Metrika (2020) 83:467–484
https://doi.org/10.1007/s00184-019-00740-7

Asymptotic distribution of modularity in networks

Yang Li[1] · Yongcheng Qi[1]

Received: 15 October 2018 / Published online: 26 August 2019

Abstract
The structure of complex networks is an important aspect in the study of the real network data. Quite often, it is desirable to know the division of the network into communities. A large number of community detection algorithms have been proposed to probe the community structure of complex networks. For a specific partition of a given network, we show that the distribution of modularity under a null hypothesis of free labeling is asymptotically normal when the size of the network gets large. The significance of the partition is defined based on this asymptotic distribution, which can help assess its goodness. Two different partitions can also be compared statistically. Simulation studies and real data analyses are performed for illustration.

Keywords Modularity · Asymptotic distribution · Network · Complex systems

1 Introduction

The analysis and exploration of network data have been performed within the modern science of complex systems for a long time (Jackson 2010; Albert and Barabási 2002; Newman 2003, 2010). The community structure, or clustering, plays an important role in the dynamics of the network. A community is a group of vertices in the network which is more tightly connected among themselves than with other vertices from outside of the group. Vertices in the same community in general share some common characteristics and interact more strongly than with vertices from other groups. Various methods already exist in finding the community structure within a complex network (Fortunato 2010). After the detection of communities, it is an important issue to assess their statistical significance. However, the literature on this crucial topic is not extensive. Rosvall and Bergstrom (2010) bootstrapped the original network to get

✉ Yang Li
yangli@d.umn.edu

Yongcheng Qi
yqi@d.umn.edu

[1] Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN 55812, USA

new samples. For each cluster in the original partition, they used the bootstrap samples to find the largest subset of vertices that are most likely to be classified in the same cluster. Lancichinetti et al. (2010) estimated the significance of single communities by comparing the original network with a random graph with similar properties. Zhang and Chen (2016) proposed a framework in which the consistency of modularity can be shown under a degree-corrected stochastic block model.

We propose a way of evaluating the significance of any given partition by considering whether this particular partition can arise simply from randomness under the assumption that there is no underlying community structure in the network. We first derive the asymptotic distribution of the modularity under that assumption when the network size gets large and further define its significance using the z-score. For small networks, a more robust approach which does not depend on the asymptotic approximation is the randomization test.

The paper is organized as follows. In Sect. 2, we introduce some basic concepts in the network along with the definition of modularity. We further describe the global null hypothesis called free labeling. Under this null hypothesis, we derive the asymptotic distribution of modularity. In Sect. 3 we perform a simulation study to validate the asymptotic behavior and further use some well-known real network data for illustration.

2 Main results

Consider an undirected graph G consisting of n vertices $\{v_1, v_2, \dots, v_n\}$ and m edges $\{e_1, e_2, \dots, e_m\}$. The degree of vertex v_i, denoted by k_i, is the number of edges connected to it. It is easy to see that $\sum_{i=1}^{n} k_i = 2m$. Let $\mathbf{A}$ be the symmetric adjacency matrix of the network in which A_{ij} denotes the number of edges between vertices i and j. For a simple graph which is the case discussed in this paper, A_{ij} is 0 or 1, and $A_{ii} = 0$ since no self-loops are allowed. We have $k_i = \sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{ji}$ for $1 \le i \le n$. Suppose that we already have a partition of the network G, denoted by C, using one of the existing community detection methods (Fortunato 2010). In other words, each vertex v_i ($1 \le i \le n$) is associated with a group label or color $c_i \in \{1, 2, \dots, K\}$ where K is the total number of communities by the partition, and we write $C = (c_1, c_2, \dots, c_n)$.

The modularity of the partition C is defined as (Newman and Girvan 2004; Newman 2006)

$$Q_n(C) = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{c_i,c_j} = \frac{1}{2m} \sum_{i,j} B_{ij} \delta_{c_i,c_j}, \tag{1}$$

where δ_{c_i,c_j} is the Kronecker delta function which takes value 1 if vertices i and j are in the same group (i.e., $c_i = c_j$), and zero otherwise; $B_{ij} = A_{ij} - k_i k_j/(2m)$ is the modularity matrix, and $Q_n(C)$ is the weighted sum of B_{ij} over all pairs of vertices i, j that fall in the same groups. It measures the extent to which vertices of the same type are connected to each other in a network (Newman 2010). Here we explicitly use a subscript n in $Q_n(C)$ to emphasize its dependence on the size of the network. Values of $Q_n(C)$ are strictly less than 1. Large values indicate that there are more edges between vertices of the same type than we would expect by chance only, while small values indicate that vertices from different groups are more likely to be connected. In general, modularity-based algorithms to divide the network try to maximize $Q_n(C)$ either directly or indirectly (Fortunato 2010).

For a given partition C of the network, we are interested in whether this partition could be obtained by randomly assigning colors to the vertices. The global null hypothesis H_0 is that the colors are assigned to vertices randomly, regardless of the structure of the network. Under this hypothesis, each vertex can be grouped into one and only one of the possible K groups. The probability that a vertex is labeled by color k is equal to the proportion of color k in the partition C. In other works, the probability that a given vertex is labeled as group 1 is $p_1 = |\text{Col}(1)|/n$ where $|\text{Col}(1)|$ is the cardinality of the set of vertices with color 1; the probability is $p_2 = |\text{Col}(2)|/n$ for group 2, and so on. Of course, $p_k \geq 0$ and $p_1 + p_2 + \cdots + p_K = 1$. The labeling of different vertices is assumed to be independent so H_0 is also called *free labeling*. The numbers of vertices labeled with K different colors, $\mathbf{n} = (n_1, n_2, \ldots, n_K)$, follow a multinomial distribution, $\mathbf{n} \sim \text{Multinomial}(n, p_1, p_2, \ldots, p_K)$. In doing this, the group sizes in C are treated as correct either they are specified *a priori* or they are chosen for the sake of convenience, e.g., dividing a network into two with equal size. This approach is similar to the statistical tests for spatial autocorrelation for areal data in spatial statistics (Cliff and Ord 1981), where the statistical significance of the observed Moran's (1950) I and Geary's (1954) C can be assessed through their asymptotic sampling distributions.

To test the null hypothesis of *free labeling* for a given or observed partition C, we need to randomize the modularity $Q_n(C)$ given in (1) by assuming that the partition $C = (c_1, c_2, \ldots, c_n)$ is a random vector, and $c_1, c_2, \ldots, c_n$ are independent and identically distributed (iid) random variables with distribution $P(c_i = j) = p_j$ for $1 \leq j \leq K$. In this case, we denote $Q_n(C)$ by Q_n to avoid confusion. We need to investigate the distribution of Q_n for testing the hypothesis of free labeling.

To get the asymptotic normality of Q_n, we will impose the following conditions on the degree sequence $k_1, \ldots, k_n$:

$$\lim_{n\to\infty} \frac{1}{m^3} \left(\sum_{i=1}^{n} k_i^2 \right)^2 = 0, \tag{2}$$

$$\lim_{n\to\infty} \frac{1}{m^2} \sum_{1\leq i,j\leq n} \left(\sum_{\ell=1}^{n} A_{i\ell} A_{j\ell} \right)^2 = 0. \tag{3}$$

Note that the probabilities $p_1, \ldots, p_K$ depend on n implicitly. We assume

$$p_{(2)} + p_{(2)}^2 - 2p_{(3)} \geq \delta > 0 \text{ where } \delta \text{ doesn't depend on } n, \text{ and } p_{(m)} = \sum_{k=1}^{K} p_k^m. \tag{4}$$

Condition (4) controls the asymptotic variance of Q_n so that Q_n is not asymptotically degenerate.

Remark 1 A sufficient condition for (3) is

$$\lim_{n\to\infty} \frac{\max_{1\le j\le n} k_j \sum_{i=1}^n k_i^2}{m^2} = 0.$$

See Lemma 4. Since

$$\frac{\max_{1\le j\le n} k_j \sum_{i=1}^n k_i^2}{m^2} \le \frac{(\max_{1\le j\le n} k_j)^2 \sum_{i=1}^n k_i}{m^2} = \frac{2(\max_{1\le j\le n} k_j)^2}{m},$$

if

$$\lim_{n\to\infty} \frac{(\max_{1\le j\le n} k_j)^2}{m} = 0, \tag{5}$$

then (3) holds. In fact, (5) is also sufficient for (2) since

$$\frac{1}{m^3}\left(\sum_{i=1}^n k_i^2\right)^2 \le \frac{1}{m^3}\left(\max_{1\le j\le n} k_j \sum_{i=1}^n k_i\right)^2 = \frac{4(\max_{1\le j\le n} k_j)^2}{m} \to 0$$

under (5). An example for (5) to hold is to assume that there exists a sequence ℓ_n such that $d_1\ell_n \le k_i \le d_2\ell_n$ for all $1 \le i \le n$, where $d_1 > 0$ and $d_2 > 0$ are two constants. Then (5) holds if and only if $\ell_n/n \to 0$ as $n \to \infty$. In this case, both (2) and (3) are satisfied. For instance, for a Poisson random network with a finite average degree λ, $\frac{1}{m^3}\left(\sum_{i=1}^n k_i^2\right)^2 \sim \lambda/n \to 0$ and $\frac{(\max_{1\le j\le n} k_j)^2}{m} = \mathcal{O}((\log n)^2/n) \to 0$; see Riordan and Selby (2000).

Remark 2 Define $p_{\max} = \max_{1\le k\le K} p_k$. Assume $K \le K_0$, where K_0 is an integer that does not depend on n. Then $p_{\max} \ge 1/K_0$, and $p_{(2)} \ge p_{\max}\sum_{k=1}^K p_k = p_{\max} \ge 1/K_0$. We have

$$\begin{aligned} p_{(2)} + p_{(2)}^2 - 2p_{(3)} &\ge p_{(2)}\left(1 + p_{(2)} - 2p_{\max}\right) \ge p_{(2)}\left(1 + p_{\max}^2 - 2p_{\max}\right) \\ &= p_{(2)}(1 - p_{\max})^2 \ge (1 - p_{\max})^2/K_0, \end{aligned}$$

which implies (4) if $p_{\max} \le \delta_1$ for some $\delta_1 < 1$. In other words, the distribution of the colors in the partition should be more or less homogeneous.

Theorem 1 *Under conditions (2), (3) and (4) we have*

$$\frac{Q_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0, 1), \tag{6}$$

where μ_n and σ_n^2 are given by

$$\mu_n := E[Q_n] = -\frac{1-p_{(2)}}{4m^2}\sum_{i=1}^{n} k_i^2, \tag{7}$$

$$\sigma_n^2 := Var(Q_n) = \frac{p_{(2)} + p_{(2)}^2 - 2p_{(3)}}{2m^2} \sum_{1\le i\ne j\le n} B_{ij}^2 + \frac{p_{(3)} - p_{(2)}^2}{m^2}\sum_{i=1}^{n} B_{ii}^2. \tag{8}$$

Remark 3 The significance of a given partition C of a graph G can be defined from the z-score

$$z(C) = \frac{Q_n(C) - \mu_n}{\sigma_n}, \tag{9}$$

which requires a large size n. If $z(C) > z_{1-\alpha}$ where $z_{1-\alpha}$ is the percentile of the standard normal distribution, the partition C is statistically significant at the level of $1-\alpha$. In other words, H_0 should be rejected. On the other hand, if $z(C) < z_{1-\alpha}$, we cannot reject H_0 and the partition C is not statistically significant. It is also possible to use the p value, $1 - \Phi(z(C))$, to evaluate the significance of partition C, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

For small graphs, a more appropriate approach would be the randomization test where realizations of simulated graphs are generated from the null hypothesis of free labeling. The modularity of the real graph then can be compared to the modularities of the simulated graphs and a p value can be obtained from its rank which then could be utilized to evaluate the significance of the partition. An example can be found in the next section.

Remark 4 In this paper, we consider the simple graph with $A_{ij} \in \{0, 1\}$. In fact, if we allow $A_{ij} = A_{ji} \in \{0, 1, \ldots, r-1\}$ for some integer $r \ge 2$ when $i \ne j$, $A_{ii} = 0$, and set $k_i = \sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{ji}$ for $1 \le i \le n$, then the conclusion in Theorem 1 is also true, and the proof is essentially the same as the proof for Theorem 1. Theorem 1 is the special case with $r = 2$. When $r > 2$, $A_{ij}^2 = A_{ij}$ is no longer true in general. But we have the following estimation

$$m \le \sum_{1\le i<j\le n} A_{ij} \le \sum_{1\le i<j\le n} A_{ij}^2 \le (r-1)\sum_{1\le i<j\le n} A_{ij} \le (r-1)m.$$

The only difference for the proof of this general case is in Lemmas 6 and 5. Lemma 6 is valid if we redefine τ_n^2 by letting $\tau_n^2 = (p_{(2)} + p_{(2)}^2 - 2p_{(3)})\sum_{1\le i<j\le n} A_{ij}^2$. Lemma 5 is true if m in (14) is replaced by $m^2/\sum_{1\le i<j\le n} A_{ij}^2$, i.e. we can show that

$$\sigma_n^2 = \frac{p_{(2)} + p_{(2)}^2 - 2p_{(3)}}{m^2}\sum_{1\le i<j\le n} A_{ij}^2(1+o(1)) = \frac{\tau_n^2}{m^2}(1+o(1)).$$

Besides, all discussions in Remarks 1–3 are still valid.

Our proof of Theorem 1 relies on construction of an array of martingale differences and application of a martingale central limit theorem. We will first prove some preliminary lemmas.

Lemma 2 *Let $X_1, \ldots, X_m$ be iid random variables with $E(X_1) = 0$ and $E(X_1^4) < \infty$ and $\{a_1, \ldots, a_m\}$ be constants. Then*

$$E\left(\sum_{j=1}^{m} a_j X_j\right)^4 \le 3\left(\sum_{1\le i\le m} a_i^2\right)^2 E(X_1^4). \tag{10}$$

Proof For any $1 \le i, j, k, \ell \le m$, we have $E(X_i X_j X_k X_\ell) \ne 0$ if and only if $i = j = k = \ell$ or $\{i, j, k, \ell\}$ form two distinct pairs of integers. Therefore, we get

$$\begin{aligned}
E\left(\sum_{j=1}^{m} a_j X_j\right)^4 &= E\left(\sum_{1\le i,j,k,\ell\le m} a_i a_j a_k a_\ell X_i X_j X_k X_\ell\right)\\
&= \sum_{1\le i,j,k,\ell\le m} a_i a_j a_k a_\ell E(X_i X_j X_k X_\ell)\\
&= \sum_{1\le i\le m} a_i^4 E(X_1^4) + 3\sum_{1\le i\ne k\le m} a_i^2 a_k^2 (E(X_1^2))^2\\
&\le \left(\sum_{1\le i\le m} a_i^4 + 3\sum_{1\le i\ne k\le m} a_i^2 a_k^2\right) E(X_1^4)\\
&\le 3\left(\sum_{1\le i\le m} a_i^2\right)^2 E(X_1^4),
\end{aligned}$$

proving (10). □

We will first define the errors of the Hájek-type projections for U-statistics in the following Lemma 3, and then we establish a martingale central limit theorem for the weighted projection errors in Lemma 6.

Lemma 3 *Let $X_1, \ldots, X_m$ be iid random variables, and $h(x, y)$ and $g(x)$ be two measurable functions defined in $\mathbb{R}^2$ and $\mathbb{R}$ such that $h(x, y) = h(y, x)$, $E(h^2(X_1, X_2)) < \infty$ and $E(g^2(X_1)) < \infty$. Define $h_1(x) = E(h(x, X_1))$, $\mu_h = E(h(X_1, X_2))$, $\mu_g = E(g(X_1))$, and set $\bar{h}(x, y) = h(x, y) - h_1(x) - h_1(y) + \mu_h$. Then $\{\bar{h}(X_i, X_j) : 1 \le i < j \le n\} \cup \{g(X_\ell) - \mu_g : 1 \le \ell \le n\}$ are orthogonal, i.e., the covariance for any two distinct random variables in the set is zero.*

Proof Note that all the random variables in $\{\bar{h}(X_i, X_j) : 1 \le i < j \le n\} \cup \{g(X_\ell) - \mu_g : 1 \le \ell \le n\}$ have a mean zero. It suffices to show the following equations:

$$E(\bar{h}(X_i, X_j)\bar{h}(X_{i'}, X_{j'})) = 0, \quad \text{if } (i, j) \ne (i', j'), \tag{11}$$

$$E(\bar{h}(X_i, X_j)(g(X_\ell) - \mu_g)) = 0, \tag{12}$$

$$E((g(X_\ell) - \mu_g)(g(X_{\ell'}) - \mu_g)) = 0, \quad \text{if } \ell \neq \ell'. \tag{13}$$

In Eqs. (11) and (12), $i < j$ and $i' < j'$.

One can easily verify that $E(\bar{h}(X_i, X_j)|X_i) = E(\bar{h}(X_i, X_j)|X_j) = 0$. If $\{i, j\} \cap \{i', j'\} = \emptyset$, an empty set, then (11) follows from the independence of $\bar{h}(X_i, X_j)$ and $\bar{h}(X_{i'}, X_{j'})$. If $\{i, j\} \cap \{i', j'\} = \{i\}$, then conditional on X_i, $\bar{h}(X_i, X_j)$ and $\bar{h}(X_{i'}, X_{j'})$ are independent with mean zero, and thus (11) follows. If $\{i, j\} \cap \{i', j'\} = \{j\}$, (11) can be verified similarly. (12) can be verified by using conditional independence or independence according to whether $\ell \in \{i, j\}$ or not. (13) is trivial. □

The following lemma is needed for the first statement in Remark 1.

Lemma 4 *Define* $o_{ij} = \sum_{\ell=1}^{n} A_{i\ell} A_{j\ell}$ *for* $1 \leq i, j \leq n$. *Then*

$$\sum_{1 \leq i, j \leq n} o_{ij}^2 \leq \max_{1 \leq j \leq n} k_j \sum_{i=1}^{n} k_i^2$$

Proof First, note that $o_{ij} \leq \min(k_i, k_j)$ for $1 \leq i, j \leq n$. Then

$$\begin{aligned} \sum_{1 \leq i, j \leq n} o_{ij}^2 &\leq \sum_{1 \leq i, j \leq n} k_i o_{ij} \\ &= \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n} k_i \sum_{1 \leq \ell \leq n} A_{i\ell} A_{j\ell} \\ &= \sum_{1 \leq i \leq n} \sum_{1 \leq \ell \leq n} k_i A_{i\ell} \sum_{1 \leq j \leq n} A_{j\ell} \\ &\leq \max_{1 \leq j \leq n} k_j \sum_{1 \leq i \leq n} k_i \sum_{1 \leq \ell \leq n} A_{i\ell} \\ &= \max_{1 \leq j \leq n} k_j \sum_{i=1}^{n} k_i^2, \end{aligned}$$

proving the lemma. □

Lemma 5 *The mean and variance of* Q_n *are given by* (7) *and* (8). *Furthermore, if* (2) *holds, then*

$$\sigma_n^2 = \frac{p_{(2)} + p_{(2)}^2 - 2p_{(3)}}{m}(1 + o(1)) \text{ as } n \to \infty. \tag{14}$$

Proof Under free labeling, $c_1, \ldots, c_n$ are iid random variables with $P(c_1 = k) = p_k$ for $1 \leq k \leq K$.

Define $h(x, y) = \delta_{x,y}$ in Lemma 3. Then $h_1(x) = E(\delta_{x,c_1}) = p_x$ for $1 \leq x \leq K$, and $\mu_h = \sum_{k=1}^{K} p_k^2 = p_{(2)}$. Now we have

$$\bar{h}(c_i, c_j) = \delta_{c_i,c_j} - p_{c_i} - p_{c_j} + p_{(2)}, \quad 1 \leq i \neq j \leq n.$$

Then we have

$$\begin{aligned}
2mQ_n &= \sum_{i=1}^{n} B_{ii} + \sum_{1\le i\ne j\le n} B_{ij}\bar{h}(c_i, c_j) \\
&\quad + \sum_{1\le i\ne j\le n} B_{ij}(p_{c_i} + p_{c_j}) - p_{(2)} \sum_{1\le i\ne j\le n} B_{ij} \\
&= (1 + p_{(2)}) \sum_{i=1}^{n} B_{ii} + 2 \sum_{1\le i<j\le n} B_{ij}\bar{h}(c_i, c_j) - 2\sum_{i=1}^{n} B_{ii} p_{c_i} \\
&= (1 - p_{(2)}) \sum_{i=1}^{n} B_{ii} + 2 \sum_{1\le i<j\le n} B_{ij}\bar{h}(c_i, c_j) \\
&\quad - 2\sum_{i=1}^{n} B_{ii}(p_{c_i} - p_{(2)}). \qquad (15)
\end{aligned}$$

Here we use the identity $\sum_{i=1}^{n} B_{ij} = \sum_{j=1}^{n} B_{ij} = 0$ and $\sum_{1\le i\ne j\le n} B_{ij} = \sum_{1\le i,j\le n} B_{ij} - \sum_{i=1}^{n} B_{ii} = -\sum_{i=1}^{n} B_{ii}$. One can also verify that

$$\mathrm{Var}(p_{c_1} - p_{(2)}) = p_{(3)} - p_{(2)}^2$$

and

$$\mathrm{Var}(\bar{h}(c_1, c_2)) = p_{(2)} + p_{(2)}^2 - 2p_{(3)}. \qquad (16)$$

It follows from Lemma 3 that $\{\bar{h}(c_i, c_j) : 1 \le i < j \le n\} \cup \{p_{c_\ell} - p_{(2)} : 1 \le \ell \le n\}$ are orthogonal with mean zero. Therefore, we have from (15) that

$$2mE(Q_n) = (1 - p_{(2)}) \sum_{i=1}^{n} B_{ii} = -\frac{1 - p_{(2)}}{2m} \sum_{i=1}^{n} k_i^2$$

and

$$\begin{aligned}
(2m)^2\mathrm{Var}(Q_n) &= 4\mathrm{Var}(\bar{h}(c_1, c_2)) \sum_{1\le i<j\le n} B_{ij}^2 + 4\mathrm{Var}(p_{c_1} - p_{(2)}) \sum_{i=1}^{n} B_{ii}^2 \\
&= 2(p_{(2)} + p_{(2)}^2 - 2p_{(3)}) \sum_{1\le i\ne j\le n} B_{ij}^2 + 4(p_{(3)} - p_{(2)}^2) \sum_{i=1}^{n} B_{ii}^2, \qquad (17)
\end{aligned}$$

which yield (7) and (8), respectively.

Note that

$$\begin{aligned}\sum_{1\le i\ne j\le n} B_{ij}^2 &= \sum_{1\le i\ne j\le n}\left(A_{ij}^2+\frac{k_i^2k_j^2}{4m^2}-2A_{ij}\frac{k_ik_j}{2m}\right)\\ &= \sum_{1\le i\ne j\le n} A_{ij}+\sum_{1\le i\ne j\le n}\frac{k_i^2k_j^2}{4m^2}-\sum_{1\le i\ne j\le n}A_{ij}\frac{k_ik_j}{m}\\ &= 2m+\sum_{1\le i\ne j\le n}\frac{k_i^2k_j^2}{4m^2}-\sum_{1\le i\ne j\le n}A_{ij}\frac{k_ik_j}{m}.\end{aligned}$$

Then by using the Cauchy-Schwarz inequality we get

$$\begin{aligned}\frac{1}{m}\left|\sum_{1\le i\ne j\le n}B_{ij}^2-2m\right| &\le \frac{1}{m}\sum_{1\le i\ne j\le n}\frac{k_i^2k_j^2}{4m^2}+\frac{1}{m^2}\sum_{1\le i\ne j\le n}A_{ij}k_ik_j\\ &\le \frac{\left(\sum_{i=1}^n k_i^2\right)^2}{4m^3}+\frac{1}{m^2}\sqrt{\sum_{1\le i\ne j\le n}A_{ij}^2}\sqrt{\sum_{1\le i\ne j\le n}k_i^2k_j^2}\\ &\le \frac{\left(\sum_{i=1}^n k_i^2\right)^2}{4m^3}+\sqrt{\frac{2\left(\sum_{i=1}^n k_i^2\right)^2}{m^3}}\\ &\to 0 \quad \text{as } n\to\infty\end{aligned}$$

by virtue of (2), and

$$\frac{1}{m}\sum_{i=1}^n B_{ii}^2 \le \frac{1}{m}\sum_{1\le i\le n}\frac{k_i^4}{4m^2}\le\frac{\left(\sum_{i=1}^n k_i^2\right)^2}{4m^3}\to 0.$$

(14) follows directly from (17) and the above estimation. □

Lemma 6 *If* (2) *and* (4) *hold, then*

$$T_n := \frac{1}{\tau_n}\sum_{1\le i<j\le n}A_{ij}\bar{h}(c_i,c_j)\xrightarrow{d}N(0,1), \tag{18}$$

where $\tau_n^2 = m(p_{(2)}+p_{(2)}^2-2p_{(3)})$.

Proof Write T_n on the left-hand side of (18) as

$$T_n = \frac{1}{\tau_n}\sum_{j=1}^n\sum_{i=1}^{j-1}A_{ij}\bar{h}(c_i,c_j)=\frac{1}{\tau_n}\sum_{j=1}^n z_{nj},$$

where $z_{nj}=\sum_{i=1}^{j-1}A_{ij}\bar{h}(c_i,c_j)$ for $2\le j\le n$.

Let $\mathcal{F}_j = \sigma(c_1, c_2, \ldots, c_j)$ denote the σ-algebra generated by $\{c_1, c_2, \ldots, c_j\}$ for $1 \le j \le n$. Since $E(\bar{h}(c_i, c_j)|\mathcal{F}_{j-1}) = E(\bar{h}(c_i, c_j)|c_i) = 0$ for any $1 \le i < j \le n$ we have

$$E(z_{nj}|\mathcal{F}_{j-1}) = 0 \quad \text{for } 2 \le j \le n.$$

Therefore, for each $n \ge 2$, $\{z_{nj},\ \ j = 2, \ldots, n\}$ forms a martingale difference with respect to $\{\mathcal{F}_j\}$.

Let $2 \le j \le n$. Note that for any $1 \le i_1, i_2 \le j - 1$

$$\begin{aligned}
&E(\bar{h}(c_{i_1}, c_j)\bar{h}(c_{i_2}, c_j)|\mathcal{F}_{j-1}) \\
&\quad = \sum_{k=1}^{K} \bar{h}(c_{i_1}, k)\bar{h}(c_{i_2}, k)p_k \\
&\quad = \delta_{c_{i_1}, c_{i_2}} \frac{p_{c_{i_1}} + p_{c_{i_2}}}{2} - p_{c_{i_1}}^2 - p_{c_{i_2}}^2 + p_{(3)} - (p_{c_{i_1}} - p_{(2)})(p_{c_{i_2}} - p_{(2)}) \\
&\quad = f(c_{i_1}, c_{i_2}),
\end{aligned}$$

where $f(x, y) = \delta_{x,y}\frac{p_x+p_y}{2} - p_x^2 - p_y^2 + p_{(3)} - (p_x - p_{(2)})(p_y - p_{(2)})$ for $1 \le x, y \le K$.

By using the orthogonality of $\{\bar{h}(c_i, c_j) : 1 \le i < j \le n\} \cup \{p_{c_\ell} - p_{(2)} : 1 \le \ell \le n\}$ we have from (16) that

$$E(f(c_1, c_2)) = E(E(\bar{h}(c_1, c_3)\bar{h}(c_2, c_3)|\mathcal{F}_2)) = E(\bar{h}(c_1, c_3)\bar{h}(c_2, c_3)) = 0 \quad (19)$$

and

$$E(f(c_1, c_1)) = E(E(\bar{h}(c_1, c_3)^2|\mathcal{F}_2)) = E(\bar{h}(c_1, c_3)^2) = p_{(2)} + p_{(2)}^2 - 2p_{(3)}. (20)$$

Define

$$s_n^2 = \sum_{j=2}^{n} E(z_{nj}^2|\mathcal{F}_{j-1}).$$

We have

$$\begin{aligned}
s_n^2 &= \sum_{j=2}^{n} E\left(\left(\sum_{i=1}^{j-1} A_{ij}\bar{h}(c_i, c_j)\right)^2 \middle| \mathcal{F}_{j-1}\right) \\
&= \sum_{j=2}^{n} E\left(\sum_{i_1=1}^{j-1} A_{i_1j}\bar{h}(c_{i_1}, c_j) \sum_{i_2=1}^{j-1} A_{i_2j}\bar{h}(c_{i_2}, c_j)) \middle| \mathcal{F}_{j-1}\right) \\
&= \sum_{j=2}^{n} E\left(\sum_{i_1=1}^{j-1}\sum_{i_2=1}^{j-1} A_{i_1j}Ai2j\bar{h}(c_{i_1}, c_j)\bar{h}(c_{i_2}, c_j) \middle| \mathcal{F}_{j-1}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=2}^{n}\sum_{i_1=1}^{j-1}\sum_{i_2=1}^{j-1} A_{i_1 j}A_{i_2 j}E(\bar{h}(c_{i_1}, c_j)\bar{h}(c_{i_2}, c_j)|\mathcal{F}_{j-1}) \\
&= \sum_{j=2}^{n}\sum_{i_1=1}^{j-1}\sum_{i_2=1}^{j-1} A_{i_1 j}A_{i_2 j} f(c_{i_1}, c_{i_2}) \\
&= 2\sum_{j=3}^{n}\sum_{1\le i_1<i_2\le j-1} A_{i_1 j}A_{i_2 j} f(c_{i_1}, c_{i_2}) + \sum_{j=2}^{n}\sum_{i=1}^{j-1} A_{ij} f(c_i, c_i) \\
&= 2\sum_{1\le i_1<i_2\le n-1}\left(\sum_{j=i_2+1}^{n} A_{i_1 j}A_{i_2 j}\right) f(c_{i_1}, c_{i_2}) + \sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} A_{ij}\right) f(c_i, c_i).
\end{aligned} \tag{21}
$$

Then it follows from (19) and (20) that

$$
\begin{aligned}
E(s_n^2) &= \sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} A_{ij}\right) E(f(c_i, c_i)) \\
&= (p_{(2)} + p_{(2)}^2 - 2p_{(3)})\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} A_{ij} \\
&= \frac{1}{2}(p_{(2)} + p_{(2)}^2 - 2p_{(3)})\sum_{1\le i<j\le n} A_{ij} \\
&= m(p_{(2)} + p_{(2)}^2 - 2p_{(3)}) \\
&= \tau_n^2.
\end{aligned}
$$

To prove (18), i.e.,

$$
\frac{1}{\tau_n}\sum_{j=1}^{n} z_{nj} \xrightarrow{d} N(0, 1) \tag{22}
$$

we will employ a martingale central limit theorem. In view of Corollary 3.1 in Hall and Heyde (1980), the martingale central limit theorem (22) holds if the following two conditions hold:

$$
\frac{1}{\tau_n^2}\sum_{j=2}^{n} E\left(z_{nj}^2 I(|z_{nj}| \ge \varepsilon\tau_n)|\mathcal{F}_{j-1}\right) \to 0 \quad \text{in probability} \tag{23}
$$

for every $\varepsilon > 0$, and

$$\frac{1}{\tau_n^2}\sum_{j=2}^{n} E\left(z_{nj}^2|\mathcal{F}_{j-1}\right) \to 1 \quad \text{in probability.} \tag{24}$$

In fact, it is easy to show (23) and (24) if we can verify the following conditions

$$\sum_{j=2}^{n} E\left(z_{nj}^4\right) = o\left(\tau_n^4\right) \quad \text{as } n \to \infty \tag{25}$$

and

$$E\left(s_n^2 - \tau_n^2\right)^2 = o(\tau_n^4) \quad \text{as } n \to \infty. \tag{26}$$

Note that $|\bar{h}(c_i, c_j)| \le 2$. Conditional on c_j, $\{\bar{h}(c_i, c_j), 1 \le i \le j-1\}$ are iid with conditional mean zero for $2 \le j \le n$. By using Lemma 2, we have

$$\begin{aligned} E(z_{nj}^4) &= E\left(E(z_{nj}^4|c_j)\right) \le E\left(3\left(\sum_{i=1}^{j-1} A_{ij}^2\right)^2 E(\bar{h}(c_1, c_j)^4|c_j)\right) \\ &\le 48\left(\sum_{i=1}^{j-1} A_{ij}\right)^2 \le 48k_j^2, \end{aligned}$$

which together with conditions (4) and (2) yields that

$$\sum_{j=2}^{n} E\left(z_{nj}^4\right) \le 48\sum_{j=2}^{n} k_j^2 = o\left(m^{3/2}\right) = o\left(\tau_n^4\right),$$

proving (25).

To show (24), we will use Lemma 3. On can verify that $E(f(x, c_1)) = 0$, $f(x, y) = f(y, x)$ for $1 \le x, y \le K$, and $E(f(c_1, c_2)) = 0$. Thus, by Lemma 3, $\{f(c_i, c_j) : 1 \le i < j \le n, f(c_\ell, c_\ell) : 1 \le \ell \le n\}$ are orthogonal. Also notice that $|f(x, y)| \le 3$ so that $\mathrm{Var}(f(c_i, c_j)) \le 9$ for $1 \le i, j \le n$. Therefore, from (21) we have

$$\begin{aligned} E\left(s_n^2 - \tau_n^2\right)^2 &= 4\sum_{1\le i_1<i_2\le n-1}\left(\sum_{j=i_2+1}^{n} A_{i_1 j}A_{i_2 j}\right)^2 \mathrm{Var}(f(c_{i_1}, c_{i_2})) \\ &\quad + \sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} A_{ij}\right)^2 \mathrm{Var}(f(c_i, c_i)) \\ &\le 36\left(\sum_{1\le i_1<i_2\le n-1}\left(\sum_{j=i_2+1}^{n} A_{i_1 j}A_{i_2 j}\right)^2 + \sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} A_{ij}\right)^2\right) \end{aligned}$$

$$\begin{aligned}
&\le 36 \sum_{1\le i_1,i_2\le n} \left(\sum_{j=1}^{n} A_{i_1 j} A_{i_2 j}\right)^2 \\
&= o(m^2) \\
&= o(\tau_n^4)
\end{aligned}$$

from (2) and (14), proving (26). □

Proof of Theorem 1 We continue to use the notation in the proof of Lemma 5. For $\tau_n^2 = m(p_{(2)} + p_{(2)}^2 - 2p_{(3)})$ defined in Lemma 6, we get from (14) that $m\sigma_n = \tau_n(1+o(1))$. It follows from (15) that

$$\begin{aligned}
\frac{Q_n - \mu_n}{\sigma_n} &= \frac{1}{m\sigma_n} \sum_{1\le i<j\le n} B_{ij}\bar{h}(c_i, c_j) - \frac{1}{m\sigma_n} \sum_{i=1}^{n} B_{ii}(p_{c_i} - p_{(2)}) \\
&= \frac{1+o(1)}{\tau_n} \sum_{1\le i<j\le n} B_{ij}\bar{h}(c_i, c_j) - \frac{1+o(1)}{\tau_n} \sum_{i=1}^{n} B_{ii}(p_{c_i} - p_{(2)}) \\
&= \frac{1+o(1)}{\tau_n} \sum_{1\le i<j\le n} A_{ij}\bar{h}(c_i, c_j) - \frac{1+o(1)}{2\tau_n} \sum_{1\le i<j\le n} \frac{k_i k_j}{m}\bar{h}(c_i, c_j) \\
&\quad + \frac{1+o(1)}{2\tau_n} \sum_{i=1}^{n} \frac{k_i^2}{m}(p_{c_i} - p_{(2)}).
\end{aligned}$$

Then (6) follows immediately from Lemma 6 if

$$\frac{1}{\tau_n} \sum_{1\le i<j\le n} \frac{k_i k_j}{m}\bar{h}(c_i, c_j) \xrightarrow{p} 0 \tag{27}$$

and

$$\frac{1}{\tau_n} \sum_{i=1}^{n} \frac{k_i^2}{m}(p_{c_i} - p_{(2)}) \xrightarrow{p} 0. \tag{28}$$

As a matter of fact, by using the orthogonality obtained in the proof of Lemma 5 we have from (2) that

$$\begin{aligned}
E\left(\frac{1}{\tau_n} \sum_{1\le i<j\le n} \frac{k_i k_j}{m}\bar{h}(c_i, c_j)\right)^2 &= \frac{1}{\tau_n^2} \sum_{1\le i<j\le n} \frac{k_i^2 k_j^2}{m^2} \mathrm{Var}(\bar{h}(c_i, c_j)) \\
&= O\left(\frac{\left(\sum_{i=1}^{n} k_i^2\right)^2}{m^3}\right) \to 0
\end{aligned}$$

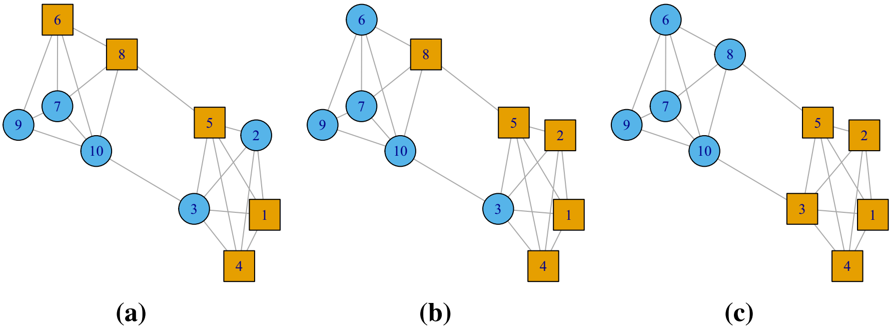

Fig. 1 Three different partitions for the same network. From left to right, the significance of the partition increases

and

$$E\left(\frac{1}{\tau_n}\sum_{i=1}^{n}\frac{k_i^2}{m}(p_{c_i}-p_{(2)})\right)^2=\frac{1}{\tau_n^2}\sum_{i=1}^{n}\frac{k_i^4}{m^2}\mathrm{Var}(p_{c_i}-p_{(2)})=O\left(\frac{\left(\sum_{i=1}^{n}k_i^2\right)^2}{m^3}\right)\rightarrow 0$$

which yield (27) and (28), respectively by Chebyshev's inequality. □

3 Numerical studies

In this section, we show some results from simulation studies and real data analysis. Computationally, we find that it is very convenient to use the R package `igraph` where many useful functions are implemented in generating, computing, and plotting complex networks (Csárdi and Nepusz 2006).

Figure 1 shows three different partitions for the same network. The network is generated by concatenating two dense subnetworks $\{1, 2, 3, 4, 5\}$ and $\{6, 7, 8, 9, 10\}$ with a couple of interconnected edges. Two colors and shapes represent two partitioned subgroups. Some related quantities are listed in Table 1. The modularities are -0.0238, 0.1667, and 0.4036, respectively. The theoretical mean and variance from (7) and (8) are -0.0510 and 0.00616 for all three partitions since they all partition the network into two equal-sized subgroups. The simulated mean and variance are close to the theoretical values. The z-scores are 0.347, 2.774, and 5.794, showing an increasing significance from three partitions. The p values also confirm that (c) is the most significant partition among these three.

To explicate the asymptotic normality when n gets large, we show three histograms of the simulated modularities from free labeling for three different network sizes $(n = 6, 12, 24)$ in Fig. 2. These networks are generated by stochastic block model (SBM) with two blocks of equal sizes (Faust and Wasserman 1992). For small size $(n = 6)$, the distribution is far from normal. However, when the network size is moderately large (around 24 in the current configuration), the sampling distribution is very close to normal. Their p values of Shapiro-Wilk normality test are nearly 0,

Table 1 Quantities from the three figures in Fig. 1

	Q_n	μ_n	σ_n^2	$\widetilde{\mu}_n$	$\widetilde{\sigma}_n^2$	z-score	p value
(a)	– 0.0238	– 0.0510	0.00616	– 0.0558	0.00533	0.347	0.364
(b)	0.167	– 0.0510	0.00616	– 0.0565	0.00568	2.774	2.77×10^{-3}
(c)	0.404	– 0.0510	0.00616	– 0.0506	0.00656	5.794	3.44×10^{-9}

Q_n is the modularity. μ_n and σ_n^2 are the expected mean and variance calculated using (7) and (8). Simulated mean $\widetilde{\mu}_n$ and variance $\widetilde{\sigma}_n^2$ are based on 1000 random assignments of colors

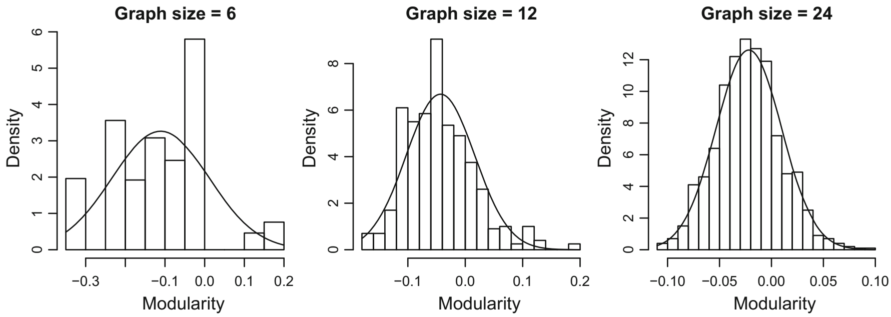

Fig. 2 Histograms for three different graph sizes (6, 12, and 24) from left to right. As the network size increases, the sampling distribution becomes more normal. The solid lines are the theoretical normal density lines with mean and variance from (7) and (8). The p values of Shapiro-Wilk normality test are nearly 0, nearly 0, and 0.07, respectively

nearly 0, and 0.07, respectively, confirming the better normality when the network size increases.

Finally, an illustration is given by using the well-known Zachary's karate club data set which has been regularly used as a benchmark to test different community detection algorithms (Zachary 1977). It contains 34 members of a karate club and an edge connecting two members represents they have interactions outside the activities of the club as shown in Fig. 3. At some point, there was a conflict between two central members, "A" and "H", who are the administrator and the instructor of the club. The club was then split into two smaller club as shown by different shapes and colors. This partition of the network has a modularity of 0.372. The theoretical sampling mean and variance are – 0.0248 and 0.00235, respectively. The simulated mean and variance are – 0.0253 and 0.00229, respectively. They are close since the sample size $n = 34$ is moderately large. This partition has a z-score of 8.175 with a p value of nearly 0, showing that it is indeed a very significant partition of the network. The histogram of the simulated modularities is shown in Fig. 4 which is approximately normally distributed.

An empirical power analysis is carried out to show the capability of detecting the departure from the null hypothesis. We construct a network using SBM containing 100 communities, each with 5 vertices. The edge probabilities are 0.3 and 0.1 for within and between communities, respectively. The original partition is the true community structure of the network. In the alternative H_a, we hold a certain number d of communities untouched while freely labeling the remaining vertices. Figure 5 shows the empirical

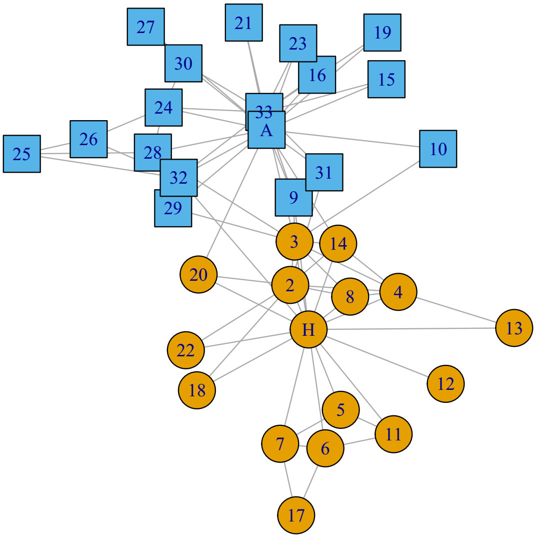

Fig. 3 Zachary's karate club network. The colors correspond to the split of the club into two separate groups. Vertices "A" and "H" are the administrator and the instructor of the club whose conflict causes the fission of the club (color figure online)

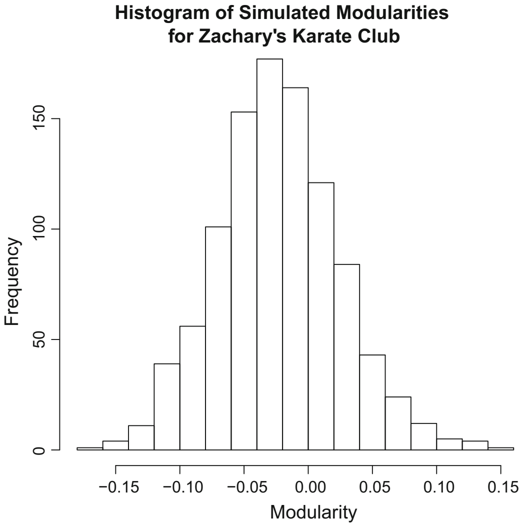

Fig. 4 Histogram of simulated modularities for Zachary's karate club network. The modularity of the actual fission of the club is 0.372, indicating a very significant partition of the network

power as a function of d. When d is small, the alternative is not very different from the null, giving a low power. As d increase, the community structure in H_a becomes stronger and the proposed method is more likely to detect it with higher power.

4 Conclusion and discussion

In this paper, we consider the asymptotic distribution of the modularity in a network under the null hypothesis of free labeling. Under free labeling, each vertex is independently assigned to different groups with probabilities equal to the proportions of

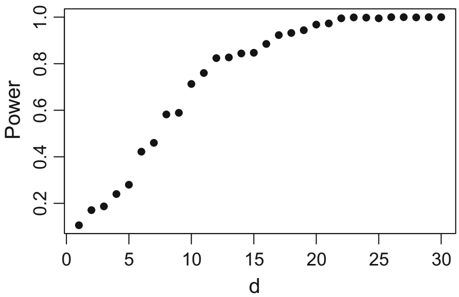

Fig. 5 Power as a function of d, the number of communities excluded from free labeling

vertices of different colors in the original network. Under some regularity conditions, the distribution is asymptotically normal when the size of the network goes to infinity. Some simulation work is performed to illustrate that the asymptotics kicks in reasonably fast and hence the asymptotic result may be used also for finite-sample situations. We also define the significance of a partition based on the z-score.

Even if a network is generated by ER model, it could have some spurious community structure (some regions are denser than other parts simply due to randomness.) A good modularity-maximization algorithm tends to find out those communities. If the obtained modularity is already the largest (or close to the largest) one, the test will almost always reject the null hypothesis and conclude the given partition (by maximizing modularity) captures some community structure of the network.

If two community detecting algorithms give the same number of communities, then we can compare their modularity directly which is equivalent to comparing the z-score in (9) since the asymptotic mean and variance in (7) and (8) are the same. On the other hand, if two algorithms result in different number of communities, then it makes more sense to compare two z-scores which take into consideration different μ_n and σ_n. Specifically, suppose that the asymptotic mean and variance of partition 1 are $\mu_{(1)}$ and $\sigma^2_{(1)}$ and partition 2 has $\mu_{(2)}$ and $\sigma^2_{(2)}$. Here we omit subscript n for brevity. Then we can use a t test type procedure and calculate

$$\frac{\mu_{(1)} - \mu_{(2)}}{\sqrt{\sigma^2_{(1)} + \sigma^2_{(2)}}}. \tag{29}$$

Two partitions will be claimed to be significantly different if this value is greater than $z_{\alpha/2}$ in magnitude. For example, Donetti and Muñoz used spectral properties of the graph Laplacian matrix and hierarchical clustering techniques to get a different partition from the one in Fig. 3 for Zachary's karate club network (Donetti and Muñoz 2004). See Figure 4 of Donetti and Muñoz (2004) for details. Its modularity is 0.4198 and $\mu_{(1)} = -0.03575$ and $\sigma^2_{(1)} = 0.001792$. The partition in Fig. 3 has modularity 0.3715 and $\mu_{(2)} = -0.02481$ and $\sigma^2_{(2)} = 0.002350$. The quantity in (29) is -0.17, indicating that these two partitions are not statistically different.

A related but different null hypothesis is also possible where the numbers of vertices with different colors are fixed and equal to the numbers of vertices with different colors in the original network. For example, if a community detection algorithm was applied to a network with size $n = 10$ and it finds out there are two communities with sizes 4 and 6, respectively. The null hypothesis is that each vertice is randomly assigned to one of the two colors under the constraints that four vertices will have color 1 and six vertices will have color 2. We call this case *non-free labeling* where the color assignments for different vertices are not independent. It turns out that the mean and variance of the distribution of modularity under non-free labeling are not difficult to obtain by following a similar argument in calculating (7) and (8). However, the theoretical derivation of the asymptotic distribution is much more complicated than free labeling. This is an on-going work and will be addressed in a separate paper.

Acknowledgements The authors would like to thank the Editor-in-Chief, an associate editor, and two anonymous reviewers for their valuable suggestions and comments which help improve the presentation of this paper.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding authors states that there is no conflict of interest.

References

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

Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion Limited, London

Csárdi G, Nepusz T (2006) The igraph software package for complex network research. InterJ Complex Syst 1695:1–9

Donetti L, Muñoz MA (2004) Detecting network communities: a new systematic and efficient algorithm. J Stat Mech Theory Exp 2004:P10012

Faust K, Wasserman S (1992) Blockmodels: interpretation and evaluation. Soc Netw 14:5–61

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

Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–146

Hall P, Heyde CC (1980) Martingale limit theory and its application. Academic Press, London

Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton

Lancichinetti A, Radicchi F, Ramasco JJ (2010) Statistical significance of communities in networks. Phys Rev E 81:046110

Moran PA (1950) Notes on continuous stochastic phenomena. Biometrika 37:17–23

Newman ME (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Newman ME (2006) Modularity and community structure in networks. Proc Nat Acad Sci 103:8577–8582

Newman ME (2010) Networks: an introduction. Oxford University Press, Oxford

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

Riordan O, Selby A (2000) The maximum degree of a random graph. Comb Probab Comput 9:549–572

Rosvall M, Bergstrom CT (2010) Mapping change in large networks. PLoS ONE 5:e8694

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

Zhang J, Chen Y (2016) A hypothesis testing framework for modularity based network community detection. Stat Sin 27:437–456

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.