Finding and Testing Network Communities by Lumped Markov Chains

Carlo Piccardi*

Department of Electronics and Information, Politecnico di Milano, Milano, Italy

Abstract

Identifying communities (or clusters), namely groups of nodes with comparatively strong internal connectivity, is a fundamental task for deeply understanding the structure and function of a network. Yet, there is a lack of formal criteria for defining communities and for testing their significance. We propose a sharp definition that is based on a quality threshold. By means of a lumped Markov chain model of a random walker, a quality measure called "persistence probability" is associated to a cluster, which is then defined as an "α-community" if such a probability is not smaller than α. Consistently, a partition composed of α-communities is an "α-partition." These definitions turn out to be very effective for finding and testing communities. If a set of candidate partitions is available, setting the desired α-level allows one to immediately select the α-partition with the finest decomposition. Simultaneously, the persistence probabilities quantify the quality of each single community. Given its ability in individually assessing each single cluster, this approach can also disclose single well-defined communities even in networks that overall do not possess a definite clusterized structure.

Citation: Piccardi C (2011) Finding and Testing Network Communities by Lumped Markov Chains. PLoS ONE 6(11): e27028. doi:10.1371/journal.pone.0027028

Editor: Yamir Moreno, University of Zaragoza, Spain

Received July 5, 2011; **Accepted** October 9, 2011; **Published** November 3, 2011

Funding: The author has no support or funding to report.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: carlo.piccardi@polimi.it

Introduction

Complex networks are currently one of the most extensively studied subjects in the field of applied mathematics. In the last fifteen years, a huge number of theoretical results have been put forward, and almost any field of science and technology has benefit from the application of such results to specific problems [1–4].

One of the most promising but challenging tasks in network science is *community analysis*, which is aimed at revealing possible partitions of a network into subsets of nodes (*communities*, or *clusters*) with dense intra- but sparse inter-group connections. Finding and analyzing such partitions often provides invaluable help in deeply understanding the structure and function of a network, as widely demonstrated by several case studies in social sciences [5,6], biology [7], ecology [8], economics [9], or information science [10,11], just to name a few.

Despite the abundance of contributions on this subject (see [12] for a thorough survey), the issue of community analysis cannot be considered satisfactorily solved. First of all, finding communities is a computationally hard task, because the "best" partition must be sought for in a set whose cardinality grows faster than exponentially with the number of nodes. The exhaustive enumeration of the partitions is thus impossible, and heuristic techniques must be employed. Secondly, and perhaps more important, there is no widespread consensus on formal criteria for defining communities and for testing their significance [12]. When a subnetwork can actually be considered to form a community, namely a group of nodes with comparatively strong internal connectivity? Many contributions, mostly coming from social sciences, computer sciences, and physics, have tried to answer this question in various ways, over the years (e.g., [13–16]). Probably the most important attempt was put forward by Newman and coworkers [5,17,18], who defined a quality index called *modularity* which quantifies, for a given partition of the network into candidate communities, to what extent the distribution of the intra-/inter-community edges is anomalous with respect to a suitably defined random network. Since high modularity values are obtained in presence of groups of nodes with comparatively large intra-community edge density, maximizing modularity should put in evidence the "best" partition. This method has been proven successful in many circumstances but, on the other hand, it has been widely demonstrated that, due to intrinsic limitations, it does not necessarily always yield a significant partition [12,15,19]. And even when it does, it quantifies the quality of a partition but not of each individual community. For that, many other methods for community analysis have been put forward in the last few years, trying to simultaneously finding a meaningful network partition and assessing its significance (we recall, e.g., [20–22]).

This paper introduces a sharp definition of community which is based on a quality threshold. More precisely, once a level $0<\alpha<1$ is specified, a node cluster is defined to be an *α-community* if the probability that a random walker, which is currently in one of the cluster's nodes, remains in the cluster in the next step is not smaller than α. Such a probability is obtained from an approximate *lumped Markov chain* model of the random walker (i.e., a reduced-order Markov chain in which the communities of the original network become nodes) which is easily derived from the original (high-order) Markov chain model. Consistently, a partition composed of α-communities is defined to be an *α-partition.*

If equipped with an effective method for generating a set of "good" candidate partitions, the notions of α-community and α-partition provide a framework for simultaneously finding communities and testing their significance. For that, the desired quality level α is first fixed. Then, a family of partitions is derived and each partition is immediately checked to assess whether it is formed by α-communities. This allows one to identify the α-partitions, and to select one of them. Typically, one searches for communities which are at the same time small (to effectively decompose the network) and significant (with much more internal than external connectivity). For that, a guideline is that of selecting, among the available α-partitions, the one with the largest number of communities.

But the notion of α-community can also be useful in a partially different way. It may happen that, for a given quality level α, no α-partitions are found. Yet, one or a few α-communities could exist. They correspond to strongly connected groups of nodes, even in a network which, overall, does not possess a definite clusterized structure. Or, finally, one can assess the significance of the results of a single-partition method, such as modularity optimization [5], and obtain an immediate assessment of the α-quality of each single community and, consequently, of the entire partition.

In the paper, we first introduce the lumped Markov chain model of the random walker and define the notions of persistence probability, α-community, and α-partition. Testing the α-quality of a given community or partition turns out to be extremely parsimonious in computational terms. Then we analyze the problem of finding communities in a given network. For that, we propose an effective algorithm for deriving a meaningful set of partitions, among which the "best" one will be selected. The algorithm, which applies hierarchical cluster analysis, is again based on the Markov chain model of a random walker and, consequently, it involves a notion of similarity/distance among nodes which is consistent with the quality criterion above introduced. The results of the application of the above approach to both synthetic and real-world networks are discussed. We finally compare this approach, which can be applied to fully general networks (i.e., directed and weighted), with other community analysis methods having a similar philosophy.

Methods

Networks, α-Communities, and α-Partitions

Consider a network with nodes $\mathbb{N} = \{1,2,\ldots,N\}$ and L edges. In the most general case the network is directed and weighted, and we denote by $W = [w_{ij}]$ the $N \times N$ weight matrix, where $w_{ij} \geq 0$ is the weight of the edge $i \rightarrow j$. The connectivity matrix $A = [a_{ij}]$ is the $N \times N$ binary matrix where $a_{ij} = 1$ if $w_{ij} > 0$, and $a_{ij} = 0$ otherwise. If the network is actually undirected we have $W = W'$ and $A = A'$, and if it is unweighted we let $W = A$ (i.e., all weights equal to 1). Since connectedness is typically required for communities ([12], p. 84), we naturally assume that the network is strongly connected (e.g., [3]), namely there exists an oriented path from any i to any j. If this is not the case, namely the network is disconnected, each strongly connected component must be separately analyzed.

If the network is directed, for each node i we define the (total) degree as $k_i = k_i^{in} + k_i^{out} = \sum_j a_{ji} + \sum_j a_{ij}$, whereas $k_i = \sum_j a_{ji} = \sum_j a_{ij}$ for undirected network. The average degree is given by $\langle k \rangle = \sum_i k_i / N$. Similarly, for a directed network the in-, out-, and total strength of node i are given by $s_i^{in} = \sum_j w_{ji}$, $s_i^{out} = \sum_j w_{ij}$, and $s_i = s_i^{in} + s_i^{out}$, respectively, and the total network weight by $S = \sum_{ij} w_{ij}$. If the network is undirected we have instead $s_i = s_i^{in} = s_i^{out} = \sum_j w_{ji} = \sum_j w_{ij}$ and $S = \sum_{ij} w_{ij}/2$.

A N-state Markov chain $\pi_{t+1} = \pi_t P$, with $\pi_t = (\pi_{1,t} \pi_{2,t} \ldots \pi_{N,t})$, can be associated to the N-node network by row-normalizing the weight matrix W, namely by letting the transition probability from i to j equal to

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{s_i^{out}}. \qquad (1)$$

The quantity p_{ij} is the probability that a random walker which is in node i jumps to node j, and $\pi_{i,t}$ is the probability of being in node i at time t. The transition matrix $P = [p_{ij}]$ is a row-stochastic (or Markov) matrix ($0 \leq p_{ij} \leq 1$ for all i,j, and $\sum_j p_{ij} = 1$ for all i). Furthermore, P is irreducible since the network is connected. This implies that the equation $\pi = \pi P$ has a unique solution π, which is strictly positive ($\pi_i > 0$ for all i) [23] and corresponds to the stationary Markov chain state probability distribution. For undirected networks one can easily check that $\pi = (s_1 s_2 \ldots s_N)/(2S)$, whereas for directed networks a general closed form does not exist and π has to be numerically computed.

We denote by $\mathbb{P}_q$ a partition of $\mathbb{N}$ in q subsets (or subnetworks), namely $\mathbb{P}_q = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_q\}$ with $\bigcup_c \mathbb{C}_c = \mathbb{N}$ and $\mathbb{C}_c \bigcap \mathbb{C}_d = \emptyset$ for all c,d. The sub-network $\mathbb{C}_c$ is called a (candidate) *community* (or *cluster*). Defining a partition $\mathbb{P}_q$ induces a q-state meta-network, where communities become meta-nodes. The rigorous description of the dynamics of the random walker at this scale by a *lumped Markov chain*, however, is not possible if not in special cases [24] - actually, the Markovian property is not even preserved in general. Despite this limitation, a q-state Markov chain can be defined, which correctly describes the random walker at the aggregate level provided the stochastic process is started at the stationary distribution π [25,26]. This lumped Markov chain is defined by the $q \times q$ row-stochastic matrix

$$U = [\mathrm{diag}(\pi H)]^{-1} H' \mathrm{diag}(\pi) P H, \qquad (2)$$

where H (*collecting matrix*) is a $N \times q$ binary matrix coding the partition $\mathbb{P}_q$, i.e., its entry h_{ic} is 1 if and only if node $i \in \mathbb{C}_c$ (see the Supporting Information S1 for the derivation of equation (2)). The lumped Markov chain $\Pi_{t+1} = \Pi_t U$ shares the stationary distribution with the original one (suitably collected), namely $\Pi = \pi H$ satisfies $\Pi = \Pi U$. On the contrary, starting from an arbitrary π_0, the lumped Markov chain $\Pi_{t+1} = \Pi_t U$ started at $\Pi_0 = \pi_0 H$ provides, in general, only an approximate description of the evolution of $\pi_t H$. The difference between the real and approximate Π_t, however, tends exponentially to zero if the two chains are regular [23], since they converge, by definition, to the same stationary state.

The ability of the lumped Markov chain to describe the random walk dynamics only at stationarity is not a limitation for our purposes, as it will be demonstrated by the examples of application. Note that the entry u_{cd} of U is the probability that the random walker is at time $(t+1)$ in any of the nodes of community d, provided it is at time t in any of the nodes of community c. We define *persistence probability* of community c the diagonal term u_{cc}. Large values of u_{cc} are expected for meaningful communities. In fact, the expected escape time from $\mathbb{C}_c$ is $\tau_c = (1 - u_{cc})^{-1}$: the walker will spend long time within the same community if the weights of the internal edges are comparatively large with respect to those pointing outside. Given a value $0 < \alpha < 1$, $\mathbb{C}_c$ is defined *α-community* if $u_{cc} \geq \alpha$. Thus α acts as a selection parameter, as sharply qualifies communities with respect to a given quality threshold. Consistently, $\mathbb{P}_q$ is defined *α-partition* if it is composed of α-communities, namely $u_{cc} \geq \alpha$ for all $c = 1,2,\ldots,q$.

Testing communities

Testing the quality of a given partition is the simplest use of persistence probabilities. The partition can be the outcome of a community detection method (e.g., max-modularity) or instead derive from some *a priori* division (e.g., countries of the same continent in a financial network, or students of the same class in a school). By computing the u_{cc}-s using equation (2), the quality of each community and of the entire partition is readily quantified.

Consider the simple 12-node network of Fig. 1 [27], which is purposely composed of three clusters. Four partitions are considered, corresponding to finer and finer divisions, and the u_{cc}-s are computed for each candidate community. As long as the communities coincide with "natural" clusters, or with the union of two of them, all the u_{cc}-s remain rather large. But as soon as a natural community is broken, some very low persistence probabilities are found. If, for example, the quality level $\alpha = 0.5$ is fixed (a value having an important interpretation - see below), only the first and second partition are such that $u_{cc} \geq \alpha$ for all c (i.e., they are α-partitions). But even if the third and fourth partition fail in meeting such a requirement, yet some of their clusters can individually be classified as α-communities.

From equation (2), one can derive the explicit expression of the persistence probability u_{cc} of cluster $\mathbb{C}_c$ (see also the Supporting Information S1):

$$u_{cc} = \frac{\sum_{i,j \in \mathbb{C}_c} \pi_i p_{ij}}{\sum_{i \in \mathbb{C}_c} \pi_i}. \tag{3}$$

Kim et al. [28] note that $\sum_{i,j \in \mathbb{C}_c} \pi_i p_{ij}$ is the fraction of time that the random walker spends *on the links* internal to $\mathbb{C}_c$. Thus u_{cc} is the ratio between the latter and the fraction of time spent *on the nodes* of u_{cc}. In the case of *undirected* network, recalling that $\pi = (s_1 s_2 \ldots s_N)/(2S)$, we obtain

$$u_{cc} = \frac{\sum_{i,j \in \mathbb{C}_c} w_{ij}}{\sum_{i \in \mathbb{C}_c} s_i} = \frac{2W_c}{S_c}, \tag{4}$$

having denoted by W_c the *total internal weight* and by S_c the *total strength* of $\mathbb{C}_c$. Thus the persistence probability has, in this case, a straightforward interpretation: it is the fraction of the strength of the nodes of $\mathbb{C}_c$ that remains within $\mathbb{C}_c$.

In the even more special case of *unweighed* networks, this has a strict relationship, in turn, with the notion of "community in a weak sense" put forward by Radicchi et al. [14], who defined a community as a set $\mathbb{C}_c$ of nodes whose edges directed within $\mathbb{C}_c$ are more than those directed toward the rest of the network. It can easily be verified that this corresponds to $u_{cc} > 0.5$. Therefore persistence probabilities generalize the above notion of "community in a weak sense" in a twofold direction: first, they extend it to weighted, directed networks; second, they allow a flexible tuning of the "strength" of the communities by fixing the desired minimum acceptable value (not necessarily 0.5) for u_{cc}.

We note that (again by restricting the attention to undirected, unweighed networks), it can easily be checked that $u_{cc} = 1 - \Phi_n(\mathbb{C}_c)$, where $\Phi_n(\mathbb{C}_c)$ is the *normalized cut* of cluster $\mathbb{C}_c$ (e.g., [12] p. 92), namely the ratio between the number of edges connecting $\mathbb{C}_c$ to the rest of the network and the sum of the degrees of the nodes of $\mathbb{C}_c$. This observation bridges our dynamical, Markov-chain-based method with traditional graph partitioning techniques. It has already been pointed out that the latter are scarcely suitable for community detection [5,12], because the number of clusters has typically to be provided *a priori* whereas, in most instances, it is part of the outcome the network analyst is seeking for (see [29] for a relationship between modularity and cut size). Nonetheless, in the next section we shall see how a flexible exploitation of persistence probabilities enables an effective community analysis.

Finding communities

In the previous section, the persistence probabilities were used for *testing* given partitions and, individually, their communities. Here, instead, we want to analyze how this tool can be exploited for *finding* communities, namely for deriving partitions composed of meaningful communities.

The starting point is to define the desired level for the quality parameter α. For example, as pointed out above, in the case of undirected, unweighed networks, the constraint $u_{cc} > \alpha = 0.5$ for all c is equivalent to require partitions composed of "communities in a weak sense", according to the definition of [14]. But the network analyst can be more or less restrictive, i.e., require a larger $(0.5 < \alpha < 1)$ or smaller $(0 < \alpha < 0.5)$ significance level.

In general, for any given α, a large set of α-partitions exist, i.e., such that $u_{cc} \geq \alpha$ for all c (e.g., the trivial partition $\mathbb{P}_1 = \{\mathbb{N}\}$, the entire network, is an α-partition for any given α). Typically one searches for small (yet significant) communities, to effectively decompose the network. Thus we can rigorously formulate the problem of community detection as follows:

$$\max_{\mathbb{P}_q \in \mathcal{P}} q \quad \text{subject to} \quad u_{cc} \geq \alpha, \quad c = 1,2,\ldots,q, \tag{5}$$

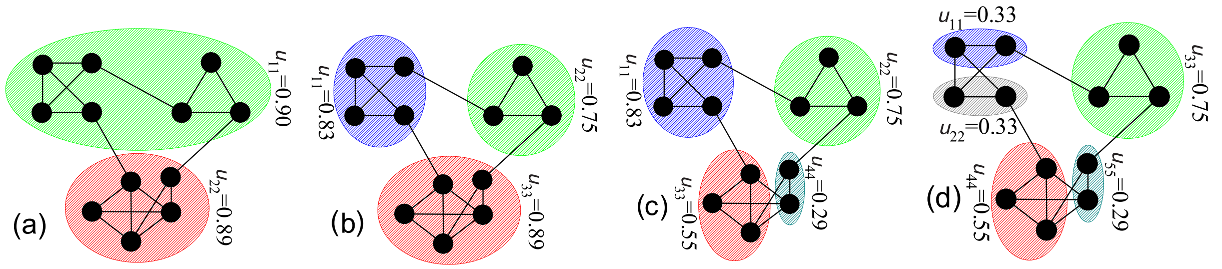

Figure 1. Four different partitions (with increasing number q of communities) of the same network. The persistence probabilities u_{cc} remain large as long as the network is partitioned into "natural" communities, or unions of them. Passing from (b) to (c), and from (c) to (d), significant communities are broken, with a sudden drop of the relevant persistence probabilities.
doi:10.1371/journal.pone.0027028.g001

where $\mathcal{P}$ denotes the set of all partitions. Notice that the admissible set of problem (5) is not empty for any given α (since $\mathbb{P}_1 = \{\mathbb{N}\}$ has $u_{11} = 1$) and that, in general, the optimal solution is not unique (if $q = \bar{q}$ attains the maximum in (5), there can be many $\mathbb{P}_{\bar{q}}$ which are α-partitions).

Analyzing the theoretical properties of problem (5) is beyond the scope of this work (see [30] for a discussion on the NP-completeness of some related optimization problems). Instead, a heuristic approach for finding a suboptimal solution to (5) can readily be derived, by restricting the optimization to a (much smaller) subset $\mathcal{P}^* \subset \mathcal{P}$ obtained by whatever "partitions generator", namely an algorithm that yields a set of partitions $\mathbb{P}', \mathbb{P}'', \ldots$ which are hopefully "good" candidates for community detection. In this way, problem (5) is readily solved by picking up the α-partitions within $\mathcal{P}^*$, and taking the one(s) with the largest q. We will make reference to this procedure in the remainder of the paper, but we anticipate that, instead of the "unsupervised" approach just outlined (with α fixed *a priori*), we will often prefer a "supervised" approach consisting in first generating a bunch of meaningful partitions, then comparatively assessing their quality, and finally selecting the preferred one, thus implicitly fixing the α value *a posteriori*. We will illustrate this procedure through many examples.

Several methods have been proposed to derive network partitions which are meaningful in the sense of community analysis (see again [12] for a thorough analysis). All of them can be used in our framework: here we adopt a method for deriving partitions which is based on cluster analysis and is consistent with the above introduced random walk modeling.

Cluster analysis can be used to group "similar nodes" into candidate communities. This needs defining a meaningful *similarity/distance* among each pair of nodes. Such a definition is by no means obvious: among the many proposals [12], a few exploit random walks to induce a suitable similarity measure (e.g., [31–35]). We follow this line by proposing an approach in which, however, we do not explicitly perform random walks in a Monte Carlo fashion, but derive analytically the global behavior of a large number M of walkers (a "fleet") started from each node i.

Consider a large number M of repetitions of a random walk started from i. For each repetition, the probability that the walker is in j after t steps is $[P^t]_{ij}$. Thus, if M random walks of length T are performed from i, the expected number of visits to j in any time instant in $1 \le t \le T$ is $M \sum_{t=1}^{T} [P^t]_{ij}$. By averaging with respect to M, we propose a (symmetric) similarity σ_{ij} defined by

$$\sigma_{ij} = \sigma_{ji} = \sum_{t=1}^{T} \left([P^t]_{ij} + [P^t]_{ji} \right). \tag{6}$$

Note that this is conceptually equivalent to an explicit random walk approach, but with an arbitrarily large number M of repetitions from each starting node instead of one only. Most notably, the results do not depend on the actual stochastic realization of the random walks. We finally define the distance $d_{ij} = d_{ji}$ between nodes (i,j) by complementing the similarity and normalizing the results between 0 and 1:

$$d_{ij} = d_{ji} = 1 - \frac{\sigma_{ij} - \min \sigma_{ij}}{\max \sigma_{ij} - \min \sigma_{ij}}. \tag{7}$$

The rationale underlying the definition of σ_{ij} and d_{ij} is to assign nodes (i,j) a large similarity if a numerous fleet of random walkers started in i (resp. j) makes a large number of visits to j (resp. i) within a sufficiently small time horizon T. The notion of community induced by this metric, therefore, is that of a subnetwork where a random walker has a large probability of circulating for quite a long time, before eventually leaving to reach another group. This is conceptually consistent with the definition of α-community above introduced.

The choice of the time horizon T is potentially critical. Cluster analysis yields a different hierarchical tree (*dendrogram*) for each time horizon T, whose choice is thus nontrivial. At the two extremes, setting $T = 1$ restricts the pairs of nodes which are candidate to nonzero similarity to neighboring pairs only, whereas larger and larger values of T tend to make any node equally similar to any other. We found that an effective selection of T can be empirically obtained by maximizing the *cophenetic correlation coefficient* C, which is defined as the linear correlation between the distances d_{ij} and the *cophenetic distances* c_{ij} [36]. The latter are a product of the hierarchical cluster analysis: for any node pair (i,j), the cophenetic distance c_{ij} is the height of the link joining (directly or indirectly) nodes (i,j) in the dendrogram. The value of C is generally used to assess whether the adopted distance d_{ij} induces an effective clusterization (notice that C qualifies the entire dendrogram, and not a network partition), although limitations have been observed in specific applications [37].

The entire procedure for finding communities is summarized in Fig. 2 with reference to the toy-network of Fig. 1. Starting from the network description, we apply cluster analysis for each T ranging from 1 to some sufficiently large $T_{\max}$ (of the order of N), eventually taking the T value that maximizes C. Horizontal top-down cross-sections of the associated dendrogram identify a sequence $\mathbb{P}_2, \mathbb{P}_3, \ldots$ of partitions with increasing number q of candidate communities. For each $\mathbb{P}_q$ we compute the lumped Markov matrix U according to (2), and plot its diagonal terms in the *persistence probabilities' diagram*. In the case of Fig. 2, the sudden drop of the least u_{cc} for q larger than 3 reveals that a meaningful community has been broken passing from $\mathbb{P}_3$ to $\mathbb{P}_4$. If we set, for instance, the quality threshold at $\alpha = 0.5$, then $\mathbb{P}_2$ and $\mathbb{P}_3$ can be qualified as α-partitions, and thus $\mathbb{P}_3$ will be our choice if we seek for the finest partition, consistently with problem (5).

Results

The analysis of four networks is now discussed. We will consider two families of synthetical benchmark networks with built-in cluster structure; a real-world network with a rather strong community structure; and another real-world network with weak clustering but with a few well-defined communities. Other examples are discussed in the Supporting Information S1.

LFR benchmarks

Lancichinetti, Fortunato, and Radicchi (LFR) [38] proposed a family of synthetically generated graphs, designed to serve as benchmarks for testing community detection algorithms. They explicitly took into account two properties found in real networks, namely the heterogeneity in the distributions of node degrees and community sizes. Both of the latter are taken as power laws, with given exponents γ and β, respectively. In addition, the network is defined by prescribing the number N of nodes, the average degree $\langle k \rangle$, and a *mixing parameter* μ such that each node shares a fraction $1 - \mu$ of its edges with the other nodes of its own community, and a fraction μ with the rest of the network. The benchmark generating method was later extended to oriented and weighted networks [39] (see the Supporting Information S1) - here we consider undirected, unweighed networks with $N = 1000$, $\langle k \rangle = 20$, $\gamma = 2$. We first let

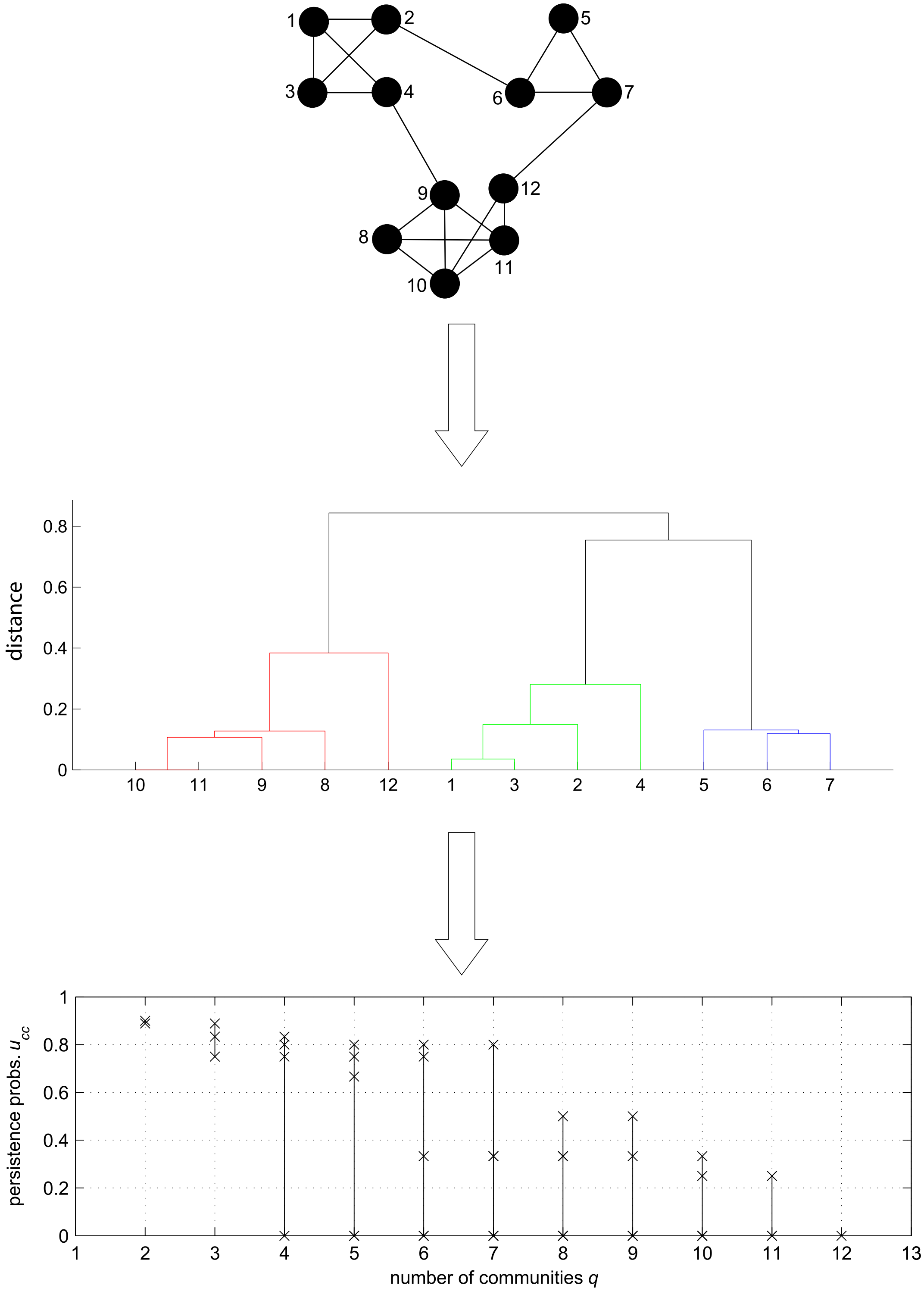

Figure 2. Summary of the procedure for community analysis. From the network description (top panel) and a suitable definition of node distance, a hierarchical tree (dendrogram) is derived by cluster analysis (middle panel). Horizontal top-down cross-sections of the dendrogram identify a sequence $\mathbb{P}_q$ of partitions with increasing number q of candidate communities. In the *persistence probabilities' diagram* (bottom panel), the q diagonal terms u_{cc} of the lumped Markov matrix U are plotted for each partition $\mathbb{P}_q$ (crosses denote the values of the u_{cc}, vertical straight lines are only for visual aid). In this example, the sudden drop of the least u_{cc} for q larger than 3 reveals that a meaningful community has been broken passing from $\mathbb{P}_3$ to $\mathbb{P}_4$.

doi:10.1371/journal.pone.0027028.g002

$\beta=1$ and $\mu=0.25$. Since the generating algorithm is stochastic, we produce 10 different network instances: the number of built-in communities q^* turns out to range from 35 to 43, and the size of each community from 10 to 77 nodes.

We now fix our desired quality level, for example $\alpha=0.5$, and solve problem (5) for each of the 10 networks. For that, we use the above described "partitions generator": in Fig. 3 we show, for illustrative purposes, the cophenetic correlation coefficient C as a function of the random walk time horizon T, as obtained analyzing one of the networks. We find a unimodal dependence, as for almost all the network studied. We take therefore $T=12$ in this case, which attains the maximum $C=0.905$. The related dendrogram is in the same figure.

The persistence probabilities' diagrams obtained for the 10 networks are shown in Fig. 4. In all instances, the diagrams reveal a sharp discontinuity. For $q\leq q^*$, all the u_{cc}-s are rather large (larger than 0.72). This indicates that meaningful communities are identified. For $q>q^*$, instead, some significant communities are broken, as revealed by a larger and larger number of small u_{cc}-s. In other words, the correct number of built-in communities is systematically revealed, in all instances, by a sudden drop of some of the persistence probabilities. This implies, in turn, that solving problem (5), i.e., taking the largest q such that $\mathbb{P}_q$ is an α-partition with $\alpha=0.5$, yields a solution with $q=\bar{q}$ which exactly recovers the number q^* of communities. Furthermore, such a solution is largely insensitive to the choice of the quality level: for example, any value in a range $0.1<\alpha<0.7$ would give the same result.

Obviously, the fact that $\bar{q}=q^*$ does not imply that the two partitions are identical. In order to quantify the ability of the method, we compare the built-in partition with that obtained by solving problem (5), in terms of the *normalized mutual information* I, a reliable and often used measure of partition similarity introduced by [40] to the network research community. The definition of I is reported in the Supporting Information S1: here we only point out that $I=1$ when the two partitions are identical, whereas I has zero expected value for independent partitions. We obtain an average of $I=0.997$ over the 10 networks, which favorably compares to the values reported by [38] after extensive tests by using modularity optimization ($I\approx0.975$) and Potts model clustering [41] ($I\approx0.925$).

In [38] it is shown that the performance of community detection algorithms deteriorates when, *ceteris paribus*, the scale parameter β of the power-law community size distribution increases (i.e., communities are less differentiated in size) and/or when the mixing parameter μ increases (i.e., communities become less isolated each other). To analyze this situation, we generate another set of 10 benchmark networks by increasing β from 1 to 2 and μ from 0.25 to 0.6 (the highest μ value considered in [38]): the resulting networks turn out to have from 47 to 58 communities, with size ranging from 10 to 61 nodes. Notice that we are generating low-quality clusters, due to the large μ: actually, none of them would met the requirement of "community in a weak sense" according to [14]. In other words, the cluster structure of the network is extremely weak, and that is obviously the reason of the scarce performance of community detection tools.

All of this is captured by the persistence probabilities' diagrams of Fig. 5. All the candidate partitions are characterized by low-quality clusters (with the u_{cc}-s accumulating in the range $0.25-0.4$), which is the obvious result of the low quality of the built-in partitions. In this situation, when analyzing one

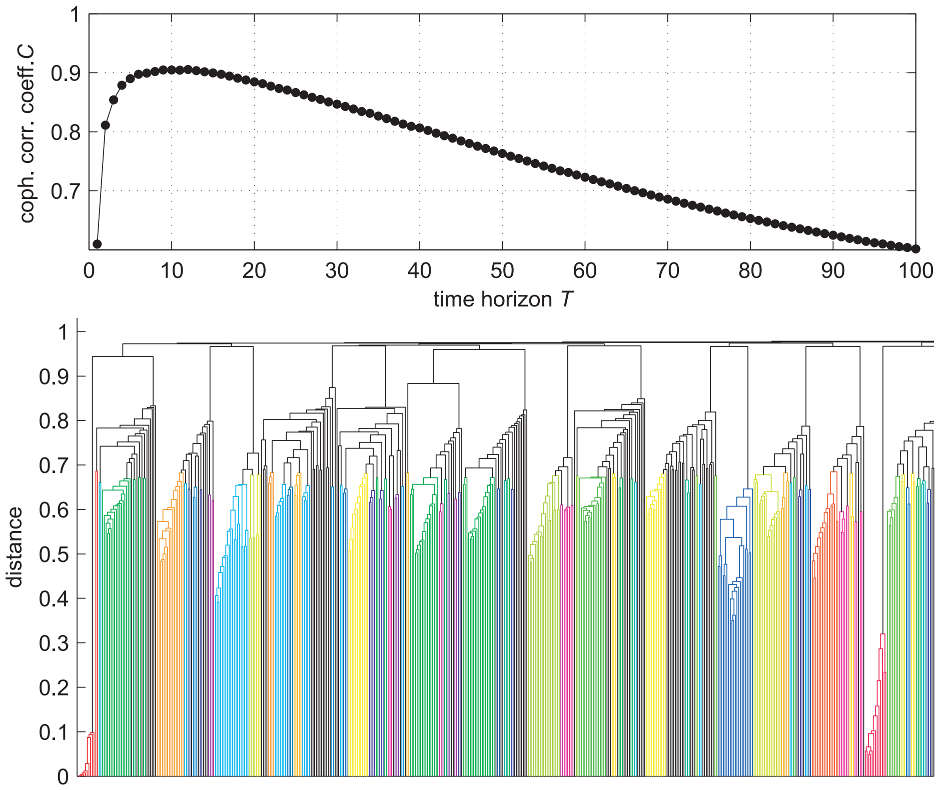

Figure 3. LFR benchmark networks. Above: The cophenetic correlation coefficient C as a function of T for one of the network instances. The maximum is attained at $T=12$. Below: The dendrogram obtained with $T=12$ (only half of the plot is shown for readability).
doi:10.1371/journal.pone.0027028.g003

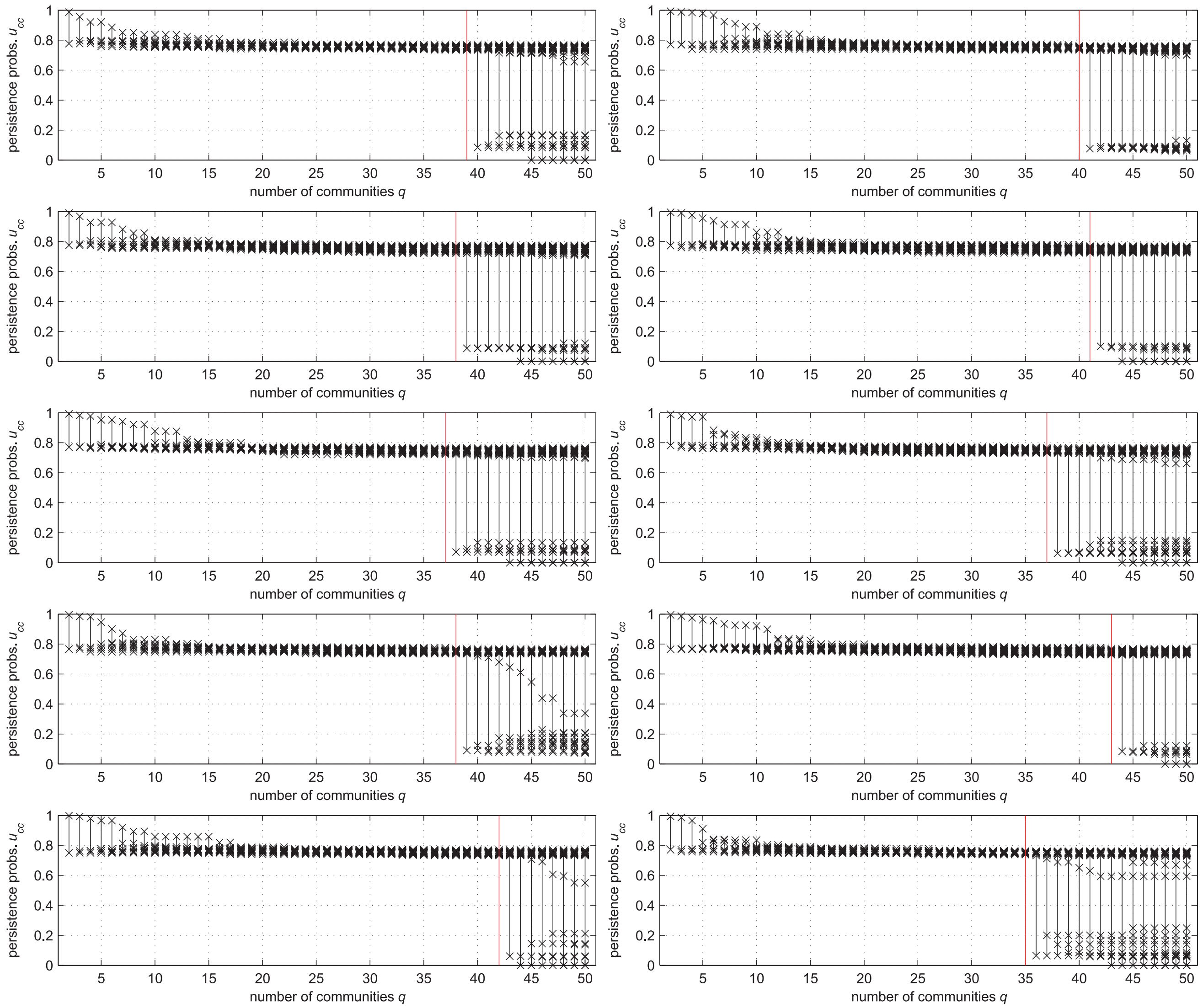

Figure 4. The persistence probabilities' diagrams of the LFR benchmark networks with $\beta = 1$, $\mu = 0.25$. (See the text for the other parameters). The vertical red line marks the built-in number of communities. In all instances, this value is revealed by a sudden drop of some of the persistence probabilities.
doi:10.1371/journal.pone.0027028.g004

of the networks, the detection procedure of [14], based on the notion of ''community in a weak sense'', would discard any candidate community; the max-modularity approach would yield a partition as outcome, but with no assessment of the quality of its clusters; and also the unsupervised solution of our problem (5) would lead to poor results: for example, setting *a priori* the value of α to the ''standard'' value of 0.5 would discard all partitions.

It is exactly in such a difficult context that persistence probabilities can be a precious decision support tool. By looking at the diagrams of Fig. 5, the analyst immediately grasp the weak cluster structure of the network under scrutiny, and can consistently *a posteriori* fix an α value not unrealistically restrictive. Alternatively, he/she can rely on the observation of a sudden drop in one or more persistence probabilities' as an indication of a (comparatively) good partition. This means selecting $q = \bar{q}$ such that $\min_c u_{cc}$ has the largest variation from $\mathbb{P}_{\bar{q}}$ to $\mathbb{P}_{\bar{q}+1}$. If we systematically apply this strategy to the 10 benchmark networks, we obtain an average mutual information between the built-in and the obtained partition of $I = 0.844$, which is intermediate with respect to the values obtained by [38] with modularity optimization ($I \approx 0.875$) and Potts model clustering ($I \approx 0.825$). But the added value of our approach is, for the selected partition $\mathbb{P}_{\bar{q}}$, the quality measure u_{cc} of each cluster and, consequently, of the entire partition.

Netscience network

The Netscience network is a weighted, undirected, social network describing the collaborations (up to year 2006) among researchers in network science, the weight of the edge connecting two researchers being proportional to the number of papers they have co-authored [18]. Its giant component has $N = 379$ nodes, and it is generally considered an example of a real network with a rather strong community structure. Many methods for network analysis, included community detection algorithms, have been tested and discussed on this example (e.g., [42-44]).

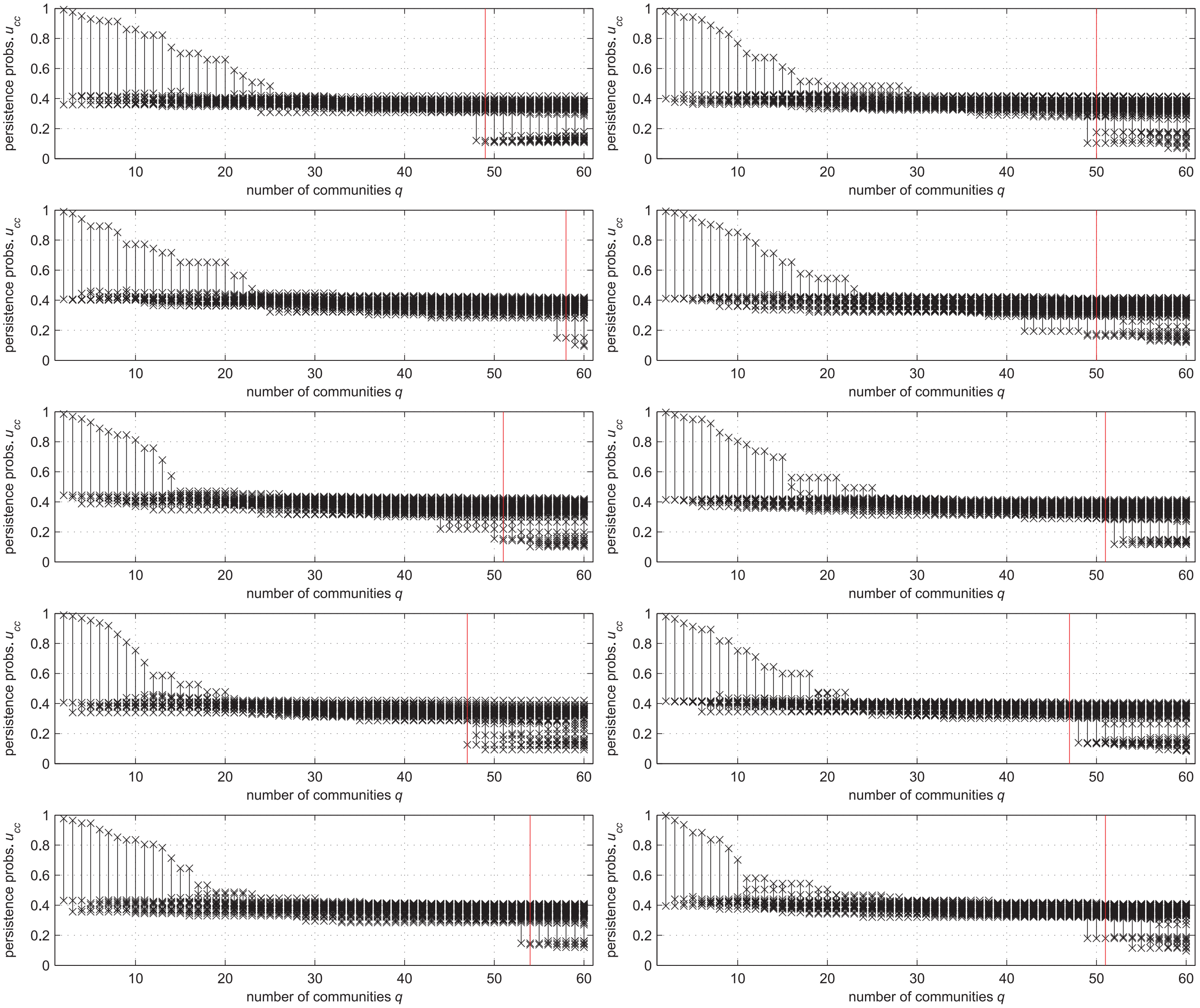

Figure 5. The persistence probabilities' diagrams of the LFR benchmark networks with $\beta = 2$, $\mu = 0.6$. (See the text for the other parameters). The vertical red line marks the built-in number of communities. The persistence probabilities accumulate around $0.25 - 0.4$, denoting the low quality of the clusters (compare with Fig. 4).
doi:10.1371/journal.pone.0027028.g005

If we run our partitions generator algorithm, at $T = 6$ we get the dendrogram attaining the largest C: the resulting persistence probabilities' diagram is in Fig. 6. The plot has a less sharp structure than that of the LFR networks of Fig. 4: if we adopt once again the criterion of [14], namely we solve problem (5) in an unsupervised fashion by letting $\alpha = 0.5$, then $\mathbb{P}_{35}$ is the optimal partition (here we have straightforwardly extended the notion of "community in a weak sense" to weighted networks). In a supervised approach, instead, the network analyst will select the proper q as a trade-off between a finer decomposition (large q) and a higher significance of the communities (small q). For example, setting α as large as 0.9 yields $\mathbb{P}_{10}$ as the optimal partition, i.e., the α-partition with largest q.

It is instructive to compare these results with those obtained, on the same case study, by the *graph stability* approach proposed by Delvenne et al. [42] (a detailed comparison of the two methods is in the next section). By means of the KVV algorithm [45] (a hierarchical, divisive, non-binary, graph clustering method), they obtain a sequence of six partitions, with $q = 2,3,5,15,17,21$. Analyzing and comparing the *stability curve* (i.e., the autocovariance function of a signal emitted by a random walker) of each of them, the authors suggest their partition with $q = 5$ as the more reliable, as it has the largest stability over a longer time span with respect to any other. Incidentally, this is also a supervised approach that leaves the analyst the choice of the preferred solution among a set of alternatives.

In order to test the six partitions of [42], we created their persistence probabilities' diagram and compared it with our results in the diagram of Fig. 7. The partition $q = 5$ of [42] confirms to be definitely more significant than those with finer decomposition (i.e., $q = 15,17,21$) according to the criterion of minimal u_{cc} too. Actually, our and their $\mathbb{P}_5$ partitions share the same minimal $u_{cc} = 0.952$, due to a common 22-node community. They are, however, partially different (their normalized mutual information is $I = 0.886$, with about 6% of differently classified node pairs).

The inspection of Fig. 7 also reveals that, for each given q, the partitions obtained with our method are superior than those proposed in [42], provided the criterion put forward in this paper

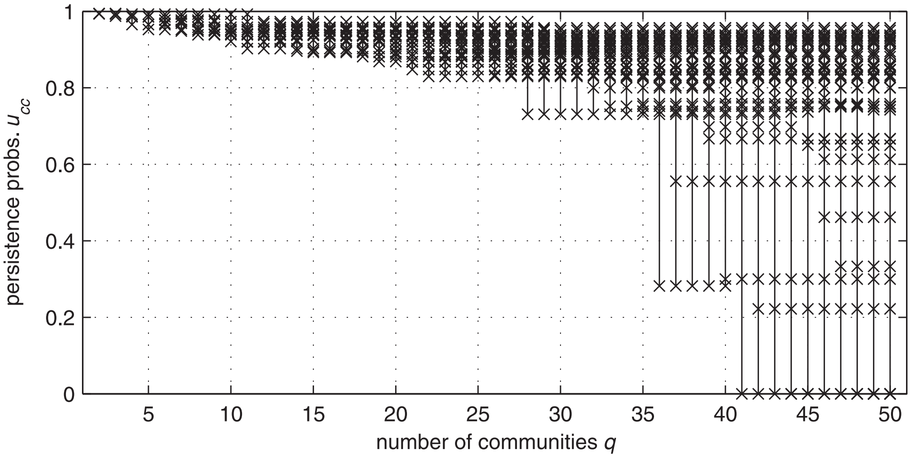

Figure 6. The persistence probabilities' diagram of the Netscience network.
doi:10.1371/journal.pone.0027028.g006

(i.e., minimal u_{cc}) is adopted. Actually, while the criterion of [42] ranks partitions by "averaging" among the communities, our approach is a "worst-case" one: by selecting an α-partition one guarantees that the "worst" community has a persistence probability not less than α. Finally, note that in the gap from $q=6$ to 15, where no partition is obtained by the KVV divisive algorithm, our partitions generating algorithm provides a set of finer and finer partitions, whose quality only slowly deteriorates as q increases. The network analyst can fruitfully select in this interval a proper trade-off between fine granularity and significance of the partition.

World trade network

The final example concerns a real-world, directed, weighted network, representing the trade flows among countries. This network, denoted as *world trade network* (or *world trade web*), has extensively been studied in recent years (e.g., [46–48]). The problem of the existence of communities, namely groups of countries with preferential partnerships, has been addressed too, although results seem to be not definitive [49,50]. This issue is obviously related to the debate about "globalization versus regionalization" in the world economy.

We consider the network derived from 2008 data, whose largest connected component has $N=181$ nodes. It does not seem to display a definite community structure: as a matter of fact, the maximum modularity (estimated as in [51]) is rather small, namely $Q=0.296$, if compared to other examples where N has the same order of magnitude (e.g., $Q=0.831$ for the Netscience network). In this situation, we show how our method is able to detect well-defined communities (if any) even in a network which overall does not possess a definite clusterized structure. Consider the persistence probabilities' diagram of Fig. 8. With the exception of the cases $q=2$ and 3, corresponding to rather trivial partitions, no α-partition exists with α reasonably large (say, $\alpha\geq0.5$).

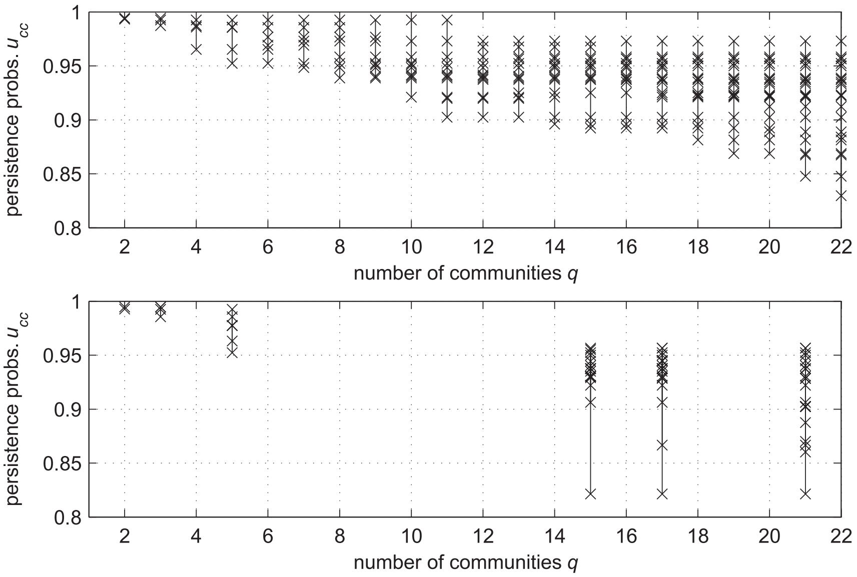

Figure 7. Comparison of two persistence probabilities' diagrams for the Netscience network. The two plots are in the same scale. Above: blow-up of the diagram of Fig. 6 (our results). Below: the diagram related to the six partitions proposed in [42].
doi:10.1371/journal.pone.0027028.g007

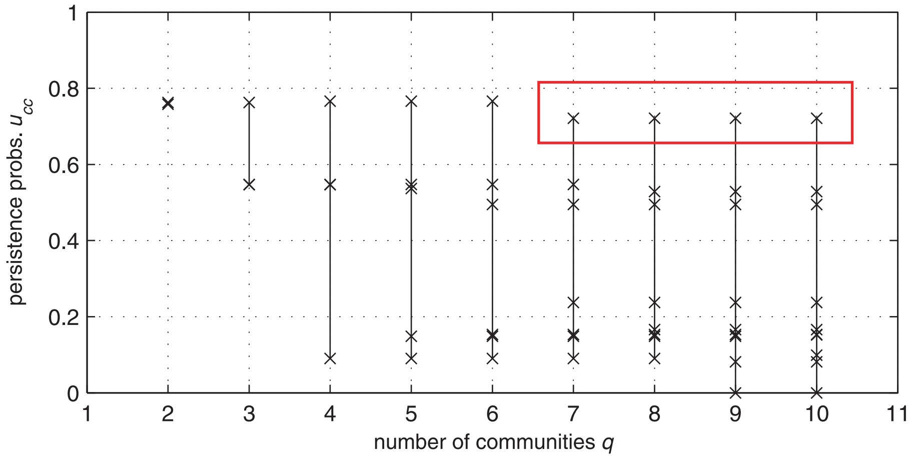

Figure 8. The persistence probabilities' diagram of the world trade network. Overall, the network does not display a definite community structure. However, a single cluster with rather strong internal connectivity ($u_{cc}=0.72$) is detected (as evidenced by the rectangle), corresponding to a subnetwork which includes almost all European countries.
doi:10.1371/journal.pone.0027028.g008

Nonetheless, an α-community which meets a restrictive quality standard is stably detected in a rather wide range of q, as highlighted in the figure. It is a cluster composed of 62 nodes which shows a rather strong internal connectivity ($u_{cc}=0.72$). Any other candidate cluster, instead, turns out to have a much smaller u_{cc} value and, therefore, it can hardly be considered to be a significant community. Interestingly, this meaningful cluster includes almost all European countries, plus a number of minor non-European partners.

Discussion

Comparison with other community detection methods

In the section "Methods" we highlighted some important connections between persistence probabilities and other quantities which are standard in graph theory. As we pointed out, for undirected networks u_{cc} reduces to the so-called *internal density*, namely the ratio between the total internal weight and the total strength of $\mathbb{C}_c$. In turn, if the graph is unweighed too, this turns out to be one minus the normalized cut of $\mathbb{C}_c$. The definition of "community in the weak sense", put forward in [14], can also be reinterpreted in terms of persistence probabilities. No straightforward connections, however, can be deducted for directed networks, where nonetheless the tool of persistence probabilities can be fully applied.

An important relationship between random walks and modularity is put forward by Kim et al. [28] who propose their *LinkRank modularity* Q^{lr} (that we denote by R for clarity), a variation to the standard modularity aimed at obtaining a better performance on directed graphs. In words, R is the difference between the fraction of time spent walking within communities (R') and the expected value of this fraction on a suitable null model (R''). Both these terms are additive with respect to communities, and it turns out that (with our notation):

$$R=\sum_{c=1}^{q} R'_c-\sum_{c=1}^{q} R''_c=\sum_{c=1}^{q}\sum_{i,j\in\mathbb{C}_c}\pi_i p_{ij}-\sum_{c=1}^{q}\sum_{i,j\in\mathbb{C}_c}\pi_i\pi_j. \quad (8)$$

In the case of undirected networks, simple computations show that $R'_c=W_c/S$ and $R''_c=S_c^2/(4S^2)$, which implies that the LinkRank modularity reduces to the standard one Q, which indeed can be written as:

$$Q=\sum_{c=1}^{q} Q'_c-\sum_{c=1}^{q} Q''_c=\sum_{c=1}^{q}\frac{W_c}{S}-\sum_{c=1}^{q}\left(\frac{S_c}{2S}\right)^2. \quad (9)$$

The comparison between $Q'_c=W_c/S$ and the persistence probability $u_{cc}=2W_c/S_c$ reveals obvious analogies but also subtle and important differences. The former is the fraction of time spent in community $\mathbb{C}_c$: being proportional to the total internal weight W_c, it will be smaller for smaller clusters, *ceteris paribus*, regardless to their cohesiveness. On the contrary, u_{cc} measures the probability of remaining in $\mathbb{C}_c$ given that the walker is currently there, regardless to the dimension of the cluster, thanks to the normalization by the total cluster strength S_c. The result is a superior capability of persistence probabilities is assessing the quality of clusters whatever their size is, a precious feature when analyzing multi-scale networks (i.e., composed of communities of different size scales).

This can be demonstrated, for example, by considering an instance of a LFR benchmark network with $\beta=1$, $\mu=0.25$ (see the section "Results" for the value of the other parameters) and analyzing the values of u_{cc} and $Q'_c(=R'_c)$ for a set of partitions. The network has 38 communities with size ranging from 11 to 77 nodes. We use our partitions generator to yield a family of $\mathbb{P}_q$ with $q\leq 50$. For each partition, we compute the set of persistence probabilities u_{cc} and the set of the fractions of time spent in the community Q'_c. The results are shown in the first and second panel of Fig. 9: as long as the considered $\mathbb{P}_q$ does not break any of the built-in clusters, all the u_{cc}-s remain large and concentrated in a rather narrow range, regardless to the cluster size. Then, some of them abruptly decreases as soon as clusters are broken. The Q'_c-s, on the contrary, are quite widespread (in a range from 1% to $6-7\%$) and vary in a rather smooth manner, since a smooth reduction of cluster sizes yields a corresponding smooth reduction of their internal weight W_c. It seems therefore that the fraction of time spent walking is not as indicative of the quality of a cluster as the persistence probability. The scenario does not modify if the *relative* fraction of time $Q_c=Q'_c-Q''_c$ (or *local modularity*) is considered, i.e., if the comparison with the performance of a null model is accounted for, as it appears from the third panel of Fig. 9. The obvious consequence is a very small sensitivity of the modularity $Q=\sum_c Q_c$, at least within this set of partitions. As shown in the bottom panel of Fig. 9, Q has almost the same value in $20\geq q\geq 40$, which makes questionable the reliability of choosing the max-modularity partition ($\mathbb{P}_{34}$, in this case). Further discussion on the use of absolute vs. relative (i.e., compared with a null model) cluster measures is in the Supporting Information S1.

The proposed approach has also important connections with two recently published community analysis methods. Delvenne et al. [42] show that the autocorrelation function of a signal emitted by a random walker, with value c as long as the walker is in a node $i\in\mathbb{C}_c$, can be expressed in terms of the clustered autocovariance matrix $R_t=H'[\mathrm{diag}(\pi)P^t-\pi'\pi]H$, and they define the stability of the partition H as $r_t^H=\min_{s=0,1,\ldots,t}\mathrm{trace}(R_s)$. Given a set of candidate partitions, the *graph stability* function $r_t=\max_H r_t^H$ puts in evidence, for each time instant t, which is the "optimal" partition. It is suggested in [42] that the most relevant partitions are those which are optimal over long time windows. It is straightforward to check that our matrix U is related to the step-1 autocovariance R_1 by $R_1+\Pi'\Pi=\mathrm{diag}(\Pi)U$. The two methods are thus based on the same ground, but our approach has two advantages: first, for each partition H we do not have to compute a long time-dependent sequence such as $R_1,R_2,\ldots,R_{t_{\max}}$ (with $t_{\max}$ of the same order as N) of $q\times q$ matrices, but the sole matrix U, with an important reduction in the computational burden. Second, the full list of the persistence probabilities u_{cc} allows one to test the quality of each single community, whereas the stability of the clustering r_t^H averages among all the communities.

Finally, a work with straightforward connections to ours is that of Weinan et al. [52], who suggest that the best q-community partition is that corresponding to the "best" (in a suitable technical sense) q-state approximated lumped Markov chain. This boils out

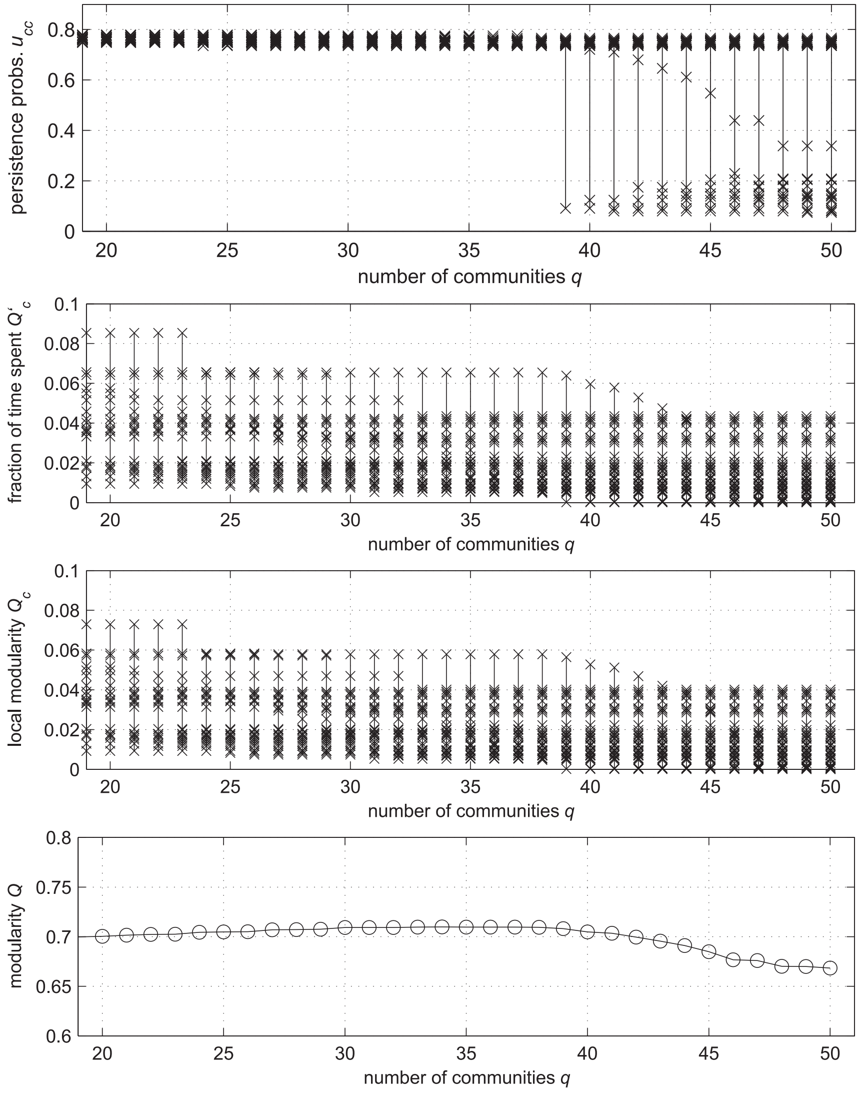

Figure 9. A comparison between persistence probabilities and fraction of time spent as indicators of the quality of a community. The test considers a LFR benchmark network with $\beta=1$, $\mu=0.25$ (see the text for the other parameters). For each candidate partitions $\mathbb{P}_q$, the four panels show, from the top to the bottom: the persistence probabilities u_{cc}; the fractions of time Q'_c spent in each community by a random walker; the difference of such fractions with those obtained in a null network model (local modularity Q_c); the modularity Q of the partition. Only the set of persistence probabilities shows a definite structural change in correspondence of the correct number of communities ($q=38$).
doi:10.1371/journal.pone.0027028.g009

to the formulation of a minimization problem, after a metric on the space of stochastic matrices is introduced. A drawback of this approach is however that q must be *a priori* specified, whereas often identifying the correct number of communities is the main goal of the analysis. For the same reason, it can hardly support the discussion of the significance and convenience of choosing one partition instead of another.

Concluding remarks

In this paper, we have shown that associating a lumped Markov chain to a given network partition (i.e., a set of communities) provides an effective tool for testing the significance of each single community and, consequently, of the entire partition. As a matter of fact, the diagonal terms (called persistence probabilities) of the lumped Markov matrix can be used as quality measures for each individual community. If a threshold level $0<\alpha<1$ is fixed, a sharp criterion for defining a community as "meaningful" is therefore that of requiring that its persistence probability is not less than α.

If an effective method for generating a set of "good" partitions is available, the above criterion can be used to rapidly select one of them among those complying with the prescribed α-quality, typically the one with the finest network decomposition (i.e., the largest number of communities). We have used a generator of partitions based on hierarchical cluster analysis, where the node distance is again defined on the basis of a Markov chain random walk model. Overall, the method has fair computational requirements, and can be applied to fully general networks (i.e., directed and weighted). Its effectiveness has been demonstrated on several medium-scale examples (see also the Supporting Information S1 for further case studies).

As already pointed out, the tool of persistence probabilities can be used to assess the quality of partitions, or single clusters, obtained with whatever method (e.g., modularity optimization) or *a priori* defined (e.g., geographical areas in the world trade network). Along this line, two possible extensions appear to be promising. One one side, several methods have recently been proposed to identify *overlapping* communities, i.e., clusters with shared nodes [53,54]. In principle, a lumped Markov chain can be associated to a *cover* as well (i.e., a clusterization with possible overlaps), although this requires a careful treatment of the shared nodes. Another extension concerns time-variant networks, namely networks whose edges (or their weights) vary in time (many examples can be found in social or economic networks). Once a community structure has been identified in a given time instant (i.e., on a "frozen" network), one may be interested in tracking the time evolution of the persistence probabilities, to reveal which communities remain significant in time or, on the contrary, which ones have a decaying cohesion [22,55]. These extensions will be the subject of future research.

Supporting Information

Supporting Information S1 **Figure S1.1.** The persistence probabilities' diagram of the Erdös-Rényi network. **Figure S1.2.** Zachary's karate club network. Above: The dendrogram obtained with $T=2$. Below: The persistence probabilities' diagram. **Figure S1.3.** The persistence probabilities' diagrams of two LFR directed, weighted benchmark networks. Top: $\mu_t=\mu_w=0.3$ (the number of planted communities is 35). Bottom: $\mu_t=\mu_w=0.6$ (42 planted communities). See the text for the other parameters. **Figure S1.4.** LinkRank benchmark network. **Figure S1.5.** The persistence probabilities' diagram of the LinkRank benchmark network. **Figure S1.6.** The persistence probabilities' diagram of the neural network. **Figure S1.7.** Absolute and relative persistence probabilities' diagrams of a LFR benchmark network. The relative persistence probability $r_{cc}=u_{cc}-Eu_{cc}$ compares the absolute one u_{cc} with the persistence probability Eu_{cc} of the same cluster in a null model.
(PDF)

Author Contributions

Conceived and designed the experiments: CP. Performed the experiments: CP. Analyzed the data: CP. Contributed reagents/materials/analysis tools: CP. Wrote the paper: CP.

References

1. Strogatz SH (2001) Exploring complex networks. Nature 410: 268–276.
2. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DH (2006) Complex networks: Structure and dynamics. Phys Rep 424: 175–308.
3. Barrat A, Barthlemy M, Vespignani A (2008) Dynamical Processes on Complex Networks. Cambridge University Press.
4. Newman MEJ (2010) Networks: An Introduction. Oxford University Press.
5. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582.
6. Guimera R, Sales-Pardo M, Amaral LAN (2007) Module identification in bipartite and directed networks. Phys Rev E 76: 036102.
7. Jonsson P, Cavanna T, Zicha D, Bates P (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7.
8. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments revealed in food-web structure. Nature 426: 282–285.
9. Piccardi C, Calatroni L, Bertoni F (2010) Communities in italian corporate networks. Physica A 389: 5247–5258.
10. Flake G, Lawrence S, Giles C, Coetzee F (2002) Self-organization and identification of web communities. Computer 35: 66–71.
11. Šubelj L, Bajec M (2011) Community structure of complex software systems: Analysis and applications. Physica A 390: 2968–2975.
12. Fortunato S (2010) Community detection in graphs. Phys Rep 486: 75–174.
13. Wasserman S, Faust K (1994) Social Network Analysis. Cambridge, UK: Cambridge University Press.
14. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci USA 101: 2658–2663.
15. Reichardt J, Bornholdt S (2006) When are networks truly modular? Physica D 224: 20–26.
16. Hu Y, Chen H, Zhang P, Li M, Di Z, et al. (2008) Comparative definition of community and corresponding identifying algorithm. Phys Rev E 78: 026121.
17. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.
18. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74: 036104.
19. Fortunato S, Barthelemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci USA 104: 36–41.
20. Hu Y, Nie Y, Yang H, Cheng J, Fan Y, et al. (2010) Measuring the significance of community structure in complex networks. Phys Rev E 82: 066106.
21. Kovacs IA, Palotai R, Szalay MS, Csermely P (2010) Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS One 5: e12528.
22. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS One 6: e18961.
23. Meyer C (2000) Matrix Analysis and Applied Linear Algebra. SIAM.
24. Kemeny JG, Snell JL (1976) Finite Markov Chains. Springer-Verlag.
25. Buchholz P (1994) Exact and ordinary lumpability in finite Markov-chains. J Appl Probab 31: 59–75.
26. Hoffmann KH, Salamon P (2009) Bounding the lumping error in Markov chain dynamics. Appl Math Lett 22: 1471–1475.
27. Fortunato S, Castellano C (2009) Community structure in graphs. In: Meyers, RA, editor, Encyclopedia of Complexity and System Science, Springer-Verlag Berlin. pp 1141–1163.
28. Kim Y, Son SW, Jeong H (2010) Finding communities in directed networks. Phys Rev E 81: 016103.
29. Reichardt J, Bornholdt S (2007) Partitioning and modularity of graphs with arbitrary degree distribution. Phys Rev E 76: 015102.
30. Sima J, Schaeffer S (2006) On the NP-completeness of some graph cluster measures. In: Wiedermann J, Tel G, Pokorny J, Bielikova M, Stuller J, eds. SOFSEM 2006: Theory and Practice of Computer Science, Proceedings. volume 3831 of *Lecture Notes in Computer Science* pp. 530–537.

31. Zhou H (2003) Distance, dissimilarity index, and network community structure. Phys Rev E 67.
32. Pons P, Latapy M () Computing communities in large networks using random walks. In: Yolum P, Gungor T, Gurgen F, Ozturan C, eds. Computer and Information Sciences - ISCIS 2005, ProceedingsYolum, P and Gungor, T and Gurgen, F and Ozturan, C, editor, Springer-Verlag Berlin, 2005 volume 3733 of *Lecture Notes In Computer Science*, pp. 284–293.
33. Fouss F, Pirotte A, Renders JM, Saerens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans Knowl Data Eng 19: 355–369.
34. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105: 1118–1123.
35. Steinhaeuser K, Chawla NV (2010) Identifying and evaluating community structure in complex networks. Pattern Recognit Lett 31: 413–421.
36. Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster Analysis, 5th ed. John Wiley & Sons.
37. Holgersson M (1978) The limited value of cophenetic correlation as a clustering criterion. Pattern Recognit 10: 287–295.
38. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78.
39. Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E 80.
40. Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech-Theory Exp. P09008 p.
41. Reichardt J, Bornholdt S (2004) Detecting fuzzy community structures in complex networks with a Potts model. Phys Rev Lett 93: 218701.
42. Delvenne JC, Yaliraki SN, Barahona M (2010) Stability of graph communities across time scales. Proc Natl Acad Sci USA 107: 12755–12760.
43. Narayanam R, Narahari Y (2011) A Shapley value-based approach to discover influential nodes in social networks. IEEE Trans Autom Sci Eng 8: 130–147.
44. Cafieri S, Hansen P, Liberti L (2011) Locally optimal heuristic for modularity maximization of networks. Phys Rev E 83: 056105.
45. Kannan R, Vempala S, Vetta A (2004) On clusterings: Good, bad and spectral. J ACM 51: 497–515.
46. Serrano MA, Boguñá M (2003) Topology of the world trade web. Phys Rev E 68: 015101.
47. Garlaschelli D, Loffredo MI (2005) Structure and evolution of the world trade network. Physica A 335: 138–144.
48. Fagiolo G, Reyez J, Schiavo S (2008) On the topological properties of the world trade web: a weighted network analysis. Physica A 387: 3868–3873.
49. He J, Deem MW (2010) Structure and response in the world trade network. Phys Rev Lett 105: 198701.
50. Barigozzi M, Fagiolo G, Mangioni G (2011) Identifying the community structure of the international-trade multi-network. Physica A 390: 2051–2066.
51. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech-Theory Exp. P10008 p.
52. Weinan E, Li T, Vanden-Eijnden E (2008) Optimal partition and effective dynamics of complex networks. Proc Natl Acad Sci USA 105: 7907–7912.
53. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435: 814–818.
54. Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11: 033015.
55. Fenn DJ, Porter MA, McDonald M, Williams S, Johnson NF, et al. (2009) Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007-2008 credit crisis. Chaos 19: 033119.