Statistical significance of communities in networks

Andrea Lancichinetti

Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, Turin, Italy and Physics Department,
Politecnico di Torino, Turin, Italy

Filippo Radicchi and José J. Ramasco

Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, Turin, Italy

Nodes in real-world networks are usually organized in local modules. These groups, called communities, are intuitively defined as subgraphs with a larger density of internal connections than of external links. In this work, we define a measure aimed at quantifying the statistical significance of single communities. Extreme and order statistics are used to predict the statistics associated with individual clusters in random graphs. These distributions allows us to define one community significance as the probability that a generic clustering algorithm finds such a group in a random graph. The method is successfully applied in the case of real-world networks for the evaluation of the significance of their communities.

PACS number(s): 89.75.Hc, 89.75.Fb, 89.70.Cf

I. INTRODUCTION

Complex networks play a crucial role in understanding physical, biological, social, and technological systems [1–3]. Interactions between proteins in cells of living organisms, relations between human actors in socioeconomic contexts, and connections between web pages in the world wide web can naturally be described as graphs. Real-world networks typically have complex topological properties, but in spite of their evident diversity, structural analysis has revealed that they share a conspicuous set of common features: scale-freeness (i.e., the number of connections per node following a wide or power-law distribution) [1] and small-worldness (i.e., the average number of hops between two nodes in the network scales logarithmically with its size) [4] are two celebrated examples of such properties. Recent studies have focused on deeper structural features of networks. Real-world networks are typically organized in local clusters of nodes which are usually denominated communities. Communities are groups of nodes with a higher level of interconnection among themselves than with the rest of the graph. In this sense, communities are groups relatively isolated from the other nodes of the network and are expected to represent elements sharing common features and/or playing similar roles within the system (see Ref. [5] for an exhaustive review). For instance, if one considers the world wide web, communities are composed by groups of web pages dealing with similar topics; in social networks, communities stand for sets of actors sharing common interests, ideas, and friendship relationships; in protein interaction networks, communities represent groups of proteins with similar functionalities.

This imbalance of in and out connections corresponds to an intuitive concept. There are some formalizations of the definition of community. The LS set [6] or strong community [7,8] stands for a group where every node belonging to the group has more internal connections than external ones. A less restrictive definition refers to a *weak community* [8] as a set of nodes where the number of intracommunity connections (summed over all nodes within the group) is larger than the number of links going out of the community. Along these lines, the well known *modularity* is a quality function able to quantify the statistical importance of a partition comparing the number of internal connections observed in the communities with its expected number in a suitable *null model* [9]. Besides the formulation of a definition, big efforts have been made for the detection of communities in networks. Since the total number of possible divisions of a network in subgraphs is a nonpolynomial function of the size of the network itself, finding and detecting communities is not a trivial issue. Many algorithms have been proposed during recent years, every of them with the same spirit of finding the best groups which maximize the internal density of links [5,9–24]. Different principles may be used, but nevertheless in all cases some property related to the community structure is locally or globally optimized. The consequence is that even in uncorrelated networks these algorithms find clusters that are supposed to be good according to the modularity function or to other quality measures.

If algorithms are able to identify communities even in random graphs, which value can we give to communities found in real networks? Or better, how to statistically determine the significance of a community? This problem has been the subject of some studies in the literature [20,25–30]. In [20,27], for example, the partition of a network maximizing the modularity is compared with the maximum modularity partition of a randomized version of the given network (i.e., all edges are randomly rewired). In [29], differently, the importance of a community partition is proportional to its robustness against random perturbations (i.e., random reshuffling of edges). Such heuristic approaches rely on the modularity function to evaluate the quality of a partition, which means that are subjected to the modularity resolution limits [21,31]. Furthermore, all the proposed methods are designed to deal with full partitions not with single communities. Even though in a network one might find some meaningful communities alongside with randomly connected node clusters. In this paper, we develop a statistical method aimed at discriminating between a single *bona fide* community and structures arising as topological fluctuations. Instead of a direct
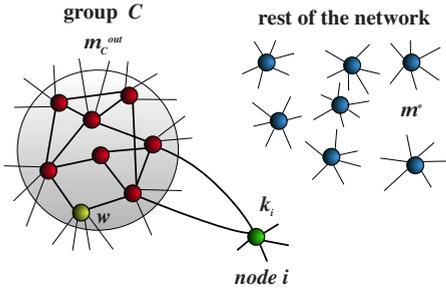
FIG. 1. (Color online) Sketch of the theoretical framework referring to the null model. Node i has k_i free ends to allocate. Each of them can connect to nodes within $\mathcal{C}$ or vertices belonging to the rest of the network.

comparison with an average outcome, the community is confronted with the best expected result for a null model. The reason for stressing this "best outcome" is that community detection algorithms will in general produce the best possible clusters given a graph even if it is random. The threshold of significance can be approximated by using extreme and order statistics [32,33] applied to null-model community fitness. A community significance can be then obtained as the extreme probability of finding a group equal or better than the one given in a set of equivalent random graphs.

II. NULL MODELS AND DEFINITION OF $\mathcal{C}$-SCORE

Consider a scenario as the one depicted in Fig. 1, with a given community $\mathcal{C}$ in a graph. k_i denotes the number of connections (degree) of the node i. Given $\mathcal{C}$, k_i can be divided in two terms: k_i^{int}, the number of links connecting i to nodes in $\mathcal{C}$, and k_i^{ext}, the number of connections outside. Similarly, we define the internal degree of $\mathcal{C}$, $m_{\mathcal{C}}^{int} = \sum_{j \in \mathcal{C}} k_j^{int}$, as well as $m_{\mathcal{C}}^{ext} = \sum_{j \in \mathcal{C}} k_j^{ext}$ and its total degree $m_{\mathcal{C}} = m_{\mathcal{C}}^{int} + m_{\mathcal{C}}^{ext}$. We consider a very simple stochastic null model: all connections inside the group are locked (the community is given so cannot be altered), while the other links are randomly reshuffled among all nodes preserving their degrees. For simplicity, we allow the rewiring operation to form multiple links (two nodes can be connected by more than one edge) or self-loops. In some weighted graphs the weights of the links are equivalent to multiple connections and so the present null model would be appropriate. Some examples are social networks (the Zachary club [34], see last section) or the *C. Elegans* metabolic network [16] that will be analyzed later. For unweighted graphs, we have checked that our results do not noticeably change by including or not multiple links as long as the graph is not condensed (a node gets a finite fraction of the links). When node degrees are much smaller than the network size, the probability of generating self-loops and multiple links by random reshuffling becomes negligible. Note also that our null model is similar to the one used for the definition of the modularity [9] and close in spirit to the configurational model [35]. It generates graphs that have no special internal structure except that given by random fluctuations, keep the degree sequence of the original network and can show degree-degree correlations only if the degree sequence and the network size determine their presence [36].
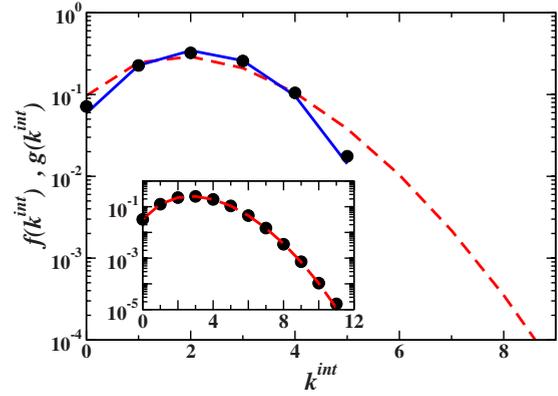

FIG. 2. (Color online) Distributions $f(k^{int})$ and $g(k^{int})$ for randomly generated networks of homogeneous degree. The (black) circles are numerical results with $N=100$, $n_{\mathcal{C}}=20$, and $k=15$. The distributions refer to external nodes of groups selected at random (inset) or by maximizing modularity (main plot). Dashed (red) curves are the approximation given by Eq. (1) and the continuous (blue) one that of Eq. (2).

This is the most general null model, appropriate when no knowledge about the system is available and simple enough to be treated from an analytical point of view. If further information regarding the constraints present in the process that generated the given network is available, other, simpler or more elaborated, null models can be employed. Our method to evaluate group significance is general enough to admit the use of different null models by altering consequently the distributions that will be described next.

Once the null model has been selected, suppose that $\mathcal{C}$ is a group composed of randomly chosen nodes and consider a generic node i not belonging to $\mathcal{C}$. The distribution of k_i^{int} is given by the hypergeometric distribution

$$f(k_i^{int}|\mathcal{C}) = \frac{\binom{m_{\mathcal{C}}^{ext}}{k_i^{int}} \times \binom{m^* - m_{\mathcal{C}}^{ext}}{k_i - k_i^{int}}}{\binom{m^*}{k_i}}, \qquad (1)$$

where $\binom{y}{x} = y!/(y-x)!x!$ is a binomial coefficient and m^* are the free ends in the network: $m^* = m - m_{\mathcal{C}}$ (m are the total ends in the graph, twice the number of links). Equation (1) states that the probability of node i to have k_i^{int} internal connections to $\mathcal{C}$ is given by the ratio of two terms: the total number of ways in which k_i^{int} links can be placed at the end of $m_{\mathcal{C}}^{ext}$ free ends multiplied by the number of ways to locate the remaining $k_i - k_i^{int}$ edges out of $m^* - m_{\mathcal{C}}^{ext}$ free ends divided by the total number of ways to place all k_i connections in the network (i.e., out of m^* free ends). If the node i belongs to $\mathcal{C}$, Eq. (1) has to be corrected to exclude i from the group. When the group $\mathcal{C}$ is composed of $n_{\mathcal{C}}$ randomly chosen nodes, Eq. (1) recovers the results obtained via numerical simulations (see inset Fig. 2).

The next, more interesting, case is when $\mathcal{C}$ is not composed of randomly chosen nodes, but it has been detected by a clustering algorithm. As can be seen in the main plot of Fig. 2, the shape of $f(k^{int})$ dramatically changes due to the

algorithm node selection. Most of the nodes populating the tail of the distribution are incorporated into the group. Correlations are also present since nodes in the community are expected to be connected among themselves. Still, it is possible to obtain an approximate expression for the probability $f(k^{int})$. We first consider the case of homogeneous graphs where all nodes have the same degree (i.e., $k_i = k, \forall \ i$) and extend later our analysis to networks with arbitrary degree sequences. We will assume that $\mathcal{C}$ has been selected to maximize k_i^{int} for each node inside $\mathcal{C}$ as well as the overall $m_{\mathcal{C}}^{int}$. This also implies that since the nodes are all equivalent, have the same degree k, they can be ranked according to their k^{int}. We indicate with w the node (or nodes) with the lowest k^{int} within the community (see Fig. 1). k_w^{int}, the internal degree of the worst node, establishes then an upper cut-off to the possible values of k^{int} of the out-group nodes. An expression similar to Eq. (1) can then be derived for the external nodes by taking into account this cutoff

$$g(k_i^{int}|\mathcal{C},k_w^{int}) = \frac{\binom{m_{\mathcal{C}}^{ext}}{k_i^{int}} \times \binom{\tilde{m} - m_{\mathcal{C}}^{ext}}{k_w^{int} - k_i^{int}}}{\binom{\tilde{m}}{k_w^{int}}}, \quad (2)$$

where $\tilde{m} = (N - n_{\mathcal{C}})k_w^{int}$. The term $\tilde{m}$ accounts for the fact that no node can connect to more than k_w^{int} internal vertices and therefore some of the free ends m^* become occupied. Equations (1) and (2) specify the null model. Our method does not depend on the particular functional shapes of $f(k_i^{int})$ and $g(k_i^{int})$. For instance, a more restricted null model without multiple links can be approximated by using Wallenius hypergeometric distribution although this considerably complicates the numerical evaluation of the functions. Another null model, less realistic but very easy to implement, is the Erdös-Rényi-like networks for which $f(k_i^{int})$ and $g(k_i^{int})$ are binomial distributions.

The worst node within the community, w, will play a central role in our method to evaluate group significance. We assume that in a random graph there is not a drastic variation between k_w^{int} and the internal degree k^{int} of the best nodes outside the group. Postulating a smooth variation in k^{int} between inside and outside of the community allows us to find an expression for the probability distribution of k_w^{int} based on Eq. (2) which only applies to external nodes. The degree of the worst node, k_w^{int}, is a given quantity in $g(k^{int}|\mathcal{C},k_w^{int})$. In order to find a formula for $P(k_w^{int})$, we need thus to alter our point of reference and consider the second worst node within the community w'. If the statistics of k_w^{int} is comparable to that of the best external nodes, k_w^{int} should follow the distribution of the extreme of $g(k_i^{int}|\mathcal{C}\backslash\{w\},k_{w'}^{int})$. This means that the probability for k_w^{int} to be lower or equal to a certain number reads as

$$\mathrm{Pr}(\leq k_w^{int}) = [G(k_w^{int}|\mathcal{C} \backslash \{w\}, k_{w'}^{int})]^{N-n_{\mathcal{C}}+1}, \quad (3)$$

where $G(\cdot)$ is the cumulative of the function $g(\cdot)$. The distribution $P(k_w^{int})$ is given by the derivative of the cumulative of Eq. (3), $P(k_w^{int}) = \partial \, \mathrm{Pr}(\leq k_w^{int})$. It must be remarked that Eq. (3) is valid for independent random variables, in our null
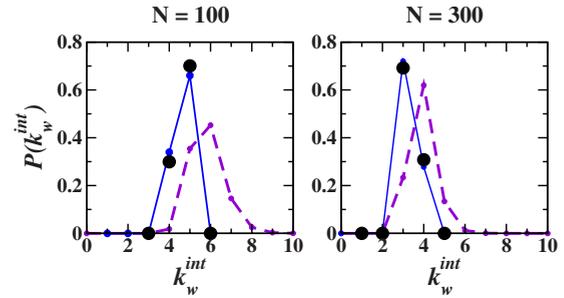

FIG. 3. (Color online) Probability distribution $P(k_w^{int})$ for the internal degree of the worst node of the community calculated for groups $\mathcal{C}$ detected by maximizing the modularity in randomly generated networks of $N=100$ (left) and $N=300$ (right) with homogeneous degree $k=15$ and for groups of $n_{\mathcal{C}}=20$. The black circles are the results of numerical simulation, while the continuous blue curves correspond to the theoretical predictions derived from Eq. (3). The extreme value distribution of $f(k^{int})$ from Eq. (1) is also plotted for comparison (violet dashed curve).

model the independence is justified for external nodes and is an approximation when refers to w. Figure 3 shows a comparison between the distribution $P(k_w^{int})$ obtained with this procedure and its counterpart from numerical simulations. Despite the approximations performed to reach an analytical form for $P(k_w^{int})$, the agreement is remarkable. The use of extreme statistics contributes in part to such agreement since under very general conditions the limit extreme-value distribution is stable and has no memory of the parental distribution.

Once a functional form for $\mathrm{Pr}(\leq k_w^{int})$ was obtained, we can define a measure of the significance of a group, the $\mathcal{C}$ score, as

$$c = \mathrm{Pr}(\geq k_w^{in}) = 1 - \mathrm{Pr}[\leq (k_w^{int} - 1)], \quad (4)$$

which corresponds to the probability that k_w^{int} for an optimized community in an equivalent random graph ensemble is higher than or equal to the value seen in $\mathcal{C}$. A point to stress here is that c contains not only information about the worst node, k_w^{int}, but also about the community external links and about the degrees of the external nodes.

In order to extend our results to heterogeneous graphs, we need to rank the nodes according to the role they play with respect to the given community $\mathcal{C}$. For regular networks, since all the nodes are equivalent, the ranking can be simply established by considering the values of the internal degrees k^{int}. However, another criterion is required to deal with heterogeneous networks. We use the probability distribution provided by Eq. (1) as the basis for such procedure. The rank for a node i can be established by the probability of finding a node with an internal degree k_i^{int} or higher in the null model given its degree k_i and $\mathcal{C}$. That is, for each node i we calculate the score $r_i = \sum_{q=k_i^{int}}^{k_i} f(q)$ and then perform comparisons on the basis of r. The values of r fall in the interval $[0,1]$ regardless of the node degree, which facilitate the comparison. w and w' correspond thus to the nodes with the highest and second highest values of r within the community, respectively. Under the hypothesis of a randomly connected net-

work, the scores r of the vertex w, r_w, and that of the external nodes can be seen as random variables uniformly distributed in the interval $[r_{w'}, 1]$. The $\mathcal{C}$ score can be then calculated as the probability of observing r_w as the minimal value of a set of $(N - n_\mathcal{C} + 1)$ random extractions from a uniform distribution defined in the interval $[r_{w'}, 1]$. An alternative to this last step is to map the internal degree of w' into $\hat{k}_{w'}^{int}$ (the internal degree that it would have if its degree was equal to k_w and its score $r_{w'}$) by inverting the distribution of Eq. (1). Once the transformation has been performed, we can proceed in the same way as for homogeneous networks with Eqs. (3) and (4).

III. BEYOND THE $\mathcal{C}$ SCORE

A low value of the $\mathcal{C}$-score (i.e., $c \leq 5\%$) is enough to consider a group as significant. However, when the $\mathcal{C}$-score is higher, one could argue that the reason is that relaying only on the worst node of the community for the full group evaluation is a too severe criterion. Algorithms may fail to place a single node and this would translate into a non significant community according to the $\mathcal{C}$-score approach. The performance of the method can be improved by a further refinement. Instead of considering only the last node, one can include a longer list of nodes and use this information for the computation of the statistical significance of the community. A way to do so is to write an algorithmic procedure. Three classes of nodes can be considered: The community $\mathcal{C}$, the "border" $\mathcal{B}$ and the rest of the network. Initially, the group $\mathcal{B}_0$ is empty and $\mathcal{C}_0 = \mathcal{C}$. Then at each algorithm step, the following actions are taken:

(i) Compute $r_i = \Sigma_{q=k_i^{int}}^{k_i} f(q)$, where the function $f(\cdot)$ is given by Eq. (1). r_i is calculated for each node $i \in \mathcal{C}$ with respect to the group $\mathcal{C}_t$;

(ii) Determine the worst node in $\mathcal{C}_t$, w_{t+1}, as the vertex with highest $r_{w_{t+1}}$. Set $\mathcal{B}_{t+1} = \mathcal{B}_t \cup \{w_{t+1}\}$ and $\mathcal{C}_{t+1} = \mathcal{C}_t \backslash \{w_{t+1}\}$;

(iii) Compute $\Pr(\langle S_{t+1} | \mathcal{C}_{t+1}, \mathcal{B}_{t+1}, r_{w_{t+2}})$, where $S_{t+1} = \Sigma_{i \in \mathcal{B}_{t+1}} r_{w_i}$ and w_{t+2} is the worst node still in $\mathcal{C}_{t+1}$;

(iv) Increase $t \to t+1$.

This algorithm explores the interior of the community trying to maintain the worst nodes always in $\mathcal{B}$, it ends when $t = n_\mathcal{C} - 1$. $\Pr(\langle S_{t+1} | \mathcal{C}_{t+1}, \mathcal{B}_{t+1}, r_{w_{t+2}})$ stands for the probability that the sum of the scores of the worst t nodes of an optimized community in an ensemble of equivalent random graphs is smaller than the given for $\mathcal{C}$. Its value for a set of independent random variables can be estimated by using order statistics (see Appendix for more details). We define then the $\mathcal{B}$ score as

$$\mathcal{B}\text{-score} = \min_t \Pr(\langle S_t | \mathcal{C}_t, \mathcal{B}_t, r_{w_{t+1}}), \qquad (5)$$

which corresponds to the lowest value of the probability $\Pr(\langle S_t | \mathcal{C}_t, \mathcal{B}_t, r_{w_{t+1}})$ observed during the iterative procedure. We take the minimum as the best approximation for the significance of the group $\mathcal{C}$ since it is evaluated in the most favorable discrimination of $\mathcal{C}$ nodes in border and core. This probability is equivalent to the $\mathcal{C}$-score for $t = 1$, while becomes a more synergic quantity as t increases. The inclusion
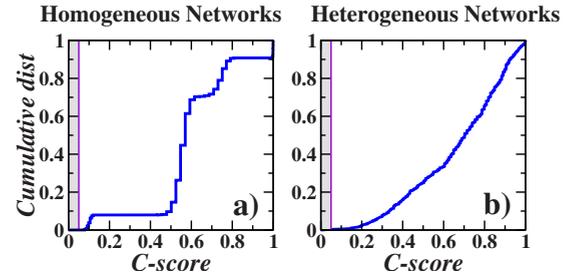

FIG. 4. (Color online) Cumulative distribution of the $\mathcal{C}$-score for groups of size $n_\mathcal{C} = 20$ obtained in random networks of $N = 100$ nodes by modularity maximization. (a) Homogeneous networks with degree $k = 15$. (b) Heterogeneous networks with average degree $\langle k \rangle = 15$. The degree distribution follows a power law $P(k) \sim k^{-\gamma}$ with $\gamma = 2$. The gray areas delimit the values of $\mathcal{C}$ score that indicate group statistical significance (probability $\leq 5\%$).

of a longer list of worst nodes in the calculation helps to correct conservative estimates due to undersampling. When communities are significant with respect to the $\mathcal{C}$ score they are significant also according to the $\mathcal{B}$ score. Vice versa, low values of the $\mathcal{B}$ score do not necessarily correspond to small $\mathcal{C}$ scores. Many concomitant bad nodes with features slightly different from the random expectations may multiply their effect and lead, if there is a real signal, to the prediction of a significant community.

IV. COMPUTATIONAL BENCHMARKS

As a first test, we applied the $\mathcal{C}$ and the $\mathcal{B}$ scores to groups found in random graphs using clustering techniques. The $\mathcal{C}$ score and the $\mathcal{B}$ score are able to identify these groups as not significant (see Fig. 4). The results confirmed that the scores are good estimators for the statistics of such groups further contributing to our confidence in the method. We consider next the performance of the scores on artificial networks with planted community structure. In order to do so, we build networks in the spirit of Girvan and Newman's benchmark [7]. Since our aim is to evaluate a single cluster, the benchmark will be composed of a group $\mathcal{C}$ with 32 nodes and of other 96 nodes in the rest of the network. Every node in $\mathcal{C}$ is connected on average with $\langle k^{int} \rangle$ nodes of its own group and $\langle k^{ext} \rangle$ outside. The external nodes are connected at random. The average total degree for all the nodes is fixed at $\langle k \rangle = 16$. $\langle k^{ext} \rangle$ acts thus as a control parameter for the strength of the community structure. The higher it is, the more prominent the disorder of the connections becomes. The scores are shown in Fig. 5(a) as a function of $\langle k^{ext} \rangle$. Both are able to detect the increasing disorder. Although, as expected, the $\mathcal{C}$ score is more conservative than the $\mathcal{B}$ score raising for earlier values of $\langle k^{ext} \rangle$ and so claiming that the group could be found in random graphs before. The (green) continuous curve in the figure represents a numerical estimation of the ideal function that we want to approximate with the scores. Before explaining how it is obtained, we need to describe the second panel of the figure. The distribution for the internal number of connections of $\mathcal{C}$ is displayed for the benchmarks at different $\langle k^{ext} \rangle$ as well as for equivalent randomized graphs in Fig.
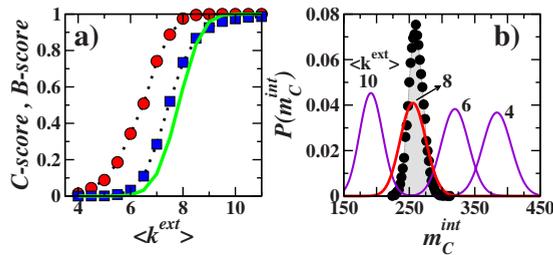
FIG. 5. (Color online) (a) $\mathcal{C}$ and $\mathcal{B}$ scores for communities in benchmarks. The disorder in the connections increases with $\langle k^{ext} \rangle$. The continuous (green) curve corresponds to the target distribution obtained numerically (see text for details). (b) Distribution $P(m_{\mathcal{C}}^{int})$, continuous curves are for the benchmark with the $\langle k^{ext} \rangle$ shown over the curve. The black circles are the numerical distribution measured for an equivalent random graph with the group found by maximizing modularity.

5(b). The randomized graphs are obtained by reshuffling the connections of the benchmark networks and the groups of 32 nodes in them are found by modularity maximization. The curves for the benchmarks start far away in the area of high $m_{\mathcal{C}}^{int}$ when $\langle k^{ext} \rangle$ is low. As $\langle k^{ext} \rangle$ increases, they move towards the left and at a certain point, close to $\langle k^{ext} \rangle \approx 8$, cross under the distribution for the randomized graphs. This point marks the end for the significance of the community. Similar (or better) groups could be found in a random graph by a clustering algorithm. The continuous curve in Fig. 5(a) is obtained by simulating this process. For each value of $\langle k^{ext} \rangle$, a set of instances of the benchmark are generated. $m_{\mathcal{C}}^{int}$ is measured for each of them, and the green curve is calculated averaging the probability of the value $m_{\mathcal{C}}^{int}$ or a higher one (cumulative distribution) in the random graph curve of Fig. 5(b). The good agreement of this curve with the $\mathcal{B}$-score proves that, despite all the approximations, the $\mathcal{B}$ score is a good measure of cluster significance.

As a final test on benchmarks, we have evaluated the scores performance on the benchmark proposed by Lancichinetti, Fortunato, and Radicchi (LFR) in Ref. [37]. This technique to generate graphs with planted community structure is a generalization of Girvan and Newman's method to networks with heterogeneous group size and degree distribution. As before, the nodes have k^{int} connections within its own group and $k^{ext} = k - k^{int}$ edges linking elsewhere. The mixing parameter k^{ext}/k indicates the strength of the communities. The scores shows a great ability in characterizing the modular structure of the benchmark as we increase the mixing parameter as can be seen in Fig. 6. Due to the absence of fluctuations all the communities are well defined until each node shares almost half of its connections with nodes of its group, while the groups become less defined for larger values of the mixing parameter. When about the 60% of the links connect with nodes outside the *a priori* established groups, the communities become equivalent to those found in random graphs.

V. EXPLORING THE INTERIOR OF A COMMUNITY

An interesting application of the scores is the exploration of the internal structure of groups. One could decide to re-
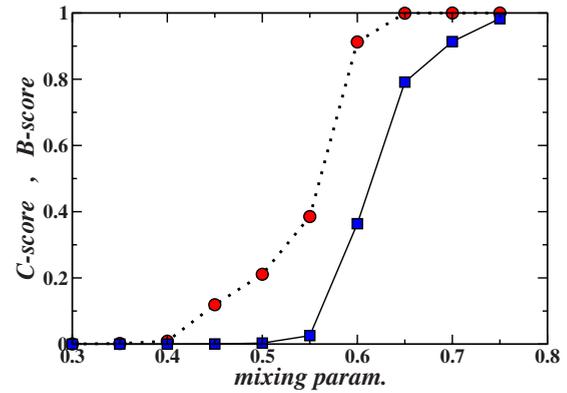

FIG. 6. (Color online) $\mathcal{C}$ score (red circles) and $\mathcal{B}$ score (blue squares) calculated for communities in LFR (heterogeneous) benchmarks. The scores are displayed as a function of the mixing parameter, k^{ext}/k. The benchmark networks size is $N = 1000$, the $\langle k \rangle = 15$ with a degree sequence exponent of $\gamma = -2$ and a size of the community $n_{\mathcal{C}} = 50$.

move the worst node from the community as we did to measure the $\mathcal{B}$ score and recompute the scores for the remaining group. The operation can be repeated iteratively as long as there are nodes remaining in the group. Interestingly, this process is able to identify the presence of internal structure in groups of vertices if the original community displays internal modularity. Figure 7 shows two examples of the described operation. The $\mathcal{B}$ score is plotted as a function of the number of removed nodes. We consider two different examples: a well-defined cluster (generated with the LFR benchmark) plus some randomly added nodes [Fig. 7(a)]; and a group composed of two clusters connected via few random links [Fig. 7(b)]. The iterative procedure is able to detect and set out the randomly added nodes [Fig. 7(a)] and also to find the deeper internal structure inside the two-element cluster [Fig. 7(b)].

This procedure also allows us to define more detailed measures for the quality of a community. We can search for deeper and deeper cores in the community that we will call $\mathcal{C}$-q or $\mathcal{B}$-q core. Fixed a level of significance q, the $\mathcal{C}$-q (or $\mathcal{B}$-q) core corresponds to the largest subgroup of a community with $\mathcal{C}$ score ($\mathcal{B}$ score) lower than q. In practical appli-
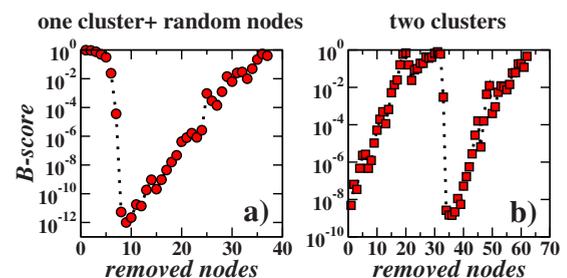

FIG. 7. (Color online) Iterative application of the $\mathcal{B}$ score in order to detect the presence of an internal organization in groups. At each stage, we remove the worst node of the community and compute the $\mathcal{B}$ score for the remaining group. We consider two examples: (a) a well-defined cluster with the addition of a few random nodes; and (b) a group composed by the union of two (good) communities.

TABLE I. Analysis of real networks with known community structure. For each network the table reports, from left to right, the name of the network, the size of its communities n_C, the C score, the B score, the size of the C-5%core and the size of the B-5%core.

Network	n_C	C score	B score	C -5%core	B -5%core
			Original partitions		
Karate club[a]	16	0.962	0.005	0	16
	18	0.989	0.004	0	18
Karate club weighted[a]	16	0.053	10^{-8}	11	16
	18	0.477	10^{-9}	17	18
College football[b]	9	10^{-10}	10^{-14}	9	9
	8	10^{-8}	10^{-11}	8	8
	11	10^{-9}	10^{-12}	11	11
	12	10^{-9}	10^{-12}	12	12
	10	1	0.941	9	9
	5	1	0.854	0	0
	13	10^{-8}	10^{-10}	13	13
	8	10^{-8}	10^{-10}	8	8
	10	10^{-10}	10^{-13}	10	10
	12	10^{-10}	10^{-13}	12	12
	7	0.937	0.407	4	4
	10	1	0.969	8	8

[a]Reference [34].
[b]Reference [7].

cations, a reasonable value of q is 5%. As we will see next, this concept turns out to be a useful tool to characterize communities in real networks. In the case of the benchmarks, the average sizes of the C-$q$cores obtained for the GN-like networks at $q=5$% are close to 32 up to $\langle k^{ext} \rangle = 8$. At this level of disorder, some nodes stop being significant for the planted communities and therefore come excluded from the q core. For higher disorder levels, the cores further reduce until eventually vanish.

VI. EMPIRICAL NETWORKS

We show now the utility and versatility of our method for the statistical evaluation of communities in real networks. An exhaustive study of the networks with modular structure in the literature has been performed, the following are only a few examples. We report results on social networks such as the Zachary karate club [34] or the one extracted for the characters of the novel Les Miserables [38] or for biological networks such as the *C. Elegans* metabolic network [16]. In two cases, the Zachary club and the college football networks, the structure of the groups is *a priori* known. In the Zachary club because the network split in two separate groups due to internal dissensions and for the college football because the conference in which the teams play is a given data. It is also important to note that some of these networks as, for instance, the Zachary club or the *C. Elegans* metabolic network are weighted graphs for which the

weights of the links are equivalent to multiple connections. We have analyzed both the weighted and unweighted versions and report both results in the case of the Zachary club. The evaluation of the groups for the *a priori* known communities is summarized in Table I. While the results for the communities obtained maximizing the modularity with a simulated annealing technique are displayed in Table II. There are some general observations valid for all networks. The C score is often able to discriminate good communities although sometimes a more sophisticated approach as the B score is needed. There are also a few cases in which the B score reverts the judge based on the C score, meaning that a deeper analysis of the communities was required. An example of this type is for instance the Zachary club two partitions. However, when the original graph with the weight information is considered its communities become more significant. This seems to apply also to the other weighted graphs, showing that there is a connection between clustering structure and weight location in these networks. We also show the sizes of the 5% cores of each community in the Tables I and II as well as detailed analysis of one of the communities of the *C. Elegans* metabolic network in Fig. 8.

VII. CONCLUSION AND DISCUSSION

Finding structure in graphs has direct implications for the study of several empirical disciplines as well as for a general understanding of the phenomena behind the evolution of the

TABLE II. Analysis of the community structure of several real networks via modularity maximization. For each network the table reports, from left to right, the name of the network, the size of its communities n_C, the C score, the B score, the size of the C-5%core and the size of the B-5%core. The community highlighted in Fig. 8 is marked as (green) in the table text.

		Maximal modularity partitions			
Network	n_C	C score	B score	C -5%core	B -5%core
Karate club[a]	11	0.987	0.029	0	11
	12	0.999	0.092	0	10
	5	0.096	0.017	0	5
	6	0.996	0.505	0	0
College football[b]	14	10^{-4}	10^{-4}	14	14
	10	0.005	0.003	10	10
	11	10^{-9}	10^{-12}	11	11
	15	0.088	0.001	14	15
	12	10^{-9}	10^{-12}	12	12
	9	10^{-10}	10^{-13}	9	9
	10	10^{-10}	10^{-13}	10	10
	9	10^{-6}	10^{-8}	9	9
	16	0.126	0.151	12	12
	9	10^{-10}	10^{-14}	9	9
Les Miserables[c]	11	0.241	10^{-4}	10	11
	17	0.954	0.455	15	16
	22	0.999	0.007	9	22
	10	0.868	10^{-6}	3	10
	11	0.999	10^{-4}	9	11
	6	10^{-9}	0	6	6
C. Elegans[d]	41 (Green)	1	0.832	32	38
	114	1	0.001	87	114
	47	0.999	0.583	35	42
	5	10^{-6}	10^{-14}	5	5
	23	1	0.704	21	21
	73	1	10^{-10}	67	73
	119	1	0.165	50	75
	31	1	0.865	25	25

[a]Reference [34].
[b]Reference [7].
[c]Reference [38].
[d]Reference [16].

systems in which such structures raise. Communities are the most direct and easy-to-envisage example of network structures. This concept is a direct heir of the intuitive idea of closer groups when considering social networks. As such, it has had a long history with a good number of algorithms proposed to detect communities in graphs. There are however two important issues missing in the literature. A firm mathematical definition of what a community means and a clear way to determine which of the outputs of the community detection algorithms are really significant.

In this work, we have focused on the second question with the hope of giving even if partially a hint of where the answer to the first one can lay. A measure able to statistically quantify the meaning of a single community in networks has been used. This measure, called C score, represents the probability of occurrence of a group with the same properties (i.e., same number of nodes, nodes with the same degree sequence and same internal connections) under the following hypothesis: (i) nodes in the network are randomly connected; (ii) the group is chosen, among all possible groups with the
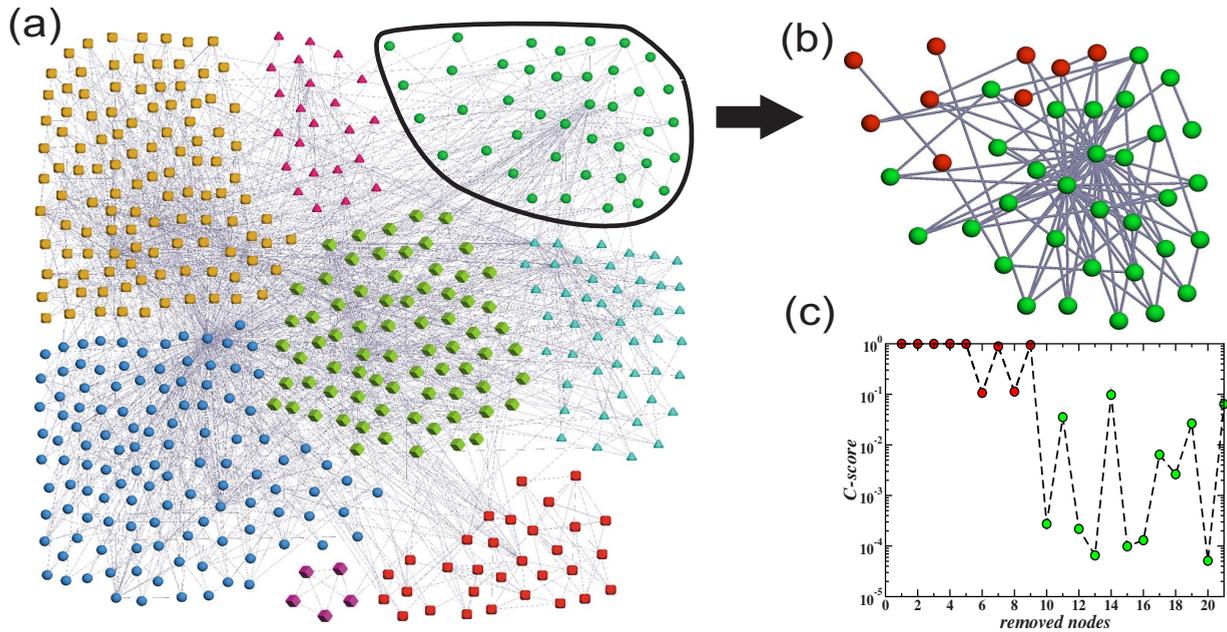
FIG. 8. (Color online) Community structure for the *C. Elegans* metabolic network [16] obtained by modularity optimization. In (a), an overview of the graph partition is shown. In (b), we display a zoom of a single community depicting in red the nodes that are not significant group members. And (c), the $\mathcal{C}$-*q*core analysis of the community.

same properties, because is the one which maximizes the density of internal connections. The first hypothesis is a natural assumption and a null model where links are randomly placed is very often used as term of comparison for the determination of correlations or other topological properties in networks. The latter one comes out from the common knowledge which prescribes communities as groups with high intraconnectivity. Thanks to the theory of extreme statistics, we approximate the values of the $\mathcal{C}$ score in the case in which our hypothesis hold. We have tested the performances of the $\mathcal{C}$ score on several networks, ranging from random graphs to artificial networks with controlled community structure, or to real networks with unknown internal organization. In all cases, we have been able to find good results. The method ability of evaluating one community at a time allows to detect situations in which only some of the communities of the graph are meaningful while the rest of the groups are equivalent to random fluctuations. This approach is also flexible enough to deal with overlapping groups that share nodes between them, providing a separate evaluation for each cluster. Two further refinements of the $\mathcal{C}$ score have been also introduced. One with the aim of exploring the internal structure of the communities, the q core, and another, the $\mathcal{B}$ score, with the intention of evaluating a community significance based on a group of nodes instead of on the worst node of the cluster. The computational complexity of the evaluation of the $\mathcal{B}$ and $\mathcal{C}$ scores grows quadratically and linearly with the community size, respectively. These tools constitute a set of statistical measures for a thorough evaluation of single communities, avoiding thus the blind acceptance of the output of clustering algorithms.

The software to calculate the $\mathcal{C}$ score and $\mathcal{B}$ score of communities is available at http://filrad.homelinux.org/cscore.

APPENDIX: THE DISTRIBUTION $\mathrm{Pr}(\langle S_t | \mathcal{C}_t, \mathcal{B}_t, r_{w_{t+1}})$

In Sec. III, we have outlined how to compute the $\mathcal{B}$ score of a community. The iterative procedure makes use of the probability $\mathrm{Prob}(\langle S_t | \mathcal{C}_t, \mathcal{B}_t, r_{w_{t+1}})$ which has yet to be described. During the procedure for the computation of the $\mathcal{B}$ score, the size of the border is increased by one at each stage. At step t, the border $\mathcal{B}_t$ is composed of the t nodes which, on the basis of their internal degrees, are less likely to belong to $\mathcal{C}_{t-1}$. We have therefore a sequence of scores, $r_{w_1} \geq r_{w_2} \geq \cdots \geq r_{w_t}$, for the t worst nodes. The score of the worst node, namely, w_{t+1}, still inside $\mathcal{C}_t$ represents a lower bound for the sequence since by definition we should have that $r_{w_t} \geq r_{w_{t+1}}$. Instead of trying to obtain the probability for the full sequence, we can simplify our problem and consider the sequence sum $S_t = \Sigma_{i=1}^{t} r_{w_i}$. Finding the distribution of S_t can be formulated as calculating the probability that, given a sequence of $N - n_{\mathcal{C}_t}$ independently identically distributed (iid) random variables (we indicate by $n_{\mathcal{F}}$ the size of a set $\mathcal{F}$), the sum of the t largest variables is less than S_t. The solution for this problem can be found in [32,39]. The cumulative probability distribution is given by the expression

$$\Pr(\langle S_t | \mathcal{C}_t, \mathcal{B}_t, r_{w_{t+1}}) = 1 - \sum_{j=1}^{\theta_t} (-1)^{j+1} \frac{(n_{\mathcal{B}_t} + 1 - j - \xi_t)^{N - n_{\mathcal{C}_t} - 1}}{(n_{\mathcal{B}_t} + 1 - j)^{N - n_{\mathcal{C}}} (n_{\mathcal{B}_t} - j)! (j-1)!}, \tag{A1}$$

where $\theta_t = \text{Integer-Value} \lfloor n_{\mathcal{B}_t} + 1 - \xi_t \rfloor$ and $\xi_t = (S_t - n_{\mathcal{B}_t} w_{t+1})/(1 - w_{t+1})$. Note that Eq. (A1) is valid under the assumption of independent variables, which is justifiable to some extent in the case of random networks.

---

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[3] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, England, 2004).
[4] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[5] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[6] S. B. Seidman, Soc. Networks **5**, 97 (1983).
[7] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[8] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[9] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[10] S. van Dongen, Ph.D. thesis, University of Utrecht, 2000.
[11] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).
[12] H. Zhou, Phys. Rev. E **67**, 061901 (2003).
[13] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[14] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[15] L. Donetti and M. A. Muñoz, J. Stat. Mech.: Theory Exp. (2004) P10012.
[16] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[17] G. Palla, I. Derényi, I. Frakas, and T. Vicsek, Nature (London) **435**, 814 (2005).
[18] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[19] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006).
[20] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Acad. Sci. U.S.A. **104**, 15224 (2007).
[21] J. S. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Eur. Phys. J. B **56**, 41 (2007).
[22] M. E. J. Newman and E. A. Leicht, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).
[23] J. J. Ramasco and M. Mungan, Phys. Rev. E **77**, 036122 (2008).
[24] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).
[25] V. Spirin and L. A. Mirny, Proc. Natl. Acad. Sci. U.S.A. **100**, 12123 (2003).
[26] R. Guimerà, M. Sales-Pardo and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[27] J. Reichardt and S. Bornholdt, Physica D **224**, 20 (2006).
[28] J. Reichardt and M. Leone, Phys. Rev. Lett. **101**, 078701 (2008).
[29] B. Karrer, E. Levina, and M. E. J. Newman, Phys. Rev. E **77**, 046119 (2008).
[30] G. Bianconi, P. Pin, and M. Marsili, Proc. Natl. Acad. Sci. U.S.A. **106**, 11433 (2009).
[31] S. Fortunato and M. Barthelémy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[32] H. A. David and H. N. Nagaraja, *Order Statistics*, Wiley Series in Probability and Statistics (Wiley and Sons, New York, 2003).
[33] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of Extremes* Wiley Series in Probability and Statistics (Wiley and Sons, New York, 2004).
[34] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[35] M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).
[36] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[37] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[38] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).
[39] A. P. Dempster and R. M. Kleyle, Ann. Math. Stat. **39**, 1473 (1968).