# Social significance of community structure: Statistical view

Hui-Jia Li[1,2,*] and Jasmine J. Daniels[3]

[1]*School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

[3]*Department of Applied Physics, Stanford University, Stanford, California 94305, USA*

(Received 16 July 2014; revised manuscript received 3 December 2014; published 5 January 2015)

Community structure analysis is a powerful tool for social networks that can simplify their topological and functional analysis considerably. However, since community detection methods have random factors and real social networks obtained from complex systems always contain error edges, evaluating the significance of a partitioned community structure is an urgent and important question. In this paper, integrating the specific characteristics of real society, we present a framework to analyze the significance of a social community. The dynamics of social interactions are modeled by identifying social leaders and corresponding hierarchical structures. Instead of a direct comparison with the average outcome of a random model, we compute the similarity of a given node with the leader by the number of common neighbors. To determine the membership vector, an efficient community detection algorithm is proposed based on the position of the nodes and their corresponding leaders. Then, using a log-likelihood score, the tightness of the community can be derived. Based on the distribution of community tightness, we establish a connection between $p$-value theory and network analysis, and then we obtain a significance measure of statistical form . Finally, the framework is applied to both benchmark networks and real social networks. Experimental results show that our work can be used in many fields, such as determining the optimal number of communities, analyzing the social significance of a given community, comparing the performance among various algorithms, etc.

DOI: 10.1103/PhysRevE.91.012801 PACS number(s): 89.75.Fb, 89.65.Ef, 05.10.−a

## I. INTRODUCTION

Community structure detection [1–3] is a main focus of social network studies. It has attracted a great deal of attention from various scientific fields. Intuitively, "community" refers to a group of nodes in a network that is more densely connected internally than with the rest of the network. A well-known exploration for this problem is the concept of modularity, which was proposed by Newman *et al.* [1–3] to quantify a network's partition. Optimizing modularity is effective for community structure detection, and it has been widely used in many real networks. However, as pointed out by Fortunato *et al.* [4], modularity suffers from the resolution limit problem, which involves the reliability of the communities detected through the optimization of modularity. In conjunction with the modularity concept, many efforts have been devoted to understanding the properties of dynamical processes taking place in underlying networks. Specifically, researchers have begun to investigate the correlation between community structure and dynamical systems such as synchronization [5] and the random-walk process [6–11]. Recently, extensive studies were performed on the phase transition [12–14] of an algorithm from an undetectable region to one where detection is possible, and the performance of a variety of partition methods was investigated.

However, despite the large volume of work on community structure detection and its applications, one important question has not been clearly addressed, i.e., that of the significance of the communities in social networks. How can we distinguish real communities from fake ones? How can we tell when the communities detected by different methods are truly significant or when they could merely be a consequence of a chance coincidence of edge positions in the network? How do we statistically determine the significance of a given social community [7,8]? Clear answers to these questions are crucial for scientists from many fields.

The value of the modularity can be used as a quality function for communities: a network with a strong community structure will have high modularity, and hence it is proposed to evaluate the community partition. However, recent studies have shown that this approach is insufficient [15–22]. Although it is true that networks with a strong community structure have high modularity, it turns out that not all networks with high modularity have a strong community structure. Researchers have found that there are networks with no obvious community structure at all that nonetheless have high modularity. In [23], Guimera *et al.* showed numerically that divisions exist in ordinary random graphs that have high modularity, even in the limit of large network size, a result confirmed in later analytic calculations by Reichardt and Bornholdt [24]. The reason for this is that the number of possible divisions of a network increases extremely fast with network size (faster than any exponential), so that it is highly improbable that any one division will, purely by chance, have high modularity. As a result, high modularity is only a necessary but not a sufficient condition for significant community structure.

If the algorithms are able to identify communities even in random graphs, which value should we give to communities found in real networks? This problem has been the subject of some studies in the literature [10–14]. In [24,25], for example, the maximum of the modularity of the network analyzed is compared with the maximum of the same function measured in a randomized version of the network itself (i.e., all edges are randomly rewired). In contrast, in [26] the importance of a community partition is proportional to its robustness against random perturbations (i.e., random reshuffling of edges). The

*Hjli@amss.ac.cn

basic idea is that, if a partition is significant, it will be recovered even if the structure of the graph is modified, as long as the modification is not too extensive. Instead, if a partition is not significant, one expects that minimal modifications of the graph will suffice to disrupt the partition. In a recent work by Bianconi *et al.* [27], the notion of entropy $\Theta$ of graph ensembles was employed to find out how likely it is for a cluster structure to occur on a graph with a given degree sequence. The entropy is computed from the number of graph configurations that are compatible with a given classification of the nodes in $q$ groups. If the entropy $\Theta \gg 1$, the cluster structure is far more likely than a random classification of the nodes, so the clustering is relevant. Lancichinetti *et al.* [28] also addressed the issue by comparing the cluster structure of the graph with that of a random graph with similar properties. They found that, in fact, not all communities are equally significant in general, so it makes a lot of sense to check them individually. In particular, it may be that real networks are not fully modular, due to their particular history or generating mechanisms, and that only portions of them display community structure. The main idea is to verify how likely it is that a community $C$ is a subgraph of a random graph with the same degree sequence of the original graph, using the proposed measure called the $C$ score.

However, these approaches rely heavily on the topology structure, and they do not incorporate the specific characteristics of social networks, such as social hierarchy and node centrality. Furthermore, most of the proposed methods are designed to deal with full partitions, i.e., they are not suitable for a single community. In this paper, we present a framework for calculating the significance of a social community. The framework does not embrace the universal approach, but instead it tries to focus on the unique properties of social networks. We model the dynamics of social interactions by identifying social leaders and corresponding hierarchical structures, as social communities are formed around those leaders. Instead of a direct comparison with the average outcome of a random model, we compute the similarity of a given node with its leader by using the number of common neighbors. To determine the membership vector, an efficient community detection algorithm is proposed based on the position of nodes and their corresponding leaders. Then, using the log-likelihood score, the tightness of the community can be derived. Based on the distribution of community tightness, a "$p$-value" form significance measure is proposed for community structure analysis. Finally, we apply our framework to both benchmark networks and real social networks. Experimental results show that it can be used to (i) determine the optimal number of a community; (ii) analyze the social significance of a given community; (iii) compare the performance among various algorithms; etc.

## II. THE FRAMEWORK

Real social networks have their own specific characteristics, which are essential to define the significance of community structure. In this section, we discuss these important characteristics, and we provide a detailed introduction of the framework.

### A. Social hierarchy and community leader

It is natural to relate social networks with hierarchical structure [29]. In one such hierarchy there are nodes that are more important and influential than other nodes, hence they are located on a higher level in the hierarchy [see Fig. 1(a)]. The leaders should have two properties: they should be well connected to the members of their group, and they should be able to communicate with other leaders when necessary. If the distributed algorithm is carried out in each group separately and the leaders communicate at a higher level, the nodes can enjoy a faster convergence rate.

Hierarchical structure and leader nodes also exist in almost all real social networks. As an example, in the famous Karate network [30] there are two significant leaders (nodes 1 and 33), and communities are built around those leaders [see Fig. 1(b)]. The removal of those leaders will result in splitting those communities, since the leaders are keeping the communities together. Since the hierarchies are a consequence of the spreading of correlation, as are the communities, we believe that the identification of these hierarchies in a network will result in a natural community detection. The area in which a leader has the most influence should define its community. Therefore, community detection can be performed by finding all natural leaders and all nodes that they influence.

There are two representative ways to define the leader nodes: the first is as a *degree leader*, which is the most natural

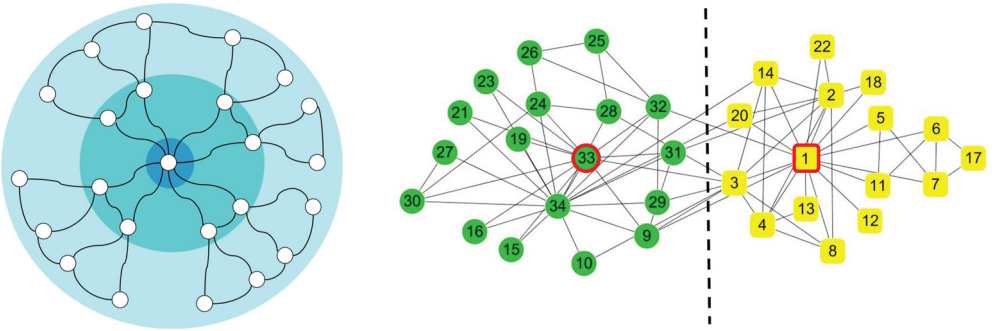



FIG. 1. (Color online) (a) Social hierarchy within a community. The leader is located on the highest level, representing the most influential node. Circles depict different levels in the hierarchy, with the darkest color denoting the highest level. (b) The Zachary karate network. Different communities are represented by different colors and shapes. Leaders with numbers 1 and 33 are highlighted in the original graph.

way in that it enables nodes with the largest degree (number of edges linked with it) to be the leader nodes; the second is as an *influence leader*, which uses the notion of relative influence defined in [31]. The influence represents how important the opinion of a given node is to its neighborhood. A leader is the node with the biggest overall influence, since the overall influence represents how close a node is to the core of its community, as well as its actual potential of becoming a leader. Also, a leader should have a bigger influence on its neighbors than they have on it. Therefore, we define leaders as those nodes for which the product (overall influence) × (relative influence) is large. More precisely, we denote the relative influences between the nodes as $T_{ij}$, and the overall influences of the nodes as $u_i^*$. $T_{ij}$ is defined as

$$T_{ij} = \frac{a'_{ij}}{\sum_k a'_{kj}}, \tag{1}$$

where $a_{ij}$ is the adjacent matrix element, $a'_{ji} = a_{ji} + \sum_k C^k_{ji}$, and $C^k_{ji} = \min\{a_{ki}, a_{jk}\}$. The overall influence of nodes $u_i^*$ is defined as

$$u_i^* = \sum_j T_{ij} = \sum_j \frac{a'_{ij}}{\sum_k a'_{kj}}. \tag{2}$$

Node $x_i$ is a leader if $T_{ij}u_i^* > T_{ji}u_j^*$ for all $x_j$. The product $T_{ij}u_i^*$ combines the relative influence of node $x_i$ toward node $x_j$ with the overall influence of node $x_i$.

### B. Identification of a community based on leaders

In this step, our goal is to devise a scheme to provide each node with a small vector that includes compact global information on how the node is located with respect to the leader nodes. We provide a definition for the membership vector based on the properties of random-walk dynamics on graphs. Consider a graph with $c$ leaders $l_1, l_2, \ldots, l_c$ and $N - c$ regular nodes. Given the leaders and the arbitrary order assigned to them, we describe the algorithm to determine the membership vectors for each regular node. We denote the membership vector of node $i$ by $\mathbf{y}_i = (y_i^1, y_i^2, \ldots, y_i^c) \in R^a$. By $y_i^k(t)$, we mean the $k$th entry of the influence vector of node $x_i$ evaluated at time $t$.

The procedure operates as follows. The membership vector of leader $l_i$ is first assigned to be the unit vector. These $c$ vectors do not vary. For regular node $x_i$, $y_i^k$ is initialized randomly, and then distributed uniformly on $[0,1] (k = 1,2,\ldots,c)$. Then we normalize each row of $\mathbf{y}_i$ so that for all leader $k$, the sum of $y_i^k$ is 1. At each iteration time $t$, the influence vector of each regular node $x_i$ is updated entrywise $(k = 1,2,\ldots,c)$ using the following rule:

$$y_i^k(t+1) = \frac{1}{\sum_j a_{ij} + 1}\left[ y_i^k(t) + \sum_j a_{ij} y_j^k(t) \right], \tag{3}$$

where $A = \{a_{ij}\}$ is the adjacency matrix in which $a_{ij} = 1$ if nodes $x_i$ and $x_j$ are connected, and $a_{ij} = 0$ otherwise.

We notice that, for all time $t$, $\sum_k y_i^k(t) = 1$. Equation (3) is equivalent to $Y(t+1) = PY(t) = (I + D)^{-1}(A + D)Y(t)$, where $P = (I + D)^{-1}(A + D)$ is a stochastic walk matrix. Actually, the influence of leader nodes $l_k$ $(k = 1,2,\ldots,c)$ on any regular node $x_i$, $y_i^k$, is the probability that a random walker that starts from $x_i$ hits $l_k$ before it hits any other leader node [9,10]. If the underlying graph is connected, the iteration $\lim_{t\to\infty} y_i(t)$ converges to a set of unique vectors, and these vectors can naturally be represented as the probability that a regular node belongs to the community with a given leader node. The membership vector in this probability form can be used to uncover soft communities with overlapping nodes. As a result, although leadership of a node only contains local information, random-walk dynamics can be used to gain membership containing a global view of the whole graph. The performance has been tested on both GN and LFR benchmarks in Sec. IV, which verify the efficiency of our algorithm.

### C. Node similarity

Nodes with large amounts of different neighbors are considered very "far" from each other. Alternatively, one could measure the similarity as the overlap between the neighborhoods $\Gamma(i)$ and $\Gamma(j)$ of nodes $x_i$ and $x_j$, given by the ratio between the intersection and the union of the neighborhoods, i.e.,

$$\mathrm{sim}(x_i, x_j) = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}. \tag{4}$$

Using this similarity measure, one can compute the expected similarity of elements to the community leader $z$, given the similarity measure $\mathrm{sim}(x,z)$,

$$E[\mathrm{sim}(x,z)] = \int_{\mathbb{R}^M} \mathrm{sim}(x,z) Q(x|z) dx, \tag{5}$$

where $Q(x|z)$ is a distribution of nodes in a community with leader $z$. Using the maximum entropy principle, we obtain a statistically unbiased distribution fulfilling constraint,

$$Q(x|z,\eta) = \frac{1}{Z_\eta} P_0(x) e^{\eta\, \mathrm{sim}(x,z)} dx. \tag{6}$$

The background distribution $P_0(x)$ is contrasted with an alternative hypothesis: node $x$ being part of a community, and a group of nodes distinguished by enhanced mutual similarity. The normalization constant $Z_\eta$ depends on the value of the *scoring parameter* $\eta$. Parameter $\eta$ is in a one-to-one relationship with the value of $E[\mathrm{sim}(x,z)]$, the expected similarity $\mathrm{sim}(x,z)$ of vectors following distribution $Q(x|z,\eta)$. This relationship can be described as

$$\frac{\partial}{\partial \eta} \log Z_\eta = E[\mathrm{sim}(x,z)]. \tag{7}$$

In other words, parameter $\eta$ determines the community's "width" in the same way that the corresponding constant $Z_\eta$ does. Intuitively, the larger the value of $\eta$, the smaller the expected width of the community. We will thus refer to $\eta$ as the width parameter. Note that when $\eta = 0$, the distribution $Q(x|z,\eta)$ is the same as the background model $P_0(x)$.

### D. Log-likelihood score and community tightness

The deviations of the community distribution from the null model define the log-likelihood score, which takes the simple

form

$$s(x|z,\eta) \equiv \log \frac{Q(x|z,\eta)}{P_0(x)} = \eta\, \mathrm{sim}(x,z) - \log Z_\eta. \quad (8)$$

Using Eq. (8), the log-likelihood score assigns positive score values to nodes that are more likely to be in a community with center $z$ and scoring parameter $\eta$ than in the null background model. The exact form of the scoring function depends on the similarity measure $\mathrm{sim}(x,z)$ and, via the normalization constant $Z_\eta$, on the background model $P_0(x)$.

Given a community with a node set $\{x_1,\ldots,x_N\}$, for a given leader $z$ and a scoring parameter $\eta$, the log-likelihood scores $s(x_i|z,\eta)$ are positive. The community tightness is the sum of the scores of the community elements,

$$S(x_1,\ldots,x_N|z,\eta) = \sum_i \max[s(x_i|z,\eta),0]. \quad (9)$$

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 have comparable tightness to looser but larger communities.

### E. Distribution of community tightness

To describe the statistics of an arbitrary tightness score $S(x_1,\ldots,x_N)$ for nodes drawn independently from the distribution $P_0(x)$, we consider the quality function

$$Z(\beta) = \Pi_{i=1}^N \int dx_i\, P_0(x_i) e^{\beta S(x_1,\ldots,x_N)} = \int dS\, p(S) e^{\beta S}. \quad (10)$$

Next, we introduce the computation procedure of $p(S)$. In the collection of all configurations of node set $X$ with energy $E$, $p(E)$ denotes the density of states as a function of energy $E$. Replacing the extensive energy with the intensive quantity, $E = Ne$, and using $p(E) = \frac{1}{N} p(e)$, we get

$$\int p(E) e^{-\beta E} dE = \frac{1}{N} \int e^{-N\beta e + \log p(e)} \simeq \frac{1}{N} e^{N \sup_e [\log p(e)/N - \beta e]}. \quad (11)$$

In next step, assuming $N$ is large, we use the saddle-point approximation and get

$$\log Z(\beta)/N = \sup_e [\log p(e)/N] - \beta e, \quad (12)$$

i.e., the normalized logarithm of the partition function, $\log Z(\beta)/N = -\beta f(\beta)$, is a Legendre transform of the normalized logarithm of the probability, $\log p(e)/N$. Exploiting the duality of the Legendre transform, we get

$$\log p(e) \simeq -N \sup_\beta [\beta f(\beta) + \beta e] = N[\beta_0 e - \beta_0 f(\beta_0)], \quad (13)$$

with $\beta_0$ the saddle point of the function in the squared brackets. Then, there is

$$\log p(E) = \log p(e) + \log\left(\frac{1}{N}\right) \simeq N[\beta_0 e - \beta_0 f(\beta_0)] + \log\left(\frac{1}{N}\right). \quad (14)$$

Using the conclusion derived above, given all configurations of the node set $X = (x_1,\ldots,x_N)$ with a community tightness $S$, $p(S)$ denotes the density of states as a function of tightness $S$. Asymptotically for large $N$, this density can be extracted from $Z(\beta)$ based on Eq. (10) as

$$\log p(S) \simeq N\Omega(s) - \tfrac{1}{2}\log(gN). \quad (15)$$

Here $\Omega(s)$ is the entropy as a function of the tightness per element, i.e., $\Omega(s) = -\max_\beta[f(\beta) + \beta s]$. $\beta f(\beta) = -\log Z(\beta)/N$ is the free-energy density. The distribution of community tightness $S$ is defined as the probability $\int_S^{+\infty} p(S')dS'$ to find a score larger than or equal to $S$. This is a typical $p$-value form, and it can be used to represent the statistical significance directly.

## III. SIGNIFICANCE OF SOCIAL COMMUNITIES

The quality of an insignificant community can also be quantified with a community tightness function, yielding some score $S_0$. To distinguish the true and random communities, we need to characterize the distribution of the tightness score $p(S)$ from the background distribution. The statistical significance of score $S_0$ is then defined in a "$p$-value" form [32] as the probability that a random chosen node set contains a community with a score greater than or equal to $S_0$. In the statistical significance analysis, we proceed as follows: given a group of nodes with some score $S_0$, we formulate a null hypothesis: "These nodes are drawn from the background distribution." To test this hypothesis, we compute the statistical significance of score $S_0$: a low value suggests that the null hypothesis is unlikely and allows for rejecting it. Importantly, a low value does not yet say that the group of nodes is indeed a significance community. A low value provides a necessary but not a sufficient condition in this direction.

However, the scoring parameter $\eta$ is hard to determine. We now rewrite the community tightness function of Eq. (9) and simplify it as

$$S(x_1,\ldots,x_N|z,\eta) = \sum_{i=1}^N \max[s(x_i|z) - \mu, 0], \quad (16)$$

where $s(x_i|z) = \mathrm{sim}(x_i,z)$. Through this transform, the width of the community can be determined simply by parameter $\mu$. If the size of the network is large enough, using the mean-field theorem, $s_i = s(x_i|z)$ is approximately Gaussian-distributed with variance $M$, $P[s(x_i|z)] = \sqrt{1/(2M\pi)}\exp\{-s^2/(2M)\}$. Computation of the distribution of the tightness $S$ is straightforward from the derivation shown in Sec. II, and it requires calculation of the quality function:

$$\begin{aligned} Z_c(\beta,\mu) &= \int_{\mathbb{R}^N} e^{\beta S(x_1,\ldots,x_N|z,\eta)} P(s_1)\cdots P(s_N) ds_1 \cdots ds_N \\ &= \left[\int_{-\infty}^{+\infty} e^{\beta \max[s_i-\mu,0]} P(s) ds\right]^N \\ &= \left[\int_{-\infty}^{\mu} P(s) ds + \int_{\mu}^{+\infty} e^{\beta(s_i-\mu)} P(s) ds\right]^N \\ &= \{[1 - H(\mu)] + e^{\frac{(\beta)^2}{2} - \beta\mu} H(\mu - \beta)\}^N, \end{aligned} \quad (17)$$

with $H(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ the complementary cumulative Gaussian distribution. In Eq. (17), the integration is divided into two intervals: 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 the score threshold, 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

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

and the entropy is

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

As described in Sec. II,

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

### A. Significance score

For a given community, the significance score $F$ can be calculated using the probability that the community tightness $S$, $p(S)$, is greater than or equal to $S$,

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

Furthermore, from a global perspective, we use the average significance score $\langle F\rangle_q$ to indicate the robustness of a partition corresponding to $q$ communities, defined as the average value among $F$ values of all $q$ communities partitioned by a particular algorithm. Since $\langle F\rangle_q$ tries to directly characterize the social significance of a specific network partition, it is very convenient to estimate the performance and function property of a given algorithm.

### B. Computational complexity

The calculation of the significance score mainly contains three steps: (i) calculate the degree or influence of every node to find the leaders of the communities; (ii) identify the communities in the network based on the positions of the nodes and the leaders; and (iii) measure the similarity between nodes with their corresponding leaders and calculate the significance score using the distribution of tightness. The computational complexity of our method depends on the highest complexity of these three steps. Obviously, part (ii) is of the highest complexity, while the complexity of the other two parts is rather low. For part (ii), the computational complexity is $O(N^2)$, where $N$ is the number of nodes in the network. Thus, we obtain that the cost of the whole algorithm is $O(N^2)$. Our method is very easy to implement and suitable for a lot of large-scale real networks.

## IV. EXPERIMENTS

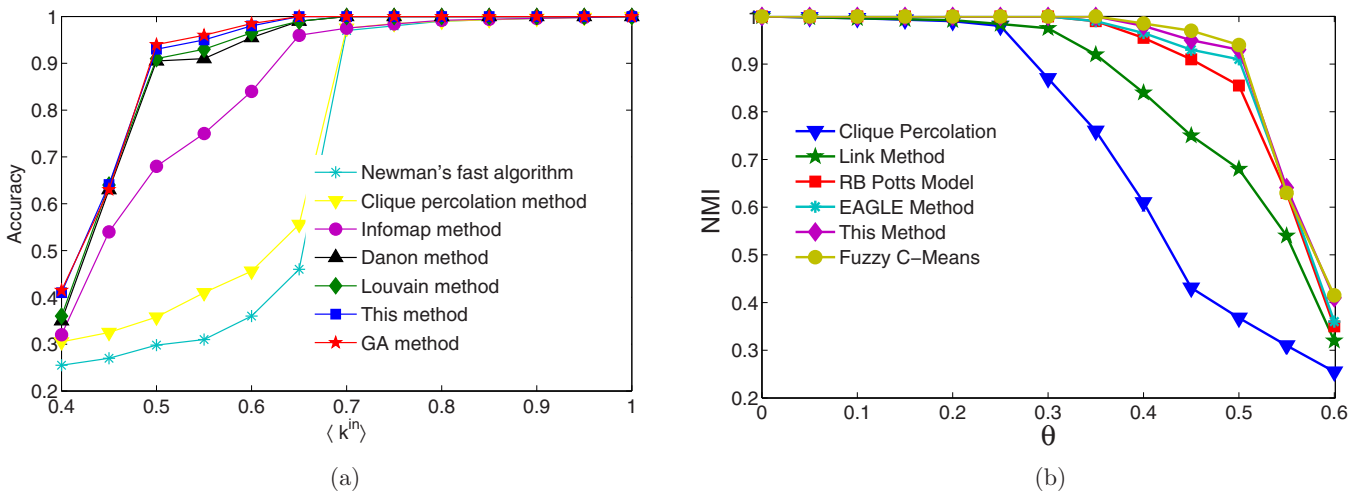
In this section, we will test the validity of our framework. Experiments are designed and implemented for two main purposes: (i) to evaluate the performance of a given algorithm, and (ii) to apply it on both artificial benchmark networks and real social networks.

### A. Benchmark network

#### 1. GN benchmark network

First, we use the classical GN benchmark presented by Girven and Newman [33]. Each network has $n = 128$ nodes that are divided into 4 communities with 32 nodes each. Edges between two nodes are introduced with different probabilities that depend on whether the two nodes belong to the same community or not, and the average degree $\langle k\rangle = 16$. Every node is connected on average with $\langle k^{\text{in}}\rangle$ nodes of its own group and $\langle k^{\text{out}}\rangle$ of the rest of the network. The total degree of each node is always kept constant and equal to $k = \langle k^{\text{in}}\rangle + \langle k^{\text{out}}\rangle$. Each group represents a well-defined community up to $\langle k^{\text{out}}\rangle = 8$, but actually communities start to become very fuzzy at lower values of $\langle k^{\text{out}}\rangle \approx 8$ due to statistical fluctuations.

We empirically demonstrate the effectiveness of our algorithm via a comparison with six other well-known algorithms on the GN networks. These algorithms include Newman's fast algorithm [1], Danon *et al.*'s method [34], the Louvain




(a) (b)

FIG. 2. (Color online) The performance of the community detection algorithm based on leaders in both the GN and LFR networks. (a) A comparison of the accuracy with six famous algorithms in the GN benchmark network. Here, accuracy is defined as the fraction of nodes correctly clustered. (b) A comparison of NMI with five fuzzy algorithms in the LFR benchmark network.

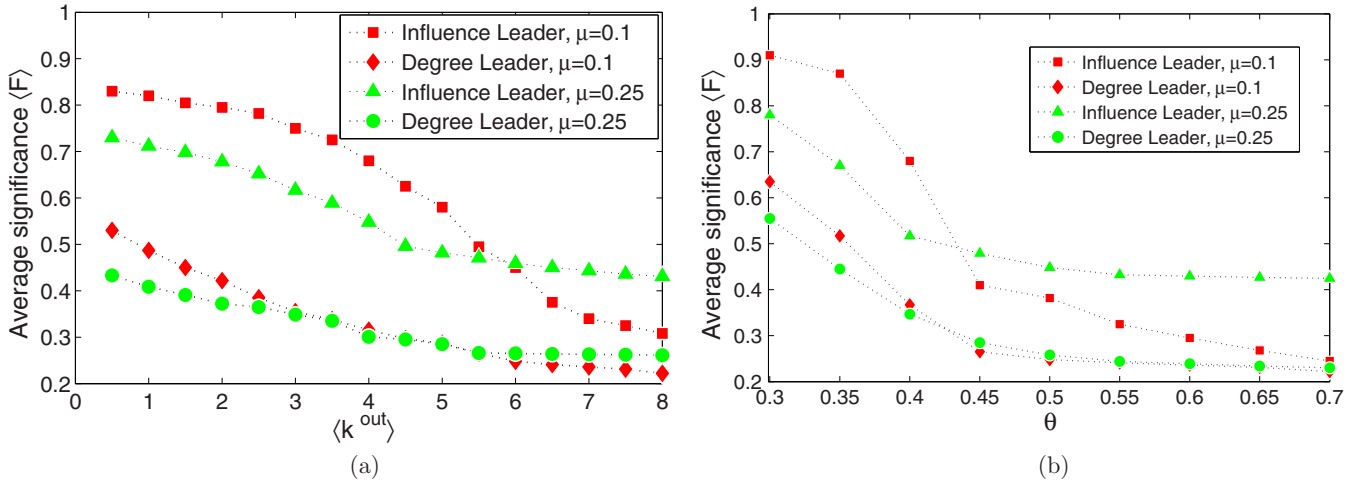



FIG. 3. (Color online) The performance of social significance $\langle F\rangle$ on both the GN and the LFR network. (a) In the GN network, $\langle F\rangle$ decreases with increasing $\langle k^{\text{out}}\rangle$. When the community structure of the network is very clear, $\langle F\rangle$ is very close to 1; when the network has almost no community structure, $\langle F\rangle$ is close to 0.3. This implies that for a given network, when $\langle F\rangle$ is less than $0.3(\langle k^{\text{out}}\rangle \approx 8)$, it is not safe to say that there exists a significant community structure. (b) In the LFR benchmark, the average degree $k = 20$, the maximum degree is 50, and $P(k) \propto k^{\gamma}$. Maximum and minimum community sizes are 50 and 20, respectively. With an increase of the mix parameter $\theta$, the $\langle F\rangle$ index decreases. When $\theta \geqslant 0.5$ (no significant community), $\langle F\rangle$ is near 0.3, which is similar to the GN network.

method [35], Infomap [36], the clique percolation method [17], and the GA method [37]. Figure 2(a) presents the experimental results, in which the $y$ axis denotes the fraction of nodes correctly clustered, and each point on the curves is obtained by testing them against 50 synthetic networks shuffled from the original network. As we observe, all algorithms work well when $\langle k^{\text{in}}\rangle$ is larger than 0.7 with accuracy larger than 0.95. Compared with the other six algorithms, our algorithm outperforms the other algorithms overall, and its accuracy is only slightly worse than that of the GA in the case of $0.5 \leqslant \langle k^{\text{in}}\rangle \leqslant 0.65$.

As is well known, the communities become fuzzier and thus more difficult to identify when $\langle k^{\text{out}}\rangle$ increases. Hence, the significance of the community structure will also tend to be weaker and the $F$ index will decrease. The numerical results of the $F$ value corresponding to both the degree leader and the influence leader are shown in Fig. 3(a). We find that the index $F$ works well in the GN benchmark: when community structure is very clear, $\langle F\rangle$ is very close to 1; when the network is nearly a random one, the corresponding $\langle F\rangle$ is near 0.2–0.3. Moreover, by comparing two kinds of leaders, we observe that $\langle F\rangle$ values corresponding to the influence leader are larger than those corresponding to the degree one, and therefore they are more effective. Furthermore, the topology becomes fuzzier when $\langle k^{\text{out}}\rangle$ increases, and the sizes of the communities will correspondingly become smaller and smaller. At the same time, as the width parameter $\mu$ increases, the significance will favor tighter communities with fewer elements. As a result, in Fig. 3(a), the value of $\langle F\rangle$ corresponding to $\mu = 0.25$ will be larger than that corresponding to $\mu = 0.1$ when $\langle k^{\text{out}}\rangle$ is larger than 6. We argue that for a given network when the corresponding $\langle F\rangle$ is larger than $0.3(\langle k^{\text{out}}\rangle \approx 8)$, there exists a significant community structure. Thus, the larger the $\langle F\rangle$ index is, the more significant the community structure will be.

### B. LFR benchmark network

We also test the index on the more challenging LRF benchmark presented by Lancichinetti, Fortunato, and Radicchi [16]. In the LFR benchmark, each node is given a degree obtained from a power-law distribution with an exponent $\gamma$, and the sizes of the communities are obtained from a power-law distribution with an exponent $\beta$. Moreover, each node shares a fraction $1-\theta$ of its links with other nodes of its community and a fraction $\theta$ with other nodes in the network; $\theta$ is the mixing parameter.

We compared with five other well-known soft community partition algorithms in the LFR networks, including the Clique percolation method [17], the Link method [38], the EAGLE method [39], the RB Potts model [40], and the Fuzzy C-Means method [41]. To evaluate a community detection algorithm, the normalized mutual information (NMI) [15,16,28–31,33] is utilized to evaluate the partition found by each algorithm. The experimental results are displayed in Fig. 2(b), where the $y$ axis represents the value of NMI, and each point in the curves is obtained by averaging the values obtained on 50 synthetic networks sampled from the above model. As we observe, all algorithms work very well when $\theta$ is less than 0.3, with NMI larger than 0.85. Compared with the five other algorithms, our algorithm performs quite well and its accuracy is only slightly worse than that of the Fuzzy C-Means in the case of $0.35 \leqslant \theta \leqslant 0.5$.

The significance of community structure can be adjusted by $\theta$ in LFR benchmark. The numerical results in the LFR benchmark are shown in Fig. 3(b). We observe that $F$ decreases with the increase of $\theta$. As in the GN network, the $F$ values corresponding to the influence leader are larger than those corresponding to the degree leader when $\theta$ is low. Furthermore, from Fig. 3(b) we notice that the value of $\langle F\rangle$ corresponding to $\mu = 0.25$ is larger than that corresponding to $\mu = 0.1$ when $\theta$ is larger than 0.43.

### C. Stochastic block model

Recently, many algorithms [42,43] have been proposed to detect communities in networks or dynamical networks based on the famous stochastic block model (SBM) first proposed by Holland *et al.* [44] and extended by Decelle *et al.* [12,45] and Zhang *et al.* [8,14], in which the connectivity between blocks is defined in terms of probabilities. In this model, each node $i$ has a hidden label, $t_i \in \{1,\ldots,q\}$, specifying which of the $q$ groups it is a member of. These labels are chosen independently, where $y_a$ is the probability that a given node has label $a \in \{1,\ldots,q\}$ (normalized so that $\sum_{a=1}^{q} y_a = 1$). If $N_a$ is the number of nodes in each group, we have $y_a = \lim_{N\to\infty} N_a/N$. Once the group assignment is chosen, the model generates a graph $G$ as follows. For each pair of nodes $i,j$ with $i < j$, we put an edge between $i$ and $j$ independently with probability $p_{t_i,t_j}$, leaving them unconnected with probability $1 - p_{t_i,t_j}$. We call $p_{ab}$ the affinity matrix. Since we are interested in the sparse case in which $p_{ab} = O(1/N)$, we will use the rescaled affinity matrix $c_{ab} = Np_{ab}$ and assume that $c_{ab} = O(1)$ in the limit $N \to \infty$. Our goal is to learn the parameters $q,\{y_a\},\{p_{ab}\}$ of the block model, as well as the true group assignments $\{ti\}$. Special cases of this model have often been considered in the literature. Planted partitioning, when $y_a = 1/q, c_{ab} = c_{\text{out}}$ for $a \neq b$ and $c_{aa} = c_{\text{in}}$ with $c_{\text{in}} > c_{\text{out}}$, is a classical problem in computer science, and it has been used as a benchmark for community detection. Here, $\varepsilon = c_{\text{out}}/c_{\text{in}}$ is used to control the fuzziness of a generated network.

To test the performance on sparse networks, we establish a large network generated by stochastic block model with low average degree. Figure 4 shows the case of a network with $N = 5000$ nodes and $q = 10$ groups, with an average degree

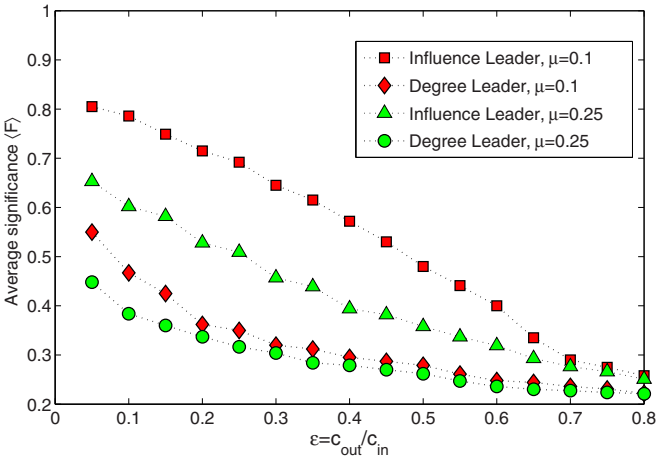
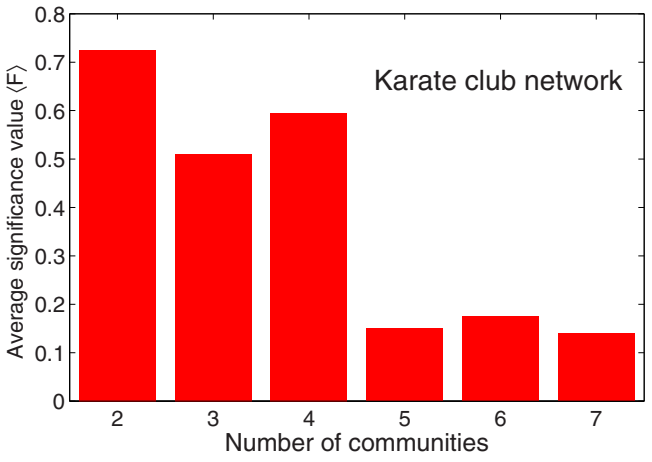
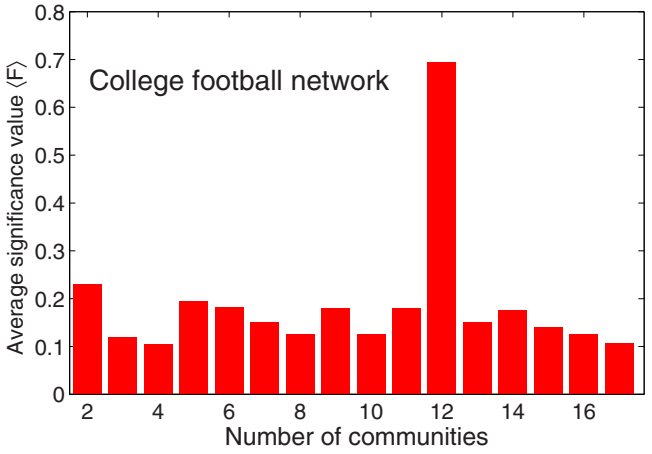


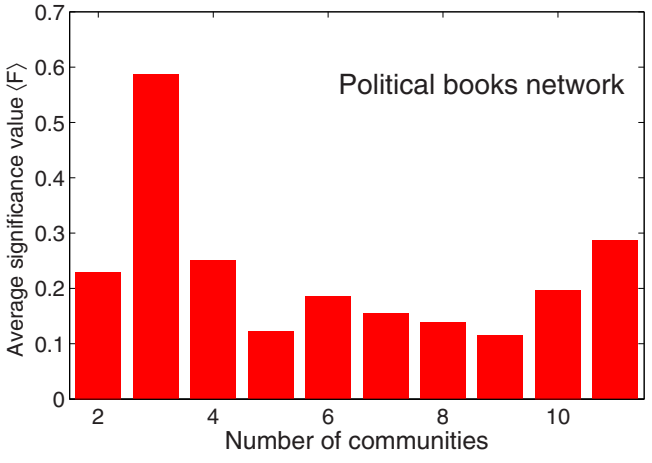

FIG. 4. (Color online) The performance of social significance $\langle F\rangle$ on the stochastic block model. In this example, there are $N = 5000$ nodes and $q = 10$ groups. The average degree $c = 8$, and the parameter $\varepsilon = c_{\text{out}}/c_{\text{in}}$ is used to control the fuzziness of the generated network. Each point on the curves is obtained by testing 50 times. With an increase of $\varepsilon$, the $\langle F\rangle$ index decreases. When $\varepsilon$ is close to 0.8, the network is nearly a random one, and the corresponding $\langle F\rangle$ values of both kinds of leaders are very low, near 0.2–0.3.




(a)

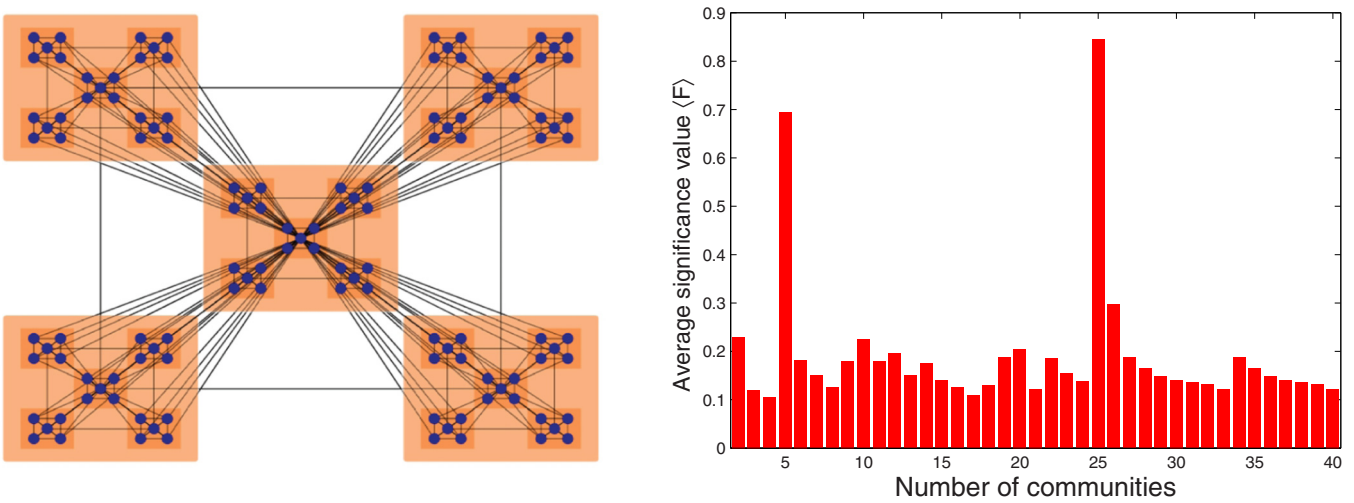



(b)




(c)

FIG. 5. (Color online) The empirical results of the optimal number of communities in the Zachary karate club network, the College football network, and the Political books network. From the plots we observe that $\langle F\rangle$ achieves its highest value when the community numbers correspond with reality: the Zachary karate club has two optimal communities, the College football network has 12 optimal communities, and the Political books network has three optimal communities.



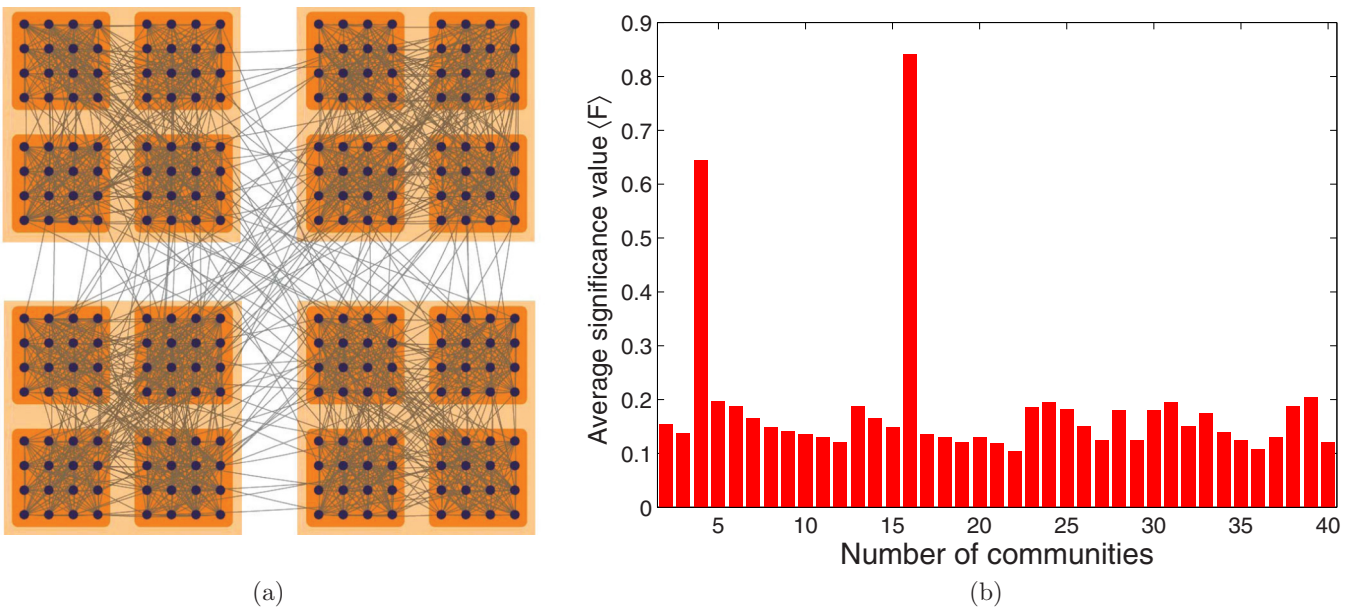

(a) (b)

FIG. 6. (Color online) (a) The structure of $RB125$, with 25 dense communities and 5 sparse communities, is highlighted in the original network. (b) The number of communities vs the average significance value $\langle F \rangle$.

$c = 8$. Each point on the curves is obtained by testing 50 times. We find that when $\varepsilon$ is close to 0, the community structure is very clear and the corresponding $\langle F \rangle$ value is close to 1. In contrast, when $\varepsilon$ is close to 0.8, the network is nearly a random one, and the corresponding $\langle F \rangle$ values of both kinds of leaders are very low, near 0.2–0.3. Furthermore, it can be observed that the value of $\langle F \rangle$ corresponding to $\mu = 0.25$ will be larger than that corresponding to $\mu = 0.1$. Specifically, we argue that for a given network when the corresponding $\langle F \rangle$ is larger than 0.32 ($\varepsilon \approx 0.4$), there exists a significant community structure that may be detectable [12]. Therefore, $F$ shows a great ability to characterize the significant modular structure as we adjust the parameter $\varepsilon$.

### D. Real network

Now we show the utility and versatility of our method for the statistical evaluation of communities in real social networks. The significance corresponding to the influence leader is used in this section. First, we find that the optimal number of community $c$ can be determined using the average significance score $\langle F \rangle_q$. For many real-world social networks, we do not know the number of communities before incorporating additive information, and the community structure will be clearest when the number is the optimal $c$. The detailed steps are as follows: (i) The degree or influence of each node is calculated and ranked. We choose the first $q$ nodes with the largest influence as leaders. (ii) We partition




(a) (b)

FIG. 7. (Color online) (a) The structure of $H13\text{-}4$, with 16 dense communities and 4 sparse communities, is highlighted in the original network. (b) The number of communities vs the average significance value $\langle F \rangle$.

the network and obtain $q$ communities, using the proposed community detection algorithm based on leaders. For each $q(1 \leqslant q \leqslant N/2)$, a specific partition with $q$ communities can be obtained. (iii) We apply our method and use $\langle F\rangle_q$ to compute the significance of the $q$ community structure. (iv) In comparison, the $q$ corresponding to the largest value of $\langle F\rangle_q$ is chosen as the optimal number of communities.

Here, three famous real examples are considered: the Zachary Karate club network [30], the College football network [33], and the Political books network [17]. The community partition of all networks has been obtained by our method in Sec. II. As shown in Fig. 5, the corresponding community numbers with the largest $\langle F\rangle_q$ are the optimal $c$ of every network. These examples show the great ability of our framework in characterizing the modular structure of the real networks. Then we analyze the partition and find that the $F$ score of the communities found is quite high. However, there are a few exceptions for which $F$ is sufficiently low, but most of the groups are not statistically significant. This occurs because the algorithm is forced to place all the nodes in some group. Since these three networks are sparse and the modularity is not strong, especially the football network, the results are very precise and verify that our framework is effective for real social networks.

Furthermore, to show that the model can uncover hierarchical structures in different scales; Figs. 6 and 7 give two examples of multilevel community structures. Figure 6(a) shows the $RB125$ network, which is a hierarchical scale-free network proposed by Ravasz and Barabási in [18]. The regions corresponding to 5 and 25 modules are the most representative in terms of resolution. Next, $H13$-4 proposed by Arenas *et al.* [19] is shown in Fig. 7(a). It is a homogeneous degree network with two predefined hierarchical scales. The first hierarchical level consists of 4 modules of 64 nodes, and the second level consists of 16 modules of 16 nodes. The partition of both levels is highlighted in the original networks.

In both examples, the significance of such levels can be quantified by their corresponding $\langle F\rangle_q$. The largest value reveals the actual number of hierarchical levels hidden in a network. From Figs. 6(b) and 7(b), we observe that 25 and 16 are the optimal numbers of communities in $RB125$ and $H13$-4 networks having the largest value, respectively. However, five modules and four modules are also reasonable partitions that show the fuzzy level of the hierarchical networks. These results are consistent with the generation mechanisms and hierarchical patterns of these two networks.

Finally, we show that significance can also be used to rank the partitions obtained by different algorithmic strategies. The Zachary Karate club network, the College football network, and the Political books network are employed as examples. Table I presents the results estimated from three algorithms chosen for their simplicity, which are all able to automatically select the number of communities: the label propagation method [20], the Wu-Huberman linear time method [21], and the Girvan-Newman betweenness algorithm [33]. Here, first we partition the network into communities using a specific method. For each community, the node with the largest influence is chosen as the leader. Then, the similarity between nodes and their corresponding leaders is measured. Finally, $\langle F\rangle$ is calculated for each algorithm. From Table I, we observe that the $\langle F\rangle$ values of all three examples are not high, due to the fuzziness and sparseness of the network's topology. However, the $\langle F\rangle$ value from the Girvan-Newman algorithm is higher than that from the other two methods, since the mechanism of the Girvan-Newman algorithm is objective function optimization. In contrast, the other two algorithms emphasize the simplicity of calculation too much while ignoring the accuracy of the results. These observations are not evidence of the overall superiority of one method over another, rather they are an example of how to compare the significance and use the different partitioning algorithms in a given network.

TABLE I. Comparison of various algorithms with $\langle F\rangle$ values.

| Networks | Algorithms | Values of $\langle F\rangle$ |
|---|---|---|
| Zachary network | label propagation method | 0.641 |
| | Wu-Huberman linear time method | 0.627 |
| | Girvan-Newman algorithm | 0.735 |
| College football network | label propagation method | 0.602 |
| | Wu-Huberman linear time method | 0.631 |
| | Girvan-Newman algorithm | 0.758 |
| Political books network | label propagation method | 0.581 |
| | Wu-Huberman linear time method | 0.617 |
| | Girvan-Newman algorithm | 0.698 |

## V. CONCLUSION

In summary, we presented a framework for calculating the significance of a social community. Our framework does not embrace the universal approach, but instead it tries to focus on the unique properties of social networks. Based on the distribution of community tightness, a "$p$-value" form significance measure is proposed for network analysis. We apply our framework to both a benchmark network and a real social network, and its efficiency has been demonstrated and verified both theoretically and experimentally. Important information related to social community structures can be mined from the significance trend, such as the social significance of a given community, the optimal number of communities, and the performance among various algorithms in detecting a meaningful community structure.

## ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their valuable suggestions, which were very helpful in improving the manuscript. The authors are supported by NSFC Grants No. 71401194, No. 91324203, and No. 11131009 (H.-J.L.), and the "121" Youth Development Fund of CUFE Grant No. QBJ1410 (J.J.D.).

[1] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[2] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[3] M. E. J. Newman, Proc. Natl. Acad. Sci. (USA) **103**, 8577 (2006).
[4] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. (USA) **104**, 36 (2007).
[5] A. Arenas, A. Diaz-Guilera, and C. J. Perez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006).
[6] J. C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. (USA) **107**, 12755 (2010).
[7] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, PloS One **6**, e18961 (2011).
[8] P. Zhang and C. Moore, arXiv:1403.5787.
[9] H. J. Li, Y. Wang, L. Y. Wu, J. Zhang, and X. S. Zhang, Phys. Rev. E **86**, 016109 (2012).
[10] H. J. Li and X. S. Zhang, Eur. Phys. Lett. **103**, 58002 (2013).
[11] H. Zhou, Phys. Rev. E **67**, 041908 (2003).
[12] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[13] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[14] P. Zhang, C. Moore, and L. Zdeborová, Phys. Rev. E **90**, 052802 (2014).
[15] X. S. Zhang, R. S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang, and L. Chen, Eur. Phys. Lett. **87**, 38002 (2009).
[16] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[17] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[18] E. Ravasz and A. L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[19] A. Arenas, A. Fernandez, and S. Gomez, New. J. Phys. **10**, 053039 (2008).
[20] U. N. Raghavan, R. Albert, and S. Kumara, Phys. Rev. E **76**, 036106 (2007).
[21] F. Wu and B. A. Huberman, Eur. Phys. J. B **38**, 331 (2004).
[22] Z. P. Li, S. H. Zhang, R. S. Wang, X. S. Zhang, and L. Chen, Phys. Rev. E **77**, 036109 (2008).
[23] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101 (2004).
[24] J. Reichardt and S. Bornholdt, arXiv:cond-mat/0606220.
[25] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Acad. Sci. (USA) **104**, 15224 (2007).
[26] B. Karrer, E. Levina, and M. E. J. Newman, Phys. Rev. E **77**, 046119 (2008).
[27] G. Bianconi, P. Pin, and M. Marsili, Proc. Natl. Acad. Sci. (USA) **106**, 11433 (2009).
[28] A. Lancichinetti, F. Radicchi, and J. J. Ramasco, Phys. Rev. E **81**, 046110 (2010).
[29] H. J. Li, Y. Wang, L. Y. Wu, Z. P. Liu, L. Chen, and X. S. Zhang, Eur. Phys. Lett. **97**, 48005 (2012).
[30] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[31] A. Stanoev, D. Smilkov, and L. Kocarev, Phys. Rev. E **84**, 046102 (2011).
[32] J. D. Wilson, S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel, Ann. Appl. Stat. **8**, 1853 (2014).
[33] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. (USA) **99**, 7821 (2002).
[34] L. Danon, J. Duch, D. Guilera, and A. Arenas, J. Stat. Mech. (2005) P09008.
[35] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. (2005) P10008.
[36] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. (USA) **105**, 1118 (2008).
[37] R. Guimera and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[38] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).
[39] H. Shen, X. Cheng, K. Cai, and M. B. Hu, Physica A **388**, 1706 (2009).
[40] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[41] S. Zhang, R. S. Wang, and X. S. Zhang, Physica A **374**, 483 (2007).
[42] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[43] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, Mach. Learn. **82**, 157 (2011).
[44] P. W. Holland, K. B. Laskey, and S. Leinhardt, Soc. Netw. **5**, 109 (1983).
[45] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Phys. Rev. E **84**, 066106 (2011).