Journal of Intelligent & Fuzzy Systems 29 (2015) 2707–2715
DOI:10.3233/IFS-151974
IOS Press



# The comparison of significance of fuzzy community partition across optimization methods

Hui-Jia Li*
*School of Management Science and Engineering, Central University of Finance and Economics, Beijing, China*

**Abstract**. The analysis of fuzzy(overlapping) community structure in complex networks is an important problem in data mining of network data sets. However, due to the exist of random factors and error edges in real networks, how to measure the significance of community structure efficiently is a crucial question. In this paper, we present a novel statistical framework comparing the significance of fuzzy community structure across various optimization models. Different from the universal approaches, we calculate the similarity between a given node and its leader and employ the distribution of link tightness to derive the significance score, instead of a direct comparison to a randomized model. Based on the distribution of community tightness, a new "$p$-value" form significance measure is proposed for community structure analysis. Specially, the well-known approaches and their corresponding quality functions are unified to a novel general formulation, which facilitate providing a detail comparison across them. To determine the position of leaders and their corresponding followers, an efficient algorithm is proposed based on the spectral theory. Finally, we apply the significance analysis to some famous benchmark networks and the good performance verified the effectiveness and efficiency of our framework.

Keywords: Fuzzy community, statistical significance, optimization methods, quality functions, benchmark networks

## 1. Introduction

A common feature observed in real networks is the presence of community structures [1–5, 18], i.e. subgraphs which are densely connected to each other while less connected to the subgraphs outside. In many scenarios, nodes in a network can belong to more than one community, called fuzzy(overlapping) communities [2–9]. In order to estimate how much a decomposition of a network which is found by a community detection algorithm is meaningful, we need a quality measure. Consequently, for a particular measure, the community detection algorithms can be ranked. To this end, various measures have been proposed in the literature, so far. The most prevalent measure which has been used extensively in the literature is due to Newman and Girvan [18]. This measure, called modularity, quantifies how much the density of the edges inside identified communities differs from the expected edge density in an equivalent network with similar number of vertices and edges but randomized edge placement, which is taken as the null model for statistical tests. Considering the modularity measure, the community detection problem is transformed to the modularity maximization problem. Modularity function can naturally extended to fuzzy form, which used to detect overlapping communities.

Recently, some optimization algorithms based on Potts models which used to detect community structure have attracted attention. Communities correspond to Potts model spin states, and the associated system

*Corresponding author. Hui-Jia Li, School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China. Tel.: +86 10 62288623; E-mail: Hjli@amss.ac.cn.

ISSN 1064-1246/15/$35.00 

energy indicates the quality of a candidate partition. It models an inhomogeneous ferromagnetic system where each node is viewed as a labeled spin in the network. Let $A$ be the adjacency matrix of graph $G$ and let $\sigma_i$ denote the label of the community that node $i$ belongs to. Furthermore, the Kronecker Delta function is defined by $\delta(\sigma_i, \sigma_j) = 1$ if $\sigma_i = \sigma_j$ and $\delta(\sigma_i, \sigma_j) = 0$, otherwise. Having the community membership labels $\sigma$, Reichardt and Bornholdt (RB) [16] proposed a generalized Hamiltonian as the core energy function,

$$H_{RB}(\{\sigma\}) = -\frac{1}{2}\sum_{i \neq j}(a_{ij} - \gamma_{RB} p_{ij})\delta(\sigma_i, \sigma_j). \quad (1)$$

where $\gamma_{RB}$ is the resolution parameter, $p_{ij} \in \mathbb{R}$ is the random form of adjacent matrix $A = (a_{ij})$. The Potts dynamical model is a powerful tool which has been wildly applied to uncover the dynamics of community structure in networks [8, 9].

Label propagation is another famous algorithm for community detection [27]. Briefly, the algorithm starts with randomly assigning a community label to each node. Then, each node updates its label by replacing it by the label most used by its neighbors. The other well-known optimization approaches used in community detection problem are Simulated Annealing (SA) [25], extermal optimization (DA) [13], expectation maximization [20], Bayesian inference [17], and variational Bayes [15]. For a comprehensive and comparative review on this topic we refer the reader to [4].

Although a lot of optimization method and their functions are proposed, How to determine the hidden properties of a given community [9] effectively remain unclearly answered. To answers these crucial questions, in this paper, we present a novel statistical framework comparing the significance of soft community structure across various optimization methods. Different from the universal approaches, we calculate the similarity of a given node to its leader and employ the distribution of link tightness to derive the significance score, instead of a direct comparison to a randomized model. A small example is shown in Fig. 1a, which illustrates that tighter the following nodes link to its leader, more significant the community is. Based on the distribution of community tightness, a new "$p$-value" form significance measure is proposed for community structure analysis. Specially, the well-known approaches and their corresponding quality functions are unified to a novel general fuzzy formulation, to provide a detail comparison across them. Then, we can choose the most suitable form of the function by set the parameters properly. To determine the position of leaders and their corresponding followers, an efficient fuzzy detection algorithm is proposed based on the spectral theory. Finally, we apply the significance analysis to some famous benchmark networks and the good performance verified the effectiveness and efficiency of our framework. The detailed procedure can be observed in Fig. 1b.

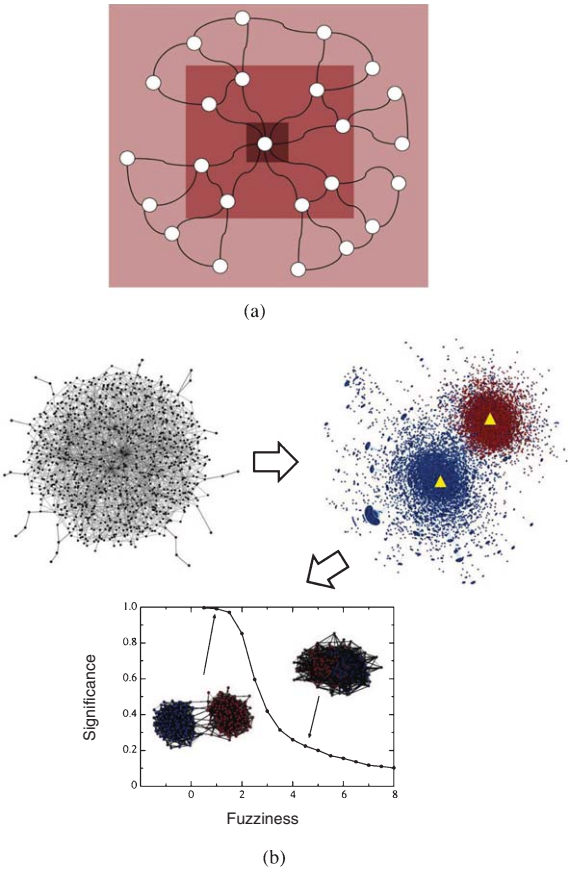
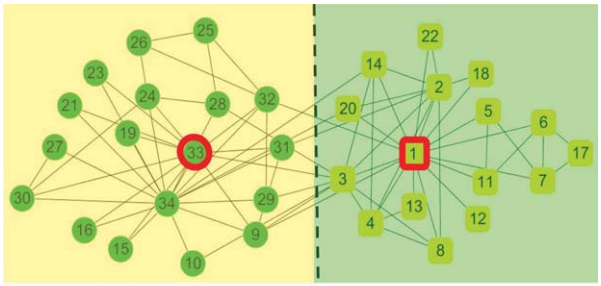


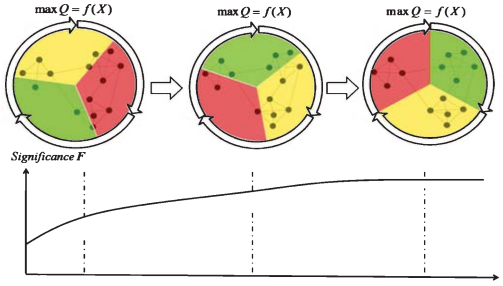

Fig. 1. (a) For a given community, the leader node usually locates on the highest level, representing the most influential node. Circles depict different levels in the network hierarchy, with the darkest color denoting the highest level. Tighter the following nodes link with its leader, more significant the community is. (b) The procedure of our framework. First, for a given network shown in the left subgraph, we derive the centers represented by triangle nodes in the right subgraph and their corresponding community partition highlighted with different colors. Then, based on the position of center and other following nodes, a new "$p$-value" form significance measure is proposed to measure the quality of community structure, which is shown in lower subgraph.

## 2. Community structure and the leader

Leader-driven algorithms [11, 12] constitute a special case of seed-centric approaches. These methods




(a)




(b)

Fig. 2. (a) The network topology of Zachary karate network. Two communities are represented by different shapes and colors. Node 1 and 34 are leaders which highlighted in the origin graph. (b) In every circle, sectors with different colors represent different communities. It can be noticed that the community partition in the rightmost circle is strongest due to the fewest intercommunity edges. When we use a given optimization method to evolve the community configure $X$ (describe by different sectors) based on maximizing the objective function max $Q = f(X)$, the significance of $X$ also evolves correspondingly. The $F$ score is utilized to measure the significance of community configure $X$. Here, the global maximum of $F$ is maybe an asymptotically stable fixed point of dynamical system associates to community configure $X$ in the rightmost circle.

show that, in many real world, especial the social networks, nodes of a network are usually classified into two categories: leaders and followers. For example, considering the famous Karate network [28], nodes 1 and 33 are two significant leaders and corresponding communities are built around them (see Fig. 2a). If two leaders are removed, these communities will be split up, as they link to most followers and keep the community together. Since community are consequence of information spreading, a given community can be defined as the area in which a leader has most influence. So, one can uncover the community partition by finding all natural leaders and their corresponding followers on which they influence. We believe if followers are more tightly linked to the leader, or leader spreads more influence on their followers, this community are more significant or robust. When we use a given optimization method to evolve the community configure, the significance of communities also evolves correspondingly, which shown in Fig. 2b.

## 3. The fuzzy community detection algorithm based on leader position

In this study, the relative positions of leader and corresponding followers are crucial to analyze the significance situation. In order to obtain the leader of corresponding community, we extract the candidate fuzzy membership by minimizing the following objective function

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}\|d_i - c_j\|^2, \tag{2}$$

where variables $x_{ij}$ is the fuzzy membership that node $i$ in community $j$, with $\sum_j x_{ij} = 1$. This method is similar as the famous $k$-means method and can be obtain both center and assignment iteratively. $d_i$ is the $i$th $n$-dimensional data point, $c_j$ is the $n$-dimensional center(leader) of the community $j$, and $\| * \|$ is any norm expressing the similarity between a given node and the center. One can use an iterative optimization of the objective function shown above, to obtain the network partition by the update of fuzzy membership $x_{ij}$ and the community leaders $c_j$ . This procedure converges to a local minimum or a saddle point of $J_m$.

Suppose $K$ is the upper bound of number of clusters and $A = (a_{ij})_{n\times n}$ is the adjacent matrix of a network, then, the detailed algorithm is shown in Algorithm 1 and stated straightforwardly as follows:

Step 1: for a given $K$

($i$) Calculate the diagonal matrix $D = (d_{ii})$, where $d_{ii} = \sum_k a_{ik}$.

($ii$) Computing the top $K$ eigenvectors based on generalized eigensystem $Ax = tDx$, and then establish the eigenvector matrix $E_K = [e_1, e_2, ..., e_K]$.

Step 2: for each number of communities $2 \le k \le K$:

($i$) Establish the matrix $E_K = [e_2, e_3, ..., e_K]$ from the matrix $E_K$.

($ii$) Normalize the rows of $E_K$ to unit length using Euclidean distance norm.

($iii$) Cluster the row vectors of $E_K$ using any community detection method by minimizing Equation.(2) to obtain a membership matrix $X_k$ and corresponding leaders.

Step 3: Maximizing the modular function: Pick the optimal number of communities $k$ and the corresponding partition $X_k$ that maximizes $Q(X_k)$.

In step 1, given the adjacent matrix $A = (a_{ij})_{n\times n}$ and a diagonal matrix $D = (d_{ii})$, $d_{ii} = \sum_k a_{ik}$, two matrices $D^{-1/2}AD^{-1/2}$ and $D^{-1}A$ are used. This is motivated by Ref. [7], which uses the top $K$ eigenvectors of the generalized eigensystem $Ax = tDx$ instead of the $K$ eigenvectors of the adjacent matrix. It shows that after normalizing the rows using Euclidean norm, their eigenvectors are mathematically identical and emphasize that this is a numerically more stable method. In step 2, we choose the initial the starting centers to be as orthogonal as possible which already used in $k$-means clustering method [6, 26]. This way of choosing centers(leaders) does not cost additional time complexity, and also improve the quality of the partition, thus at the same time reduces the need for restarting the random initialization process. In step 3, the $Q$ function measures the quality of a given community structure organization of a network and can be used to automatically select the optimal number of communities $k$ according to the maximum $Q$ value [26], we will discuss the multiple optimization methods and their corresponding $Q$ function in detail in the following section.

**Algorithm 1** The fuzzy community detection algorithm

**Require:** Graph $G$ with size $n$ and volume $m$, the algorithm parameters, i.e. $f_\mu^+$, $f_\mu^-$ and $R_\mu$ which shown in Equation. (3)

**Ensure:** The fuzzy membership matrix $X$;

1: For a given number of communities $K$
2: **repeat**
3: Calculate the top $K$ eigenvector matrix $E_K = [e_1, e_2, ..., e_K]$ and initiate the community membership $X(0) = E_K$.
4: Update the position of center and corresponding community membership matrix $X$ to minimize the Equation.(2)
5: **Until** exceeding the maximum number of iterations
6: Select the optimal number of communities $K$ and corresponding community membership according to the maximum of $Q$ defined in Equation. (3)

## 4. The general and expanded formation of function $Q$

For many community detection algorithms, the target function $Q$ is critical. Here, $Q$ can be tried to be optimized has the following general fuzzy form:

$$Q_{i\mu} = \sum_{j=1}^{n} f_\mu^+ a_{ij} x_{j\mu} - \sum_{j=1}^{n} f_\mu^-(1 - a_{ij}) x_{j\mu} + R_{i\mu}, \tag{3}$$

and choose $R_{i\mu}$ such that $\partial R_{i\mu}/\partial x_{i\mu} = 0$ and $R_{mu} = \sum_{i=1}^{n} R_{i\mu}$, e.g. $R_{i\mu} = \frac{2}{l_\mu} R_{mu}$. Interestingly, when all $x_{i\mu}$ are in fuzzy membership state, the $H$ function with $Q_{i\mu}$ defined as Equation. (3) can be reduced to well-known measures by following considerations:

(1) Hofman and Wiggins [15]

$$f_\mu^+ = \log \frac{p^{in}}{p^{out}}, f_\mu^- = \log \frac{1 - p^{out}}{1 - p^{in}}, \quad R_\mu = l_\mu \log \pi_\mu. \tag{4}$$

(2) Ronhovde and Nussinov [23]

$$f_\mu^+ = 1, f_\mu^- = \min_\mu p_{in,\mu}, R_\mu = 0. \tag{5}$$

(3) RB Potts model (Erdős-Rényi null model) [16]

$$f_\mu^+ = 1 - \gamma_{RB} p, f_\mu^- = \gamma_{RB} p, R_\mu = 0. \tag{6}$$

(4) RB Potts model (Configuration null model)[16]

$$f_\mu^+ = 1 - \frac{\gamma_{RB}}{2m}, f_\mu^- = \frac{\gamma_{RB}}{2m}, \quad R_\mu = \sum_{i>j} \frac{\gamma_{RB}}{2m}(k_i k_j - 1) x_{i\mu} x_{j\mu}. \tag{7}$$

where $k_i$ is the degree of node $i$ and $m$ is the number of all edges in the network.

(5) Modularity [18]

$$f_\mu^+ = 1, f_\mu^- = \frac{k_i k_j}{2m}, R_\mu = \sum_{i>j} \frac{1}{2m}(k_i k_j - 1) x_{i\mu} x_{j\mu}. \tag{8}$$

where $k_i$ is the degree of node $i$ and $m$ is the number of all edges in the network.

(6) Label propagation [27]

$$f_\mu^+ = 1, f_\mu^- = 0, R_\mu = 0. \tag{9}$$

where $k_i$ is the degree of node $i$ and $m$ is the number of all edges in the network.

## 5. Significance of community structure

It is essential to establish a detail framework analyzing the significance of community structure, since real

networks own specific characteristics [5, 10, 29]. In this section, we discuss these important characteristics and give a detailed introduction of the framework.

**Node similarity.** We define the similarity of nodes $i$ and $j$, $sim(i, j)$, as the ratio between the intersection and the union of their neighborhoods $\Gamma(i)$ and $\Gamma(j)$,

$$sim(i, j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}, \tag{10}$$

By employing Equation. (10), we can calculate the expected similarity between a given node and the community leader $z$,

$$E[sim(x, z)] = \int_{\mathbb{R}^M} sim(x, z) Q(x|z) dx, \tag{11}$$

where $Q(x|z)$ is a distribution of nodes in a community with leader $z$.

Next, using the maximum entropy principle, the statistical unbiased distribution fulfilling constraint can be obtained using the maximum entropy principle:

$$Q(x|z, \eta) = \frac{1}{Z_\eta} P_0(x) e^{\eta sim(x,z)} dx, \tag{12}$$

where $P_0(x)$ is the background distribution used to contrast with an alternative hypothesis: node $x$ being part of a community, a group of nodes distinguished by enhanced mutual similarity. $Z_\eta$ is the normalisation constant depends on the value of the *scoring parameter* $\eta$:

$$\frac{\partial}{\partial \eta} \log Z_\eta = E[sim(x, z)]. \tag{13}$$

$\eta$ is the parameter which used to control the "width" of a community and the larger the value of $\eta$, the smaller the expected width or scale of a given community. Specially, the distribution $Q(x|z, \eta)$ is the same as the background model $P_0(x)$ when $\eta = 0$.

**Log-likelihood score and community tightness.** We define the log-likelihood score as the deviations of the community distribution from the null model

$$s(x|z, \eta) \equiv \log \frac{Q(x|z, \eta)}{P_0(x)} = \eta sim(x, z) - \log Z_\eta. \tag{14}$$

By Equation. (14), nodes which are more likely to be in a community with center $z$ and scoring parameter $\eta$ own larger positive value, than in the null background model. Given a community with nodes set $\{1, ..., N\}$, for a given leader $z$ and a scoring parameter $\eta$, the log-likelihood scores $s(i|z, \eta)$ are positive. The community tightness is the sum of the scores of the community elements,

$$S(1, ..., N|z, \eta) = \sum_i \max[s(i|z, \eta), 0]. \tag{15}$$

However, we can't determine the scoring parameter $\eta$ easily. Here, the tightness function of Equation.(15) can be simplified as:

$$S(1, ..., N|z, \eta) = \sum_{i=1}^{N} \max[s(i|z) - \mu, 0], \tag{16}$$

where $s(i|z) = sim(i, z)$. By this transformation, one can control the width of community using parameter $\mu$ simply. The community tightness is determined both by the number of elements and by their similarities with the leader, that is, tighter communities with fewer elements own comparable more tightness to looser but larger communities.

**Calculation of Significance score.** We can the quantified the quality of the true and random communities by characterize the distribution of the tightness score $p(S)$ from the background distribution. A new "$p$-value" form measure [14] can be used to define the statistical significance of score $S_0$, as the probability that a random chosen nodes set contains a community with score greater than or equal to $S_0$. This "$p$-value" form significance can be explained by a null hypothesis: "These nodes are drawn from the background distribution". To test this hypothesis, we compute the statistical significance of score $S_0$: low value suggests that the null hypothesis is unlikely and allows for rejecting it. This method provides a new connection between statistical $p$-value theory and network analysis and then get an interesting significance measure.

If the network is large enough, according to the mean field theory, $s_i = s(i|z)$ owns an approximate Gaussian-distribution with variance $M$, $P(s(i|z)) = \sqrt{1/(2M\pi)} \exp\{-s^2/(2M)\}$. The distribution of the tightness $S$ can be calculated straightforwardly using the derivation. Specifically, we need to compute the following quality function:

$$\begin{aligned} Z_c(\beta, \mu) &= \int_{\mathbb{R}^N} e^{\beta S(1,...,N|z,\eta)} P(s_1)...P(s_N) ds_1...ds_N \\ &= \left[ \int_{-\infty}^{+\infty} e^{\beta \max[s_i - \mu, 0]} P(s) ds \right]^N \end{aligned}$$

$$= \left[\int_{-\infty}^{\mu} P(s)ds + \int_{\mu}^{+\infty} e^{\beta(s_i-\mu)}P(s)ds\right]^N$$

$$= \left[(1 - H(\mu)) + e^{\frac{(\beta)^2}{2}-\beta\mu}H(\mu-\beta)\right]^N, \qquad (17)$$

where $H(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ is the complementary cumulative Gaussian distribution. In Equation.(17), two intervals are divided: below the score threshold $\mu$, the score is zero, which contributes the cumulative distribution $\int_{-\infty}^{\mu} ds/(2\pi)^{1/2} \exp[-s^2/2]$ to the generating function. Above $\mu$, the score is positive, which generates a contribution of $\int_{\mu}^{+\infty} ds/(2\pi)^{1/2} \exp[-s^2/2 + \beta(s-\mu)]$. The free energy function reads

$$-\beta f(\beta, \mu) = \log[(1 - H(\mu)) + e^{\frac{(\beta)^2}{2}-\beta\mu}H(\mu-\beta)], \qquad (18)$$

and the entropy is

$$\omega(s, \mu) = -\max_{\beta}[\beta s + \beta f(\beta, \mu)]. \qquad (19)$$

Using the distribution of community tightness, there is

$$\log p(S, \mu) \simeq N\omega(S/N, \mu) - \frac{1}{2}\log N. \qquad (20)$$

Given a specifical community, we can calculated the significance score $F$ using the probability that the community tightness $S$, $p(S)$, larger than or equal to $S$,

$$F(S, \mu) = \int_{S}^{+\infty} p(S', \mu)dS'. \qquad (21)$$

Furthermore, from the perspective of the whole network, we use the average significance score $\langle F\rangle_Q$ to indicate the robustness of a partition, defined as the average value among $F$ values of all communities partitioned by maximizing a particular quality function $Q$ shown in Section 4.

## 6. Experiments on benchmark network

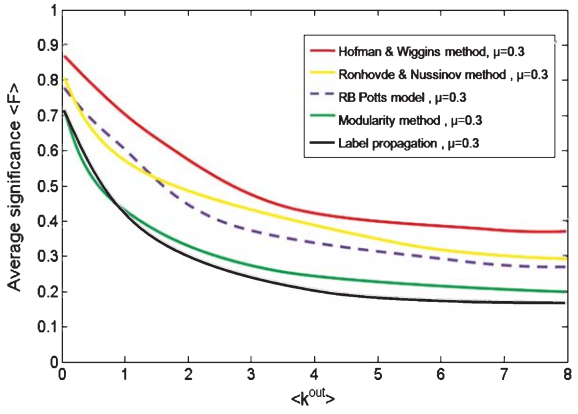
In this section, we will test the validity of our framework on some famous benchmark network and real networks.

**GN benchmark network.** First, we apply to the classical Girven-Newman benchmark [21], where the network with $n = 128$ nodes are divided into four 32 nodes communities. According to the establish mechanism, the community structure will fuzzier and thus when $\langle k^{out}\rangle$ increases, it is more difficult to identify them correctly. Hence, the significance of communities will tend to be weaker and the value of $F$ index will also decrease. The comparison results of $F$ value corresponding to all five optimization algorithms are shown in Fig. 3a when $\mu = 0.3$. It can be observed that the index $F$ has a great performance on GN benchmark: when $\langle k^{out}\rangle$ approaching 0, the community structure is quite strong and all corresponding $\langle F\rangle$ value is close to 1; while when the network is fuzzy enough, the corresponding $\langle F\rangle$ value of all algorithm is low, extremely for Modularity optimization method and Label propagation method, only near 0.2–0.3.

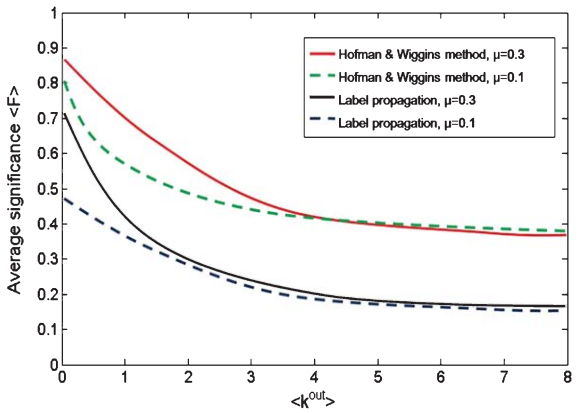
Moreover, by comparing five algorithms, we find in Fig. 3a that the $\langle F\rangle$ values corresponding to Hofman & Wiggins method is largest, and the Label propagation method is the lowest. This may because Label propagation method emphasize the simplicity of calculation too much while ignoring the accuracy of results. Furthermore, the $\langle F\rangle$ values between Modularity optimization method and Label propagation method are similar when $\langle k^{out}\rangle$ becomes lower. This result is similar as Ref. [22], which verifies the inner correlation between these two methods. These observations are no evidence of overall superiority of one method over another, but an example of how to compare the significance and use the different partitioning algorithms on a given network.

Furthermore, when $\langle k^{out}\rangle$ increases, the topology becomes fuzzier and the sizes of communities will become more and more smaller correspondingly. At the same time, as the width parameter $\mu$ increases, the significance will favor tighter communities with fewer elements. We test the Hofman & Wiggins method and Label propagation method in Fig. 3b, the value of $\langle F\rangle$ corresponding to $\mu = 0.3$ will be larger than $\mu = 0.1$ for all two examples. As a conclusion, we argue that when the corresponding $\langle F\rangle$ is smaller than 0.3 on average ($\langle k^{out}\rangle \approx 4$), it is not safe to say there exists significant community structure for a given network.

**LFR benchmark network.** We also test the index on the more challenging LRF benchmark presented by Lancichinetti, Fortunato and Radicchi [3]. In this network, the average degree $k = 20$, maximum degree is 50 and $P(k) \propto k^{\gamma}$. Maximum and minimum community sizes are 50 and 20 respectively. The significance score changes when we adjust the value of $\theta$ in LFR benchmark, and numerical results in the LFR-benchmark are shown in Fig. 4a. It can be observed that with the augment of $\theta$, $F$ decreases for all five optimization methods when $\mu = 0.3$. Same as GN network, the $\langle F\rangle$ values corresponding to Hofman &

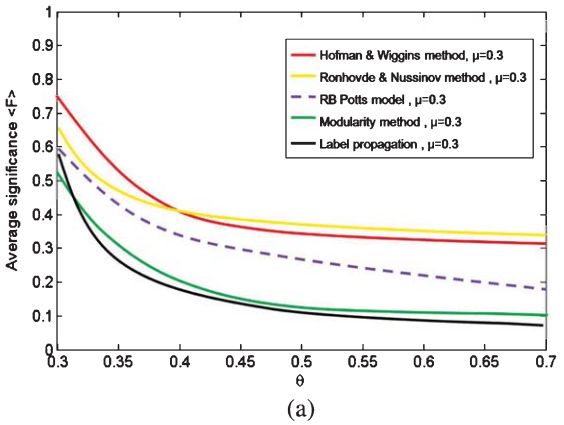



(a)

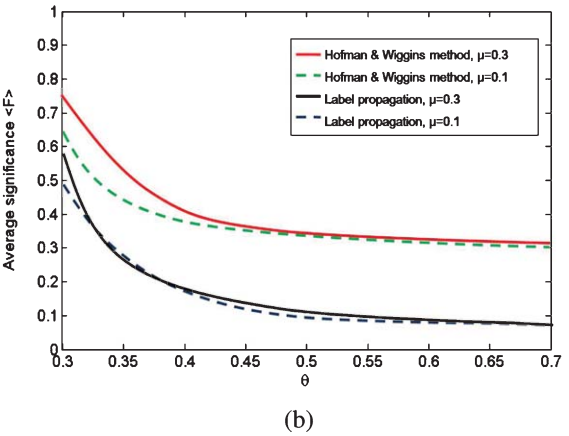



(b)

Fig. 3. The experimental results of significance $\langle F\rangle$ on GN benchmark network and each point in curves is obtained by testing 100 times. (a) For all five optimization methods, $\langle F\rangle$ decreases with increasing of $\langle k^{out}\rangle$. For a given network, when $\langle F\rangle$ is larger than 0.3 on average($\langle k^{out}\rangle \approx 4$), one can say there exit significant community structure. (b)The value of $\langle F\rangle$ corresponding to $\mu = 0.3$ will be larger than $\mu = 0.1$ for the Hofman & Wiggins method and Label propagation method. This implies as the width parameter $\mu$ increases, the significance favors tighter communities with fewer elements.




(a)




(b)

Fig. 4. The performance of significance $\langle F\rangle$ on LFR benchmark network and each point in curves is obtained by testing 100 times. (a) In this network, the average degree $k = 20$, maximum degree is 50 and $P(k) \propto k^{\gamma}$. Maximum and minimum community sizes are 50 and 20 respectively. For all five algorithms, the $\langle F\rangle$ index decreases with the increasing of mix parameter $\theta$. When $\theta \geq 0.5$ on average (no significant community), $\langle F\rangle$ is near 0.3 which is similar with GN network. (b) The value of $\langle F\rangle$ corresponding to $\mu = 0.3$ will be larger than $\mu = 0.1$ for the Hofman & Wiggins method and Label propagation method.

Wiggins method is largest at the begining, and the Label propagation method is the lowest. However, the $\langle F\rangle$ values corresponding to Ronhovde & Nussinov method will exceed Hofman & Wiggins method when when $\theta$ is larger than 0.4. Furthermore, when $\theta$ larger than 0.32, the $\langle F\rangle$ value corresponding to Label propagation method is close to Modularity optimization method. In addition, from Fig. 4b, it can be observed the value of $\langle F\rangle$ corresponding to $\mu = 0.3$ will larger than $\mu = 0.1$ when we take the Hofman & Wiggins method and Label propagation method as examples.

**Stochastic block model.** Furthermore, we consider the famous stochastic block model which used to detect community structure by Decelle and Zhang et al. [1, 2, 24]. In this benchmark, $\varepsilon = c_{out}/c_{in}$ is the parameter used to control the fuzziness of generated network. To verify the performance on sparse stochastic block model with low average degree, we generate a large network with $N = 5000$ nodes and $q = 10$ groups with average degree $c = 8$, which shown in Fig. 5. Each point in curves is the result averaged by testing 100 times. When $\varepsilon$ is close to 0, it can be observed the community structure is quite strong and the corresponding $\langle F\rangle$ value of all five algorithms are very high when

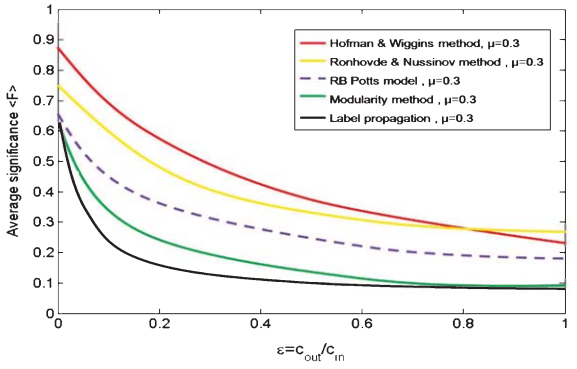



Fig. 5. The performance of social significance $\langle F\rangle$ on stochastic block model. In this example, there are $N = 5000$ nodes and $q = 10$ groups. The average degree $c = 8$ and parameter $\varepsilon = c_{out}/c_{in}$ is used to control the fuzziness of generated network. Each point in curves is obtained by testing 100 times. With the increasing of $\varepsilon$, the $\langle F\rangle$ index decreases. For all algorithm, when the corresponding $\langle F\rangle$ is nearly larger than 0.3 on average ($\varepsilon \approx 0.4$), there exists significant community structure which may detectable.

Table 1
Comparison of various algorithms with $\langle F\rangle$ values

| Networks | Algorithms | Values of $\langle F\rangle$ |
|---|---|---|
| Zachary | Label | 0.641 |
| | GN | 0.735 |
| | RB Potts | 0.827 |
| Collage football | Label | 0.602 |
| | GN | 0.758 |
| | RB Potts | 0.831 |
| Political books | Label | 0.581 |
| | GN | 0.698 |
| | RB Potts | 0.717 |

$\mu = 0.3$. In contrast, when $\varepsilon$ is increased close to 0.8, the network is nearly a fuzzy random one, and all $\langle F\rangle$ values are very low, near 0.1–0.3. Furthermore, we find that the $\langle F\rangle$ value of Hofman & Wiggins method will larger than all others when $\varepsilon < 0.81$, while lower than Ronhovde & Nussinov method when $\varepsilon > 0.81$. Specifically, we argue that for all algorithm when the corresponding $\langle F\rangle$ is nearly larger than 0.3 on average($\varepsilon \approx 0.4$), there exists significant community structure which may detectable [1]. From the results, the $F$ shows a great ability in characterizing the significant modular structure for optimization methods as we adjust the parameter $\varepsilon$.

**Real network.** Finally, we show significance can also be used to rank the real network partitions obtained by different algorithmic strategies. Zachary karate club network, Collage football network and Political books network are employed as the examples. Table 1 presents the results estimated from three algorithms and we observed that they are coincided with the analysis in artificial networks. These observations are no evidence of overall superiority of one method over another, but an example of how to compare the significance and use the different partitioning algorithms on a given network.

## 7. Discussion

In this paper, we present a novel framework comparing the significance of fuzzy community structure revealed by multiple optimization functions. Based on the distribution of community tightness, a new "$p$-value" form significance measure is proposed for analysis. As part of the future work, it is necessary to take a deeper look into how different similarity measures impact the results of this method. Additionally, this framework can be easily extended to a weighted and directed form, which only needs to modify the formation of the quality function $Q$. As a conclusion, this method shows a great performance and deserves more attention from us.

## Acknowledgments

We are grateful to the anonymous reviewers for their valuable suggestions which are very helpful for improving the manuscript. The authors are separately supported by NSFC grants 71401194, 91324203, 11131009 and "121" Youth Development Fund of CUFE grants QBJ1410.

## References

[1] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Phase transition in the detection of modules in sparse networks, *Physical Review Letters* **107**(6) (2011), 065701.
[2] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, *Physical Review E* **84**(6) (2011), 066106.
[3] A. Lancichinetti, S. Fortunato and F. Radicchi, Benchmark graphs for testing community detection algorithms, *Physical Review E* **78**(4) (2008), 046110–046115.
[4] A. Lancichinetti and S. Fortunato, Community detection algorithms: A comparative analysis, *Physical Review E* **80**(5) (2009), 056117.
[5] A. Lancichinetti, F. Radicchi, J.J. Ramasco and S. Fortunato, Finding statistically significant communities in networks, *PloS One* **6**(4) (2011), e18961.

[6] A.Y. Ng, M.I. Jordan and Y. Weiss, On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems* **2** (2002), 849–856.
[7] D. Verma and M. Meila, A comparison of spectral clustering algorithms, *University of Washington Tech Rep UWCSE* **1**(1-18) (2003), 030501.
[8] S. Fortunato, Community detection in graphs, *Physics Reports* **486**(3-5) (2010), 75–174.
[9] H.J. Li, H. Wang and L. Chen, Measuring robustness of community structure in complex networks, *Europhysics Letters* **108**(6) (2015), 68009.
[10] H.J. Li and J.J. Daniels, Social significance of community structure: Statistical view, *Physical Review E* **91**(1) (2015), 012801.
[11] H.J. Li, Y. Wang, L.Y. Wu, J. Zhang and X.S. Zhang, Potts model based on a Markov process computation solves the community structure problem effectively, *Physical Review E* **86**(1) (2012), 016109.
[12] H.J. Li and X.S. Zhang, Analysis of stability of community structure across multiple hierarchical levels, *Europhysics Letters* **103**(5) (2013), 58002.
[13] J. Duch and A. Arenas, Community detection in complex networks using extremal optimization, *Physical Review E* **72**(2) (2005), 027104.
[14] J.D. Wilson, S. Wang, P.J. Mucha, S. Bhamidi and A.B. Nobel, A testing based extraction algorithm for identifying significant communities in network, *The Annals of Applied Statistics* **8**(3) (2013), 1853–1891.
[15] J.M. Hofman and C.H. Wiggins, Bayesian approach to network modularity, *Physical Review Letters* **100**(25) (2008), 258701.
[16] J. Reichardt and S. Bornholdt, Statistical mechanics of community detection, *Physical Review E* **74**(1) (2006), 016110.
[17] M.B. Hastings, Community detection as an inference problem, *Physical Review E* **74**(3) (2006), 035102.
[18] M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical Review E* **69**(2) (2004), 026113.
[19] M.E.J. Newman, Fast algorithm for detecting community structure in networks, *Physical Review E* **69**(6) (2004), 066133.
[20] M.E.J. Newman and E.A. Leicht, Mixture models and exploratory analysis in networks, *Proceedings of the National Academy of Sciences of the United States of America* **104**(23) (2007), 9564–9569.
[21] M. Girvan and M.E.J. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences of the United States of America* **99**(12) (2002), 7821–7826.
[22] M.J. Barber and J.W. Clark, Detecting network communities by propagating labels under constraints, *Physical Review E* **80**(2) (2009), 026129.
[23] P. Ronhovde and Z. Nussinov, Local resolution-limit-free Potts model for community detection, *Physical Review E* **81**(4) (2010), 046114.
[24] P. Zhang and C. Moore, Scalable detection of statistically significant communities and hierarchies, using message-passing for modularity, *Proceedings of the National Academy of Sciences of the United States of America* **111** (2014), 18144.
[25] R. Guimera and L.A.N. Amaral, Functional cartography of complex metabolic networks, *Nature* **433**(7028) (2005), 895–900.
[26] S. Zhang, R.S. Wang and X.S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, *Physica A: Statistical Mechanics and its Applications* **374**(1) (2007), 483–490.
[27] U.N. Raghavan, R. Albert and S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E* **76**(3) (2007), 036106.
[28] W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **1** (1977), 452–473.
[29] Y. Hu, Y. Nie, H. Yang, J. Cheng, Y. Fan and Z. Di, Measuring the significance of community structure in complex networks, *Europhysics Letters* **82**(6) (2010), 066106.