm{true}}, P_D) = 0.897$, whereas $I_{\mathrm{NMI}}(P_{\mathrm{true}}, P_{\mathrm{QMC}}) = 0.928$.

5.5 The jazz musician network

The jazz musician network (Gleiser and Danon 2003) is a social network describing the collaboration among jazz bands. It consists of 198 bands that performed from 1912 to 1940 with 1,275 jazz musicians. Each band is represented by one node and two bands with at least one shared musician are linked by an edge. Due to the black/white racial segregation and the cities that bands recorded in, the network can be divided into three communities in reality. Both optimization of Q and optimization of D partition the jazz musician network into four communities (Zhang et al. 2009), with $Q = 0.445$ and $D = 52.84$. One of the communities identified by D has 22 nodes consisting of several connected components and dose not satisfy the weak community definition. A 4-node community identified by Q has 5 inner edges and 17, 30, 32 edges towards other three communities, respectively and violates the most

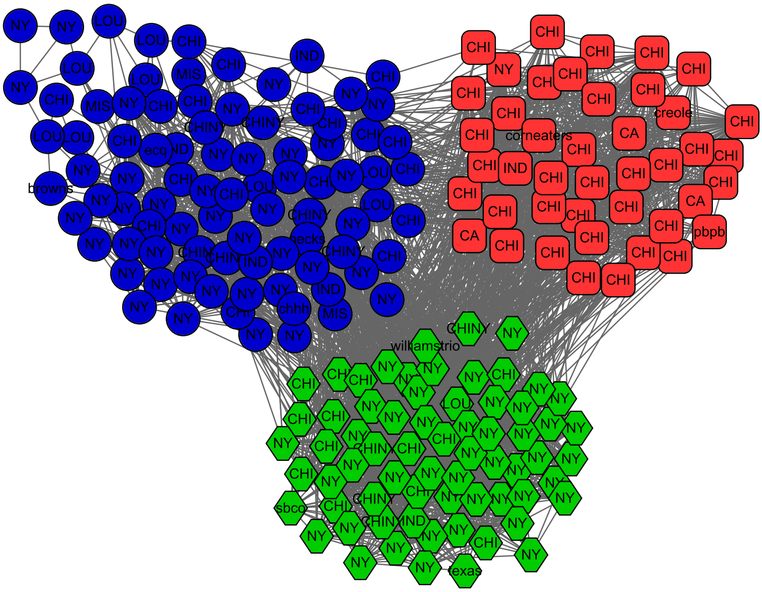

Fig. 7 The community structure in the jazz musician network detected by the QMC algorithm

weak community definition. By using our QMC algorithm, we can partition the jazz musician network into three communities shown in Fig. 7. All of them satisfy the weak community definition. The jazz bands in a same community tend to come from the same city and share more musicians: The first community includes 43 jazz bands, more than 80% coming from Chicago (CHI); The second community has 67 jazz bands, more than 70% coming from New York (NY); The third community includes 88 jazz bands, where 43 bands come from New York, 19 come from Chicago, the rest bands come from other states. Furthermore, more than 90% bands coming from Louisiana (LOU), 100% bands coming from Mississippi (MIS) and more than 67% bands coming from Indiana (IND) are in the third community.

6 Conclusion and discussion

Community structure is one of the main characteristics of complex networks and helps to understand the functions of these networks. In this paper, we propose a combinatorial model for the community identification problem. We formulate this model into a linear integer programming problem and prove its NP-completeness. We also give a qualified min-cut based bisecting algorithm. The extensive computational results demonstrate that our model and algorithm can overcome the resolution limit and misidentification problem existing in Q or D optimization for community detection.

The combinatorial model for the community identification problem closely depends on the community definition. In this paper we focus on the weak definition given in Radicchi et al. (2004). If some other definition is adopted, we can similarly build the corresponding combinatorial model and give related algorithms. In addition, the qualified min-cut problem formulated in this paper is an interesting combinatorial optimization problem. The designed bisecting algorithm is adopted from an algorithm for min-cut problems. Investigating more combinatorial properties and algorithms directly for the qualified min-cut problem is a promising research topic.

To enhance the ability of the proposed combinatorial model and the QMC algorithm in community detection, there are several potential extensions that are worth exploring. For example, a network may simultaneously exist multiple partitions that all satisfy the weak community definition. We can introduce a parameter into the constraint corresponding to the weak definition in the integer linear programming, to select a partition satisfying the weak definition that has more edges within communities compared to inter-community edges. In addition, the model and the QMC algorithm seek the largest number of communities, which may make the scale of community structure be too small or too large, depending on the network. This probably could be overcome by modifying the stop criterion of QMC. Actually the QMC algorithm is similar to hierarchical clustering: constantly divide a (sub)network into two parts. Although we only reported the subnetworks that are located into the bottom of the hierarchical tree, the process in which the QMC algorithm divides the network can reflect the community structure of the network at the whole mesoscale. Finally, it is easy to extend the combinatorial model and the QMC algorithm to cope with weighted networks. This can be achieved by modifying the constraints in the linear integer programming model and solving weighted min-cut problems in the QMC steps. For directed networks, currently there is no a consistent definition for community structure. We believe that such an extension is possible when a consistent community definition for directed networks is available.

Acknowledgements The authors are separately supported by the Natural Science Foundation of China under Grant No. 10631070, 60873205, 10801131, 10701080, the grant kjcx-yw-s7 from Chinese Academy of Sciences, Beijing Natural Science Foundation under Grant No. 1092011, Foundation of Beijing Education Commission under Grant No. SM200910037005, and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR(IHLB)).

References

Agarwal G, Kempe D (2008) Modularity-maximizing graph communities via mathematical programming. Eur Phys J B 66(3):409–418

Albert R, Barabási A (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97

Alimonti P, Kann V (2000) Some APX-completeness results for cubic graphs. Theor Comput Sci 237(1–2):123–134

Arenas A, Fernandez A, Gomez S (2008) Analysis of the structure of complex networks at different resolution levels. New J Phys 10:053039

Cafieri S, Hansen P, Liberti L (2010) Edge ratio and community structure in networks. Phys Rev E 81(2):026105

Danon L, Duch J, Diaz-Guilera A, Arenas A (2005) Comparing community structure identification. J Stat Mech, P09008

Everett M, Borgatti S (1998) Analyzing clique overlap. Connections 21(1):49–61

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Fortunato S, Barthelemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci USA 104(1):36–41

Girvan M, Newman M (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

Gleiser P, Danon L (2003) Community structure in jazz. Adv Complex Syst 6(4):565–573

Guimerà R, Amaral L (2005) Functional cartography of complex metabolic networks. Nature 433(7028):895–900

Hu Y, Chen H, Zhang P, Li M, Di Z Fan Y (2008) Comparative definition of community and corresponding identifying algorithm. Phys Rev E 78(2):026121

Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11:033015

Li Z, Zhang S, Wang RS, Zhang XS, Chen L (2008) Quantitative function for community detection. Phys Rev E 77(3):036109

Newman M (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582

Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818

Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci USA 101(9):2658–2663

Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Reichardt J, Bornholdt S (2004) Detecting fuzzy community structures in complex networks with a Potts model. Phys Rev Lett 93(21):218701

Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. Phys Rev E 80(1):016109

Rosvall M, Bergstrom C (2007) An information-theoretic framework for resolving community structure in complex networks. Proc Natl Acad Sci USA 104(18):7327–7331

Rosvall M, Bergstrom C (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123

Schuetz P, Caflisch A (2008) Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. Phys Rev E 78(17):026112

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905

Šíma J, Schaeffer S (2006) On the NP-completeness of some graph cluster measures. Lect Notes Comput Sci 3831:530–537

Stoer M, Wagner F (1997) A simple min-cut algorithm. J ACM 44(4):585–591

Wang RS, Zhang S, Wang Y, Zhang XS, Chen L (2008) Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. Neurocomputing 72:134–141

Zachary W (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452–473

Zhang S, Wang RS, Zhang XS (2007) Uncovering fuzzy community structure in complex networks. Phys Rev E 76(4):046103

Zhang XS, Wang RS, Wang Y, Wang J, Qiu Y, Wang L, Chen L (2009) Modularity optimization in community detection of complex networks. Europhys Lett 87(3):38002

Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21(9):1010–1024