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, CA 94305, USA

(Dated: March 30, 2015)

Abstract

Community structure analysis is a powerful tool for social networks, which 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 community structure partitioned is an urgent and important question. In this paper, integrating the specific characteristics of real society, we present a novel framework analyzing the significance of social community specially. 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 nodes and their corresponding leaders. Then, using log-likelihood score, the tightness of community can be derived. Based on the distribution of community tightness, we establish a new connection between $p$-value theory and network analysis and then get a novel statistical form significance measure. 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 and so on.

PACS numbers:

*Electronic address: Hjli@amss.ac.cn
I. INTRODUCTION

Community structure detection \cite{1-3} is a main focus of social network studies. It has attracted a great deal of attentions from various scientific fields. Intuitively, community refers to a group of nodes in the 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 is proposed by Newman et al \cite{1-3} to quantify a network’s partition. Optimizing modularity is effective for community structure detection and has been widely used in many real networks. However, as pointed out by Fortunato et al \cite{4}, modularity suffers from the resolution limit problem which is concerned about the reliability of the communities detected through the optimization of modularity. Complementary to the modularity concept, many efforts are devoted to understanding the properties of the dynamical processes taking place in the underlying networks. Specifically, researchers have begun to investigate the correlation between the community structure and the dynamical systems, such as synchronization \cite{5} and random walk process \cite{6, 11}. Recently, phase transition \cite{12, 13, 14} of an algorithm from an undetectable region to one where detection is possible has been extensively studied, which used to investigate the performance of a variety of partition methods.

However, despite the large volume of works on community structure detection and its applications, one important question remains not clearly addressed, that of the significance of the communities in social networks. How to distinguish real communities from fake ones? How can we tell when the communities detected by different methods are truly significant and when they could be merely the consequence of a chance coincidence of edge positions in the network? How to statistically determine the significance of a given social community \cite{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 strong community structure will have high modularity and hence it is proposed to evaluate the community partition. However, recent studies show that this approach is insufficient \cite{25, 32}. Although it is true that networks with strong community structure have high modularity, it turns out that not all networks with high modularity have strong community structure. Researchers found that there exist networks with no obvious community structure at all that nonetheless have high modularity. In \cite{15}, 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 [17]. 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 although 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 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 [16][17], 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). Differently, in [18], the importance of a community partition is proportional to its robustness against random perturbations (i.e., random reshuffling of edges). The 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. [19], the notion of entropy $\Theta$ of graph ensembles is 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 which 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. [20] as well addressed the issue by comparing the cluster structure of the graph with that of a random graph with similar properties. Authors 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 happen 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 $C$-score.

However, these approaches rely heavily on the topology structure and don’t 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, not
suitable for a single community. In this paper we present a novel framework calculating social community significance specially. The framework does not embrace the universal approach but instead tries to focus on the unique properties of social networks. we model the dynamic 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 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 log-likelihood score, the tightness of community can be derived. Based on the distribution of community tightness, a new “p-value” form significance measure is proposed for community structure analysis. Finally, we apply our framework to both benchmark networks and real social networks. Experiments results show it can be used to (1) determine the optimal number of community; (2) analyze the social significance of a given community; (3) compare the performance among various algorithms and so on.

II. THE FRAMEWORK

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

Social hierarchy and community leader. It is natural to relate social networks with hierarchical structure[24]. In one such hierarchy there are nodes that are more important and influential than other nodes, hence locate on a higher level in the hierarchy (see Fig.1(a)). The leaders should have two properties: they are well connected to the members of their group, and they are 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 faster convergence rate.

Hierarchical structure and leader nodes are also exist in almost all real social networks. By taking the famous Karate network [21] as an example, there are 2 significant leaders(nodes 1 and 33) and communities are build around those leaders (see Fig.1(b)). The removal of those leaders will result in splitting these communities since they are keeping these communities
FIG. 1: (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 id 1 and 34 are highlighted in the origin graph.

Together. Since the hierarchies are consequence of the spreading of correlation, and so 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 most influence should define its community. So, community detection can be performed by finding all natural leaders and all nodes on which they influence.

There are two representative ways defining the leader nodes: the first is degree leader, the most natural way that let nodes with largest degree (number of edges linked with it) are the leader nodes; the second is influence leader that using the notion of relative influence defined in [22]. The influence represents how important the opinion of a given node to its neighborhood. A leader is the node that owns biggest overall influence, since the overall influence represents how close a node is to the core of its community, and the actual potential of becoming a leader. Also, a leader should have more 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, denoting the relative influences between the nodes as $T_{ij}$, and the overall influences of nodes as $u_i^*$. $T_{ij}$ is defined as
\[ T_{ij} = \frac{a'_{ij}}{\sum_k a'_{kj}}, \]  

(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 influences of nodes \( u^*_i \) is defined as

\[ u^*_i = \sum_j T_{ij} = \sum_j \frac{a'_{ij}}{\sum_k a'_{kj}}. \]  

(2)

Node \( x_i \) is a leader if \( T_{ij} \cdot u^*_i > T_{ji} \cdot u^*_j \) for all \( x_j \). The product \( T_{ij} \cdot u^*_i \) combines the relative influence of node \( x_i \) towards node \( x_j \) with the overall influence of node \( x_i \).

**Identification of community based on leaders.** At 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 dynamic on graphs. Consider a graph with \( c \) leaders \( l_1, l_2, ..., l_c \) 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 \( y_i = (y^1_i, y^2_i, \ldots, y^c_i) \in \mathbb{R}^c \).

By \( y^k_i(t) \), 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 \( x_i \), \( y^k_i \) is initialized randomly, distributed uniformly on \([0, 1]\)(\( k = 1, 2, ..., c \)). Then we normalize each row of \( y_i \) so that for all leader \( k \), the sum of \( y^k_i \) is 1. At each iteration time \( t \), the influence vector of each regular node \( x_i \) is updated entry-wise \((k = 1, 2, ..., c)\) using the following rule:

\[ y^k_i(t + 1) = \frac{1}{\sum_j a_{ij} + 1}[y^k_i(t) + \sum_j a_{ij}y^k_j(t)], \]  

(3)

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

We notice that, for all time \( t \), \( \sum_k y^k_i(t) = 1 \). Eq. (3) is equivalents 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 \( l_k(k = 1, 2, ..., c) \) on any regular node \( x_i \), \( y^k_i \), is the probability that a random walker that starts from \( x_i \) hits \( l_k \) before it hits any other leader node. 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 a regular node belongs to the community that a given leader node in. 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 contain local information, random walk dynamic 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 section IV, which verify the efficiency of our algorithm.

**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.

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

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

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

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,

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

The background distribution $P_0(x)$ is contrasted with an alternative hypothesis: node $x$ being part of a community, a group of nodes distinguished by enhanced mutual similarity. The normalisation 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 $E[sim(x, z)]$, the expected similarity $sim(x, z)$ of vectors following distribution $Q(x|z, \eta)$. This relationship can be described as:

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

In other words, parameter $\eta$ equally determines the community’s “width” as 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 $P_0(x)$.
**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

\[
s(x|z, \eta) \equiv \log \frac{Q(x|z, \eta)}{P_0(x)} = \eta \text{sim}(x, z) - \log Z_\eta.
\]  

(8)

By Eq(8) the log-likelihood score assigns positive score values to nodes which 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 \(\text{sim}(x, z)\) and, via the normalisation constant \(Z_\eta\), on the background model \(P_0(x)\).

Given a community with nodes set \(\{x_1, ..., x_N\}\), 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,

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

(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 own comparable tightness to looser but larger communities.

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

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

(10)

Next, we introduce the computation procedure of \(p(S)\). The collection of all configurations of nodes set \(X\) with energy \(E\), so \(p(E)\) 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

\[
\int p(E) e^{-\beta E} dE = \frac{1}{N} \int e^{-N \beta e + \log p(e)}
\]

\[
\approx \frac{1}{N} e^{N \sup_e (\log p(e)/N - \beta e)}.
\]  

(11)

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

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

(12)
i.e. the normalised logarithm of the partition function, \( \log Z(\beta)/N = -\beta f(\beta) \), is a Legendre transform of the normalised logarithm of the probability, \( \log p(e)/N \). Exploiting the duality of the Legendre transform, we get

\[
\log p(e) \simeq -N \sup_{\beta} [\beta f(\beta) + \beta e] = N[\beta_0 e - \beta_0 f(\beta_0)].
\]  \tag{13}

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

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

Using the conclusion derived above, given all configurations of node set \( X = (x_1, ..., x_N) \) 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

\[
\log p(S) \simeq N \Omega(s) - \frac{1}{2} \log(gN).
\]  \tag{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 or equal to \( S \). This is a typical \( p \)-value form and 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 \([33] \) as the probability that a random chosen nodes set contains a community with 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 \( S_0 \): 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. Low value provides a necessary but not a sufficient condition in this direction.

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

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

where $s(x_i|z) = \text{sim}(x_i, z)$. Through this transform, the width of community can be determined by parameter $\mu$ simply. Suppose the size of network is large enough, by the mean field theorem, $s_i = s(x_i|z)$ is approximately Gaussian-distributed with variance $M$,

$$ P(s(x_i|z)) = \sqrt{\frac{1}{2\pi M}} \exp\{-s^2/(2M)\}. $$

Computation of the distribution of the tightness $S$ is straightforward from the derivation shown in Section II and requires calculation of the quality function:

$$ Z_c(\beta, \mu) = \int_{\mathbb{R}^N} e^{\beta S(x_1, ..., x_N|z, \eta)} P(s_1) ... P(s_N) ds_1 ... ds_N $n$

$$ = \int_{-\infty}^{+\infty} e^{\beta \max[s_i - \mu, 0]} P(s) ds]^N $$n$

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

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

with $H(x) = \int_{x}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$, the complementary cumulative Gaussian distribution. In Eq.(17) the integration is divided in 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

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

and the entropy is

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

As described in section II,

$$ \log p(S, \mu) \simeq N \omega(S/N, \mu) - \frac{1}{2} \log N. $$
**Significance score.** For a given community, the significance score \( F \) can be calculated using the probability that the community tightness \( S \), \( p(S) \), greater than or equal to \( S \),

\[
F(S, \mu) = \int_S^{+\infty} p(S', \mu) dS'.
\] (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, thus it is very convenient to estimate the performance and function property of a given algorithm.

**Computational complexity.** The calculation of significance score mainly contains three parts: 1. calculate the degree or influence of every node, to find the leaders of communities. 2. identify the communities in network based on the position of nodes and leaders. 3. measure the similarity between nodes with their corresponding leaders and calculate the significance score by the distribution of tightness. The computational complexity of our method depends on the highest complexity of these three parts. Obviously, the part 2 is of the highest one, while other two parts are rather low. For part 2, the computational complexity is \( O(N^2) \), where \( N \) is number of nodes in the network. Thus, we obtain 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: (1) to evaluate the performance of a given algorithm; (2) to apply it on both artificial benchmark networks and real social networks.

**A. Benchmark network**

**GN benchmark network.** First, we use the classical GN benchmark presented by Girven and Newman\(^\text{[23]}\). 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 which 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 $\langle k^{in} \rangle$ nodes of its own group and $\langle k^{out} \rangle$ of the rest of the network. The total degree of each node is always kept constant and equal to $k = \langle k^{in} \rangle + \langle k^{out} \rangle$. Each group represents a well defined community up to $\langle k^{out} \rangle = 8$, but actually communities start to become very fuzzy at lower values of $\langle k^{out} \rangle \approx 8$, due to statistical fluctuations.

We empirically demonstrate the effectiveness of our algorithm through comparison with other five well-known algorithms on the GN networks. These algorithms include: Newman’s fast algorithm\cite{1}, Danon et al.’s method\cite{38}, the Louvain method\cite{39}, Infomap\cite{40}, the clique percolation method\cite{27} and the GA method\cite{41}. Fig. 2(a) presents the experimental results, in which y-axis denotes the fraction of nodes correctly clustered, and each point in curves is obtained by testing them against 50 synthetic networks shuffled from the original network. As we observe, all algorithms work well when $\langle k^{in} \rangle$ is larger than 0.7 with accuracy larger than 0.95. Compared with other six algorithms, our algorithm overall outperforms other algorithms and its accuracy is only slightly worse than that of the GA in the case of $0.5 \leq \langle k^{in} \rangle \leq 0.65$.

As is well known, the communities become fuzzier and thus more difficult to be identified when $\langle k^{out} \rangle$ increases. Hence, the significance of the community structure will also tend to be weaker and the $F$ index will decrease. The numerical results of $F$ value corresponding
FIG. 3: The performance of social significance $\langle F \rangle$ on both GN and LFR network. (a) In GN network, $\langle F \rangle$ decreases with increasing of $\langle k^{out} \rangle$. When the community structure is very clear $\langle F \rangle$ close to 1 very much, and the network close to no community structure network $\langle F \rangle$ close to 0.3, which implies that for a given network when $\langle F \rangle$ is less than 0.3 ($\langle k^{out} \rangle \approx 8$) it is not safe to say there exit significant community structure. (b) In LFR-benchmark, 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. With the increasing of mix parameter $\theta$, the $\langle F \rangle$ index decreases. When $\theta \geq 0.5$ (no significant community) $\langle F \rangle$ is near 0.3 which is similar with GN network.

both degree leader and 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, the $\langle F \rangle$ 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 the degree one, and therefore more effective. Furthermore, the topology becomes fuzzier when $\langle k^{out} \rangle$ increases, the sizes of communities will become more and more small correspondingly. 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 $\mu = 0.1$ when $\langle k^{out} \rangle$ larger than 6. We argue that for a given network when the corresponding $\langle F \rangle$ is larger than 0.3 ($\langle k^{out} \rangle \approx 8$), there exists significant community structure. Thus, the larger the $\langle F \rangle$ index is, the more significant community structure will be.

LFR benchmark network. We also test the index on the more challenging LRF
benchmark presented by Lancichinetti, Fortunato and Radicchi \[26\]. In the LFR benchmark, each node is given a degree took from a power law distribution with an exponent $\gamma$, and the sizes of the communities are took 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 other five well-known soft community partition algorithms on the LFR networks, including Clique percolation method \[27\], Link method \[42\], EAGLE method \[43\], RB Potts model \[44\] and Fuzzy C-Means method \[45\]. To evaluate a community detection algorithm, the normalized mutual information (NMI) \[20\] \[26\] is utilized to evaluate the partition found by each algorithm. The experimental results are displayed in Fig.2(b), where y-axis represents the value of NMI, and each point in curves is obtained by averaging the values obtained on 50 synthetic networks sampled from 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 other five algorithms, our algorithm performs quite well and its accuracy is only slightly worse than that of the Fuzzy C-Means in the case of $0.35 \leq \theta \leq 0.5$.

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 augment of $\theta$. Same as GN network, the $F$ values corresponding influence leader are larger than degree leader when $\theta$ is low. Furthermore, from Fig.3(b) we notice the value of $\langle F \rangle$ corresponding to $\mu = 0.25$ will larger than $\mu = 0.1$ when $\theta$ larger than 0.43.

**Stochastic block model.** Recently, many algorithms \[34\] \[35\] have been proposed to detect communities from networks or dynamical networks based on the famous stochastic blockmodel (SBM) first proposed by Holland et al \[36\] and extended by Decelle et al \[37\] \[12\] 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 $t_i \in \{1, ..., q\}$, specifying which of $q$ groups it is a member of. These labels are chosen independently, where $y_a$ is the probability that a given node has label $a \in \{1, ..., q\}$ (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 where $p_{ab} = O(1/N)$, we will use the rescaled affinity
matrix $c_{ab} = N p_{ab}$ 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 $y_a = 1/q, c_{ab} = c_{out}$ for $a \neq b$ and $c_{aa} = c_{in}$ with $c_{in} > c_{out}$, is a classical
problem in computer science and has been used as a benchmark for community detection. Here, $\varepsilon = c_{out}/c_{in}$ is used to control the fuzziness of generated network.

To test the performance on sparse networks, we establish a large network generated by
stochastic block model with low average degree. Fig[4] shows the case of network with
$N = 5000$ nodes and $q = 10$ groups with average degree $c = 8$. Each point in 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. On the contrary, when $\varepsilon$ is close
to 0.8, the network is nearly a random one, and the corresponding $\langle F \rangle$ values of both two
kinds of leaders are very low, near 0.2-0.3. Furthermore, it can be observed the value of $\langle F \rangle$
corresponding to $\mu = 0.25$ will be larger than $\mu = 0.1$. 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 significant
community structure which may detectable[12]. Therefore, the $F$ shows a great ability in
characterizing the significant modular structure as we adjust the parameter $\varepsilon$.

B. Real network

Now we show now the utility and versatility of our method for the statistical evaluation
of communities in real social networks. The significance corresponding to 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 don’t know the number of communities before incorporating additive information and the
community structure will be most clearest when the number is the optimal $c$. The detailed
steps are: 1. The degree or influence of each node is calculated and ranked. We choose
the first $q$ nodes with largest influence as leaders. 2. We partition the network and obtain
$q$ communities, using proposed community detection algorithm based on leaders. For each
$q(1 \leq q \leq N/2)$, a specific partition with $q$ communities can be got. 3. We apply our method
and use $\langle F \rangle_q$ to compute the significance of $q$ community structure. 4. By comparison, the $q$
corresponding to the largest value of $\langle F \rangle_q$ is chosen as the optimal number of communities.
FIG. 4: 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 $\varepsilon = c_{\text{out}}/c_{\text{in}}$ is used to control the fuzziness of generated network. Each point in curves is obtained by testing 50 times. With the increasing 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 two kinds of leaders are very low, near 0.2-0.3.

Here, three famous particular real examples are considered: Zachary karate club network [21], Collage football network [23] and Political books network [27]. The community partition of all networks has been obtained by our method in section II. As shown in Fig.5, the corresponding community numbers with the largest $\langle F \rangle_q$ are the optimal $c$ of every networks. 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 the $F$ score of the communities found are quite high. However, there are a few exceptions for which the $F$ 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 are not strong, especially the football network, thus the results are very precise and verify our framework are effective for real social networks.

Furthermore, to show that the model can uncover hierarchical structures in different scales, Fig.6 and Fig.7 give two examples of the multi-level community structures. Fig.6(a) shows the $RB125$ network, which is a hierarchical scale-free network proposed by Ravasz and Barabási in [28]. The regions corresponding to 5 and 25 modules are the most representative
FIG. 5: The empirical results of optimal number of communities on Zachary karate club network, College football network and Political books network. From the plots we observe that $\langle F \rangle$ achieves its highest value when the community numbers correspond the reality: Zachary karate club has 2 optimal communities, College football network has 12 optimal communities and Political books network has 3 optimal communities.
FIG. 6: (a) Structure of RB125, with 25 dense communities and 5 sparse communities, are highlighted in the original network. (b) The number of communities versus the average significance value $\langle F \rangle$.

FIG. 7: (a) Structure of H13-4, with 16 dense communities and 4 sparse communities, are highlighted in the original network. (b) The number of communities versus the average significance value $\langle F \rangle$.

In terms of resolution. Next, $H13-4$ proposed by Arenas et al [29] is shown in Fig.7(a) which 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 on 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 Fig.6(b) and Fig.7(b), we observe 25 and 16 are the optimal numbers of communities.
in RB125 and H13-4 networks owning the largest value, respectively. However, 5 modules and 4 modules are also reasonable partitions which 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 significance can also be used to rank the 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 chosen for their simplicity which are all able to automatically select the number of communities: the label propagation method [30], Wu-Huberman linear time method [31] and the Girvan-Newman betweenness algorithm [23]. Here, first, we partition the network into communities using a specific method. For each community, the node with 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 the $\langle F \rangle$ values of all three examples are not high, due to the fuzziness and sparseness of network’s topology. However, the $\langle F \rangle$ value which Girvan-Newman algorithm own are higher than other two methods, since the mechanism of Girvan-Newman algorithm is objective function optimization. In contrast, other two algorithms emphasize the simplicity of calculation too much while ignoring the accuracy of results. 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.

V. CONCLUSION

In summary, we present a novel framework calculating social community significance which doesn’t embrace the universal approach but instead tries to focus on the unique properties of social networks. Based on the distribution of community tightness, a new “p-value” form significance measure is proposed for network analysis. We apply our framework on both benchmark network and 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 trend of significance, such as the social significance of a given community, the optimal number of communities and the performance among various algorithms on detecting meaningful community structure.
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.

[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. 103, 8577-8582(2006).
[4] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. 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. 107(29), 12755-12760(2010).
[7] A. Lancichinetti, F. Radicchi, J. J. Ramasco and S. Fortunato, PloS one 6(4), e18961(2011).
[8] P. Zhang and C. Moore, arXiv preprint, arXiv:1403.5787(2014).
[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).

| Networks               | Algorithms                       | Values of $⟨F⟩$ |
|------------------------|----------------------------------|----------------|
| Zachary network        | label propagation method         | 0.641          |
|                        | Wu-Huberman linear time method   | 0.627          |
|                        | Girvan-Newman algorithm          | 0.735          |
| Collage 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          |

TABLE I: Comparison of various algorithms with $⟨F⟩$ values.
[12] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Phys. Rev. Lett. 107(6), 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, arXiv preprint, arXiv:1404.7789(2014).
[15] R. Guimera, M. Sales-Pardo and L. A. N. Amaral, Phys. Rev. E 70, 025101(2004).
[16] M. Sales-Pardo, R. Guimera, A. A. Moreira and L. A. N. Amaral, Proc. Natl. Acad. Sci. 104, 15224-15229(2007).
[17] J. Reichardt and S. Bornholdt, Preprint cond-mat/0606220 (2006).
[18] B. Karrer, E. Levina and M. E. J. Newman, Phys. Rev. E 77, 046119(2008).
[19] G. Bianconi, P. Pin and M. Marsili, Proc. Natl. Acad. Sci. 106(28), 11433-11438(2009).
[20] A. Lancichinetti, F. Radicchi and J. J. Ramasco, Phys. Rev. E 81, 046110(2010).
[21] W. W. Zachary, J. Anthropol. Res. 33, 452-473(1977).
[22] A. Stanoev, D. Smilkov and L. Kocarev, Phys. Rev. E 84, 046102(2011).
[23] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. 99, 7821-7826(2002).
[24] H. J. Li, Y. Wang, L. Y. Wu, Z. P. Liu, L. Chen and X. S. Zhang, Eur. Phys. Lett. 97, 48005(2012).
[25] X. S. Zhang, R. S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang and L. Chen, Eur. Phys. Lett. 87, 38002(2009).
[26] A. Lancichinetti, S. Fortunato and F. Radicchi, Phys. Rev. E 78, 046110(2008).
[27] G. Palla, I. Derényi, I. Farkas and T. Vicsek, Nature 435, 814-818(2005).
[28] E. Ravasz and A. L. Barabási, Phys. Rev. E 67, 026112(2003).
[29] A. Arenas, A. Fernandez and S. Gomez, New. J. Phys. 10, 053039(2008).
[30] U. N. Raghavan, R. Albert and S. Kumara, Phys. Rev. E 76(3), 036106(2007).
[31] F. Wu and B. A. Huberman, Eur. Phys. J. B 38, 331-338(2004).
[32] Z. P. Li, S. H. Zhang, R. S. Wang, X. S. Zhang and L. Chen, Phys. Rev. E 77, 036109(2008).
[33] J. D. Wilson, S. Wang, P. J. Mucha, S. Bhamidi and A. B. Nobel, arXiv preprint, arXiv:1308.0777(2013).
[34] B. Karrer and M. E. J. Newman, Phys. Rev. E 83(1), 016107(2011).
[35] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, Mach. Learn. 82(2), 157C189(2011).
[36] P. W. Holland, K. B. Laskey and S. Leinhardt, Social Networks, 5(2), 109-137(1983).
[37] A. Decelle, F. Krzakala, C. Moore and L. Zdeborová, Phys. Rev. E 84(6), 066106(2011).
[38] L. Danon, J. Duch, D. Guilera and A. Arenas, J. Stat. Mech. 29, P09008(2005).
[39] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, J. Stat. Mech. 10, P10008(2005).

[40] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. 105(4), 1118-1123(2008).

[41] R. Guimera and L. A. N. Amaral, Nature 2, 895-900(2005).

[42] Y. Y. Ahn, J. P. Bagrow and S. Lehmann, Nature 466, 761-764(2010).

[43] H. Shen, X. Cheng, K. Cai and M. B. Hu, Physica A 388, 1706-1712(2009).

[44] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. 93, 218701(2004).

[45] S. Zhang, R. S. Wang and X. S. Zhang, Physica A 374, 483-490 (2007).