Community analysis in directed networks: In-, out-, and pseudocommunities

Pietro Landi and Carlo Piccardi*

Department of Electronics, Information and Bioengineering, Politecnico di Milano, I-20133 Milano, Italy

(Received 2 July 2013; revised manuscript received 3 December 2013; published 28 January 2014)

When analyzing important classes of complex interconnected systems, link directionality can hardly be neglected if a precise and effective picture of the structure and function of the system is needed. If community analysis is performed, the notion of "community" itself is called into question, since the property of having a comparatively looser external connectivity could refer to the inbound or outbound links only or to both categories. In this paper, we introduce the notions of *in-*, *out-*, and *in-/out-community* in order to correctly classify the directedness of the interaction of a subnetwork with the rest of the system. Furthermore, we extend the scope of community analysis by introducing the notions of *in-*, *out-*, and *in-/out-pseudocommunity*. They are subnetworks having strong internal connectivity but also important interactions with the rest of the system, the latter taking place by means of a minority of its nodes only. The various types of (pseudo-)communities are qualified and distinguished by a suitable set of indicators and, on a given network, they can be discovered by using a "local" searching algorithm. The application to a broad set of benchmark networks and real-world examples proves that the proposed approach is able to effectively disclose the different types of structures above defined and to usefully classify the directionality of their interactions with the rest of the system.

DOI: 10.1103/PhysRevE.89.012814 PACS number(s): 89.75.Hc, 89.75.Fb

I. INTRODUCTION

Networks are becoming more and more important for modeling, analyzing, and controlling large-scale complex systems [1–5]. The availability of effective methods and algorithms is crucial to disclose their main structural features, a fundamental step to understanding a number of static and dynamic properties, such as functional roles, robustness to failures, spreading dynamics, or collective behaviors.

One of the most studied problems concerning network structure is *community analysis*, which is aimed at revealing possible subnetworks (i.e., groups of nodes called *communities*, *clusters*, or *modules*) characterized by comparatively large internal connectivity, namely whose nodes tend to connect much more with the other nodes of the group than with the rest of the network. A huge number of contributions have explored the theoretical aspects of community analysis and proposed a broad set of algorithms for community detection [6]. Most notably, community analysis has revealed to be a powerful tool for deeply understanding the properties of a number of real-world complex systems in virtually any field of science, including biology [7], ecology [8], economics [9], information [10,11], and social sciences [12,13].

Despite the extremely large number of contributions on community analysis in general, only limited effort has been done on *directed networks* [14–18]. It is true that several methods, naturally designed for undirected networks, lend themselves to be automatically applied to directed networks with no modifications (e.g., Ref. [19]) and that others can be straightforwardly extended (e.g., Ref. [20]). But rarely the very meaning of community is discussed in the new context, although the link directionality calls into question basic notions such as internal versus external connectivity of a subnetwork, not to mention the effectiveness of algorithms which were not purposely designed.

Yet, when considering several classes of complex interconnected systems, directedness can hardly be neglected if one wants to get a realistic picture of the structure and functioning of the system. Examples can be found in many fields, including biogenetics [21], neural sciences [22], natural resource management [23], transportation systems [24], and information sciences [25], just to name a few. The abundance of real-world examples of directed networks has stimulated theoretical developments, too, specifically those oriented to highlighting the properties of this class of networks and contrasting them with those of the simpler undirected counterpart (e.g., Refs. [26,27]). For all the above reasons, it follows that community analysis in directed networks is an issue that still needs to be explored in detail.

Several methods for community analysis are based on information processing related to a random walker: examples include Infomap [28], Linkrank [18], Stability [19], and lumped Markov chains [29], among others. The standard model is the following: at each (discrete) time step, the walker which is in node i randomly selects one of the out-links $i \to j$ with a probability proportional to the link weight and, accordingly, follows the selected out-link to reach the neighbor j. The induced notion of community is that of a subnetwork with a large escape time, i.e., such that the walker at each step has a large probability of remaining within the subnetwork. If the network is directed, however, this approach introduces a bias in that such a community has surely *weak out-links* towards the rest of the network, but it might have *strong in-links*, too, i.e., links with large weights pointing from the outside towards the community. For example, think about the network of trades among countries, where the weight of $i \to j$ is the monetary value of the export from country i to country j. According to the above standard notion, the countries of a strongly significant community would *export* much more within the community rather than outside. But this does not exclude that they might have large *import* flows from the outside, which is in contrast with the idea of a community as a set weakly connected to the rest of the network. This

*Corresponding author: carlo.piccardi@polimi.it

calls for the new, distinct notions of *out-* and *in-community*: The former is a subnetwork whose nodes direct most of their out-strength to nodes lying within the community rather than to the rest of the network, whereas the latter is such that nodes receive most of their in-strength from within the community rather than from the outside. A subnetwork with both features will be denoted as *in-/out-community*.

The other new notion introduced by this paper is that of *pseudocommunity*. Several case studies reveal the existence of peculiar structures, namely "starlike" subnetworks, in which most of the nodes direct most of their out-strength within the subnetwork (often towards a single "central" node), but a few of them (typically the "central" node only) mainly direct their out-strength to the outside. Examples are found in the worldwide air transportation network, where regional "hubs" collect all the traffic originating from domestic airports and forward it to the rest of the world, or in the trade web of specific commodities, where a leading country monopolizes the purchase of raw material from many producers and exports finished products to the rest of the world (see Sec. V for details on these and other examples). In a subnetwork with such features, a random walker has a small escape time due the large out-strength from the "central" node to the outside: the subnetwork cannot therefore be qualified as a (out-)community. Yet such a structure is worth being revealed and classified, since it has a special form of strong intraconnectivity: We will denote it as *out-pseudocommunity*. An *in-pseudocommunity* will be a subnetwork where most of the nodes receive most of their in-strength from the community rather than from the rest of the network, but a few nodes have instead a large in-strength from the outside. A subnetwork with both features will be denoted as *in-/out-pseudocommunity*.

Two indicators will be used to quantify the out-properties of a subnetwork S, namely the *persistence probability* α_S and the *average internal strength* β_S. Their values will be used to possibly classify S as out-community or out-pseudocommunity. Another two indicators α'_S and β'_S will be used to quantify the in-properties of S and to possibly classify it as in-community or in-pseudocommunity. If we combine the in- and out-features, namely we consider the value of the 4-tuple $(\alpha_S,\beta_S,\alpha'_S,\beta'_S)$, we discover that a directed network can contain eight different nontrivial types of *structures*, i.e., subnetworks with peculiar properties: *out-community*; *in-community*; *in-/out-community*; *out-pseudocommunity*; *in-pseudocommunity*; *in-/out-pseudocommunity*; *in-pseudocommunity*/*out-community*; *in-community*/*out-pseudocommunity*. Each one of them corresponds to a specific combination of in-/out- as well as intra-/interconnectivity.

Having defined the types of structures we are interested in, we need an algorithm to discover them in a given network. We use a local approach, similar in spirit to a few recent proposals [30–32], namely an algorithm that, starting from each node, is aimed at finding the smallest (pseudo-)community which is, at the same time, significant (as measured by the above indicators) and locally maximal (i.e., it is worsened by any further node inclusion). Differently from many community analysis methods, this algorithm does not yield a partition of the network, i.e., a node might be not included in any structure. This seems perfectly reasonable, however, as it is not uncommon to discover strongly connected groups of nodes even in networks which, overall, do not possess a definite clusterized structure. Forcing a partition in such networks places side by side high- and low-quality clusters, often without the capability of discriminating among them. On the other hand, the above algorithm highlights significant structures only and fully allows for overlapping, as one node may belong to more than one (pseudo-)community. But there is also a "second level" of overlapping, remarkably, since a node could be at the same time part of structures of different types, e.g., an out-community and an in-pseudocommunity, sharing these memberships with different sets of partners (see examples in Sec. V).

The paper is organized as follows. First, we formally introduce and discuss the notions of in-/out-communities and in-/out-pseudocommunities, illustrating them with the help of simple networks. Then we propose an algorithm for finding the eight types of structures in a given network. Finally, to demonstrate the power of the method, we analyze a large number of benchmark and real-world networks: Several examples of the structures that are discovered are discussed in detail, proving that the proposed methodology has the capability of revealing structures undetected by previously available approaches (e.g., pseudocommunities) or to classify them more effectively (e.g., clarifying their role of in-, out-, or in-/out-communities).

II. COMMUNITIES AND PSEUDOCOMMUNITIES

Consider a directed, weighted network with nodes $N = \{1,2,\ldots,n\}$ and *weight matrix* $W = [w_{ij}]$, i.e., $w_{ij} > 0$ denotes the weight of the link $i \to j$, which is set to 1 when the network is binary (i.e., unweighed), while $w_{ij} = 0$ if the link $i \to j$ does not exist. Denote by $s_i^{\text{in}} = \sum_j w_{ji}$ and $s_i^{\text{out}} = \sum_j w_{ij}$, respectively, the *in-* and *out-strength* of node i, which reduce to the *in-* and *out-degree* if the network is binary. We set $w_{ii} = 0$ for all i, namely we remove self-loops, if any. This does not affect the internode connectivity, which will be quantified by a pair of indicators that would otherwise be distorted (see below).

A n-state discrete-time Markov chain can be associated to the network in a standard fashion. For that, we denote by $p_{ij} = w_{ij}/s_i^{\text{out}}$ the probability that, at each time step, a random walker which is in node i jumps to j, so the probability $\pi_{i,t}$ of finding the walker in node i at time t is governed by $\pi_{t+1} = \pi_t P$, with $\pi_t = \left(\pi_{1,t}\ \pi_{2,t}\ \ldots\ \pi_{n,t}\right)$ and $P = [p_{ij}]$. We assume, for the moment, that the network is strongly connected [2,3] (we will remove this assumption later). This implies that P is an irreducible matrix, so the stationary probability distribution $\pi = \pi P$ is unique and strictly positive ($\pi_i > 0$ for all i) [33]: Its entry π_i is the fraction of time steps spent by the random walker, in the long run ($t \to \infty$), on node i.

Let us denote by S the *subnetwork* formed by a subset $N_S \subset N$ of the nodes of the original network and by all the links of the latter connecting pairs of nodes of S (induced subgraph). Subnetworks are candidates to be (pseudo-)communities, so we need a set of suitable indicators to quantify their features.

The first of such indicators is the *persistence probability* α_S: It is the probability that a random walker, which is in any of the nodes of S at time t, remains in S at time $t+1$. If we assume that the Markov chain $\pi_{t+1} = \pi_t P$ is in the stationary

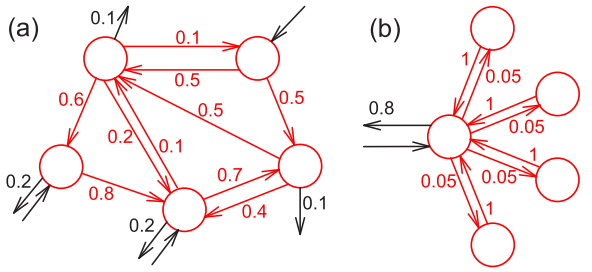

FIG. 1. (Color online) (a) A subnetwork S with large persistence probability α_S (*community*) and (b) a subnetwork S with large average internal strength β_S but small persistence probability α_S (*pseudocommunity*). The values of $p_{ij} = w_{ij}/s_i^{\text{out}}$ are shown on the graph.

state π, then we have [29]

$$\alpha_S = \sum_{i\in N_S} \frac{\pi_i}{\Pi_S} \sum_{j\in N_S} p_{ij} = \sum_{i\in N_S} \frac{\pi_i}{\Pi_S} \sum_{j\in N_S} \frac{w_{ij}}{s_i^{\text{out}}}, \qquad (1)$$

where $\Pi_S = \sum_{i\in N_S} \pi_i$ is the aggregated stationary probability of the subnetwork S. The persistence probability α_S is a measure of cohesiveness of S [indeed, the expected escape time from S is $(1-\alpha_S)^{-1}$] and it proved to be an effective tool for the structural analysis of networks [29,34]. As it is apparent from (1), α_S is a convex combination of the fraction of the out-strength of the nodes of S that is directed within S: The coefficient of the term i is the (normalized) stationary probability π_i, a well-known measure of centrality of node i [3]. Figure 1(a) shows a five-node subnetwork where, due to the balance of internal and external weights, a random walker has a large probability of circulating for long before exiting: indeed $\alpha_S \geqslant 0.8$, since 0.8 is the minimal value, over all five nodes, of the fraction $\sum_{j\in N_S} w_{ij}/s_i^{\text{out}}$. Notice that the exact value of α_S cannot be computed without knowing π, hence, the entire network. As we will discuss precisely in the following, subnetworks qualified as *communities* will be characterized by large values of α_S. Notice that, having ruled out self-loops, α_S only depends on the internode connectivity.

Measuring the persistence probability alone may fail to reveal some interesting structures in the network. The subnetwork S of Fig. 1(b) is a case in point: Due to the large probability of escaping from the central node, α_S is likely to be small. Yet the structure appears to be definitely interesting and worth revealing. We try to capture it by associating a second indicator to S, which we call *average internal strength* β_S,

$$\beta_S = \frac{1}{n_S} \sum_{i,j\in N_S} p_{ij} = \sum_{i\in N_S} \frac{1}{n_S} \sum_{j\in N_S} \frac{w_{ij}}{s_i^{\text{out}}}, \qquad (2)$$

where n_S is the number of nodes of S. The quantity $0 \leqslant \beta_S \leqslant 1$ is simply the *arithmetic* mean, over the nodes of S, of the fraction of the out-strength directed internally to S (we recall that α_S is a *weighted* mean of the same quantities). Thus β_S will be large when *most of the nodes of S direct most of their out-strength within S*, although a few others could do the opposite, yielding a small α_S: This is the case of the subnetwork of Fig. 1(b), which has $\beta_S = 0.84$, whereas α_S could be as small as 0.2. We define *pseudocommunity* a subnetwork which has small α_S but large β_S. Indeed, it is not a community in the usual sense (i.e., with strong intra- and weak interconnectivity) but it has nonetheless a special form of strong intraconnectivity, as most of the nodes have their most important connections inside the subnetwork rather than outside. We will encounter several pseudocommunities in the case studies of Sec. V.

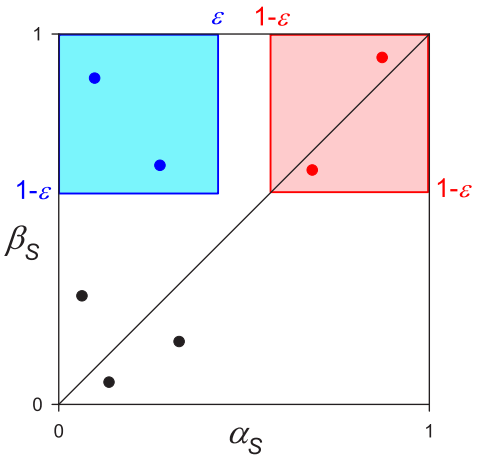

FIG. 2. (Color online) Each subnetwork S is mapped into a point in the unit square of the (α_S,β_S) plane. Points ε close to the $(1,1)$ vertex are ε communities, whereas points ε close to the $(0,1)$ vertex are ε pseudocommunities.

Once α_S and β_S are defined, each subnetwork S can be represented by a point in the unit square $[0,1]\times[0,1]$ of the plane (α_S,β_S) (see Fig. 2). When this point falls close to one of the vertices of the square, we have the most interesting cases for the classification of S as follows:

(i) $(\alpha_S,\beta_S)\to(1,1)$: most of the nodes of S, including those with large centrality π_i, direct internally most of their out-strength: S is a *community*.

(ii) $(\alpha_S,\beta_S)\to(0,1)$: most of the nodes of S, but not those with largest centrality π_i, direct internally most of their out-strength: S is a *pseudocommunity*.

(iii) $(\alpha_S,\beta_S)\to(1,0)$: $\beta_S\to 0$ reveals that most of the nodes direct externally most of their out-strength (with the exception of a few nodes where the centrality π_i concentrates, hence, $\alpha_S\to 1$): as it is not cohesive, this subnetwork is therefore not interesting for (pseudo-)community analysis (we also anticipate that structures like this are never found in real-world case studies).

(iv) $(\alpha_S,\beta_S)\to(0,0)$: in this trivial case, S is a subnetwork with no special properties.

In practice, a quality threshold $\varepsilon > 0$ is needed to analyze and classify a concrete subnetwork: we will define S as an ε *community* when

$$\|(\alpha_S,\beta_S)-(1,1)\|_\infty \leqslant \varepsilon, \qquad (3)$$

namely when $\alpha_S \geqslant 1-\varepsilon$ and $\beta_S \geqslant 1-\varepsilon$. Similarly, we will define S as an ε *pseudocommunity* when

$$\|(\alpha_S,\beta_S)-(0,1)\|_\infty \leqslant \varepsilon, \qquad (4)$$

namely when $\alpha_S \leqslant \varepsilon$ and $\beta_S \geqslant 1-\varepsilon$. Geometrically, constraints (3) and (4) define two square regions in the (α_S,β_S) plane, close to the respective vertices of the unit square (Fig. 2).

A. In- and out- (pseudo-)communities

The indicators so far introduced are biased on the properties of *out-connectivity* of the subnetwork S. This is a consequence of the definition of the random walk dynamics, which defines p_{ij} as a function of the *out-strength* s_i^{out}. In other words, α_S and β_S try to capture whether the nodes of S "influence" mostly the other nodes of S, rather than the rest of the network. Obviously, it would equally be interesting to know whether they are also mostly "influenced" by the other nodes of S rather than by the external nodes. Since the network is directed, the two properties are fully independent. Think, for example, about the world trade network: The quantities α_S and β_S tell us whether the countries of a given set export preferentially within the set itself but give us no information on whether also the import flow comes preferentially from countries of the same set or from the outside. This calls for a definition of suitable indicators of *in-connectivity*.

First, we rename α_S and β_S as *out-persistence probability* and *out-average internal strength*, respectively. Now the most natural way to define the corresponding *in-* indicators α'_S and β'_S is to reverse the direction of each link in the original network and then to define the new indicators by using, on this new network, the same definitions used above for α_S and β_S. This means considering the network defined by the weight matrix $W' = W^T$, so the random walker dynamics is now governed by the Markov matrix $P' = [p'_{ij}]$, with $p'_{ij} = w_{ji}/s_i^{\text{in}}$ (notice that $P' \neq P^T$).

Revisiting now the definitions of the previous section, we will say that S is an *out-community* when $(\alpha_S,\beta_S) \to (1,1)$, whereas it is an *out-pseudocommunity* when $(\alpha_S,\beta_S) \to (0,1)$. We will say that S is an *in-community* when $(\alpha'_S,\beta'_S) \to (1,1)$, whereas it is an *in-pseudocommunity* when $(\alpha'_S,\beta'_S) \to (0,1)$. But obviously each subnetwork should be contemporarily characterized by its in and out attributes. To get a complete picture, thus, we have to associate the 4-tuple $(\alpha_S,\beta_S,\alpha'_S,\beta'_S)$ to S and assess whether it is close to special vertices or edges of the unit *tesseract*, i.e., the unit hypercube $[0,1]^4$. As a matter of fact, by extending the discussion of the previous section we arrive at the definition of eight possible types of structures of interest for (pseudo-)community analysis (summarized in Fig. 3) as follows:

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (1,1,-,0)$: *out-community*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (-,0,1,1)$: *in-community*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (1,1,1,1)$: *in-/out-community*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (0,1,-,0)$: *out-pseudocommunity*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (-,0,0,1)$: *in-pseudocommunity*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (0,1,0,1)$: *in-/out-pseudocommunity*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (1,1,0,1)$: *in-pseudocommunity/out-community*

$(\alpha_S,\beta_S,\alpha'_S,\beta'_S) \to (0,1,1,1)$: *in-community/out-pseudocommunity*.

In the above list, a dash indicates that the value of the corresponding quantity is irrelevant since, as discussed above, S has no out-relevance [respectively, in-relevance] when the pair (α_S,β_S) [respectively, (α'_S,β'_S)] tends either to (0,0) or to (1,0).

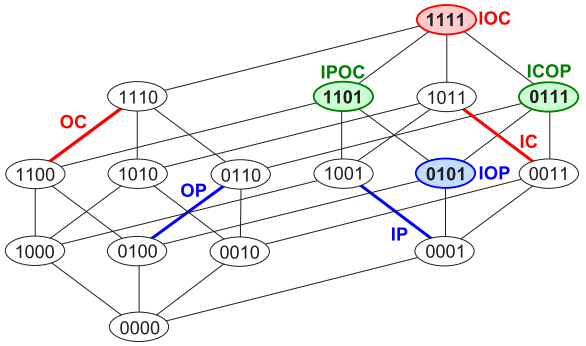

FIG. 3. (Color online) A projection of the unit tesseract, i.e., the unit hypercube $[0,1]^4$. (Pseudo-)communities are found when the 4-tuple $(\alpha_S,\beta_S,\alpha'_S,\beta'_S)$ is close to one of the four vertices or to one of the four edges highlighted in the figure (I = in, O = out, C = community, P = pseudocommunity).

We point out that, in the above definitions (and throughout the paper), the slash symbol denotes that *both* qualifiers apply. Thus, for instance, "in-/out-community" means that S is both "in-community" and "out-community," and "in-pseudocommunity/out-community" means that S is both "in-pseudocommunity" and "out-community."

Finally, we say that S is an ε "structure", where the term "structure" denotes one of the eight types of (pseudo-) communities above defined, if $(\alpha_S,\beta_S,\alpha'_S,\beta'_S)$ has (max-) distance ϕ_S not larger than ε from the relevant vertex or edge of the unit hypercube, i.e.,

ε *out-community*:
$\phi_S^{\text{oc}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (1,1,-,0)\|_\infty = \max\{1-\alpha_S, 1-\beta_S,\beta'_S\} \leqslant \varepsilon$

ε *in-community*:
$\phi_S^{\text{ic}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (-,0,1,1)\|_\infty = \max\{\beta_S,1-\alpha'_S, 1-\beta'_S\} \leqslant \varepsilon$

ε *in-/out-community*:
$\phi_S^{\text{ioc}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (1,1,1,1)\|_\infty = \max\{1-\alpha_S, 1-\beta_S,1-\alpha'_S,1-\beta'_S\} \leqslant \varepsilon$

ε *out-pseudocommunity*:
$\phi_S^{\text{op}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (0,1,-,0)\|_\infty = \max\{\alpha_S, 1-\beta_S,\beta'_S\} \leqslant \varepsilon$

ε *in-pseudocommunity*:
$\phi_S^{\text{ip}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (-,0,0,1)\|_\infty = \max\{\beta_S,\alpha'_S, 1-\beta'_S\} \leqslant \varepsilon$

ε *in-/out-pseudocommunity*:
$\phi_S^{\text{iop}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) -,(0,1,0,1)\|_\infty = \max\{\alpha_S,1-\beta_S, \alpha'_S,1-\beta'_S\} \leqslant \varepsilon$

ε *in-pseudocommunity/out-community*:
$\phi_S^{\text{ipoc}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (1,1,0,1)\|_\infty = \max\{1-\alpha_S, 1-\beta_S,\alpha'_S,1-\beta'_S\} \leqslant \varepsilon$

ε *in-community/out-pseudocommunity*:
$\phi_S^{\text{icop}} = \|(\alpha_S,\beta_S,\alpha'_S,\beta'_S) - (0,1,1,1)\|_\infty = \max\{\alpha_S, 1-\beta_S,1-\alpha'_S,1-\beta'_S\} \leqslant \varepsilon.$

It is worthwhile to mention that our ϕ_S's generalize and extend, in many directions, previous proposals for quantifying the cohesiveness of S. With reference to undirected binary

networks, Radicchi *et al.* [35] defined *community in a weak sense* a set S of nodes whose connections within S are more than those outside S: a very mild requirement for a baseline level of significance. It can easily be verified that this corresponds to $\alpha_S > 0.5$. Equivalent notions (e.g., the *normalized cut* of S, Ref. [6], p. 92) are used by traditional graph partition techniques or, more recently, by local community detection algorithms [36]. Our generalization operates in many directions. First, we extend the scope of application to the general case of directed, weighted networks and, consequently, we separately assess the in- and out- cohesiveness of S. Second, we introduce a new indicator, β_S, which is truly independent of α_S and, as such, allows us to enhance the analysis of the internal versus external link balancing. Third, we allow a flexible selection of the quality (i.e., cohesiveness) of (pseudo-)communities through ε, instead of having it rigidly fixed as in the above-recalled definition. In this respect, it is natural to set ε not larger than 0.5 to generalize the notion of community in a weak sense (e.g., for in-/out-communities, accepting that $\phi_S > 0.5$ could mean $\alpha_S < 0.5$). This means that $\phi_S \leqslant 0.5$ is a minimal requirement for significance, which can, however, possibly be strengthened by fixing lower values of ε. In the following (Secs. IV and V) we will see that fixing ε does not affect the (pseudo-)community detection procedure but allows a selection *a posteriori* among the set of identified (pseudo-)communities.

B. Nonconnected networks

So far we have assumed the strong connectedness of the network, namely the existence of a directed path $i \to j$ for any node pair (i, j). This guarantees that π_i is univocally defined and positive for any node i, which is necessary to have α_S well defined for any possible subnetwork S. Several real-world networks turn out not to have this property, and restricting the analysis to the sole strongly connected component could lead to overlooking interesting structures. Consider, for instance, the subnetwork S of Fig. 4: We would spontaneously tend to classify it as an out-pseudocommunity, since the escape time of a random walker is surely small but, on the other hand, four nodes of five direct all their out-strength within the subnetwork itself. Indeed, $\beta_S = 0.8$, but α_S is not defined since $\Sigma_{i\in S}\pi_i = 0$. Or consider again Fig. 4 but reverse all the directions: Then almost all nodes receive all their in-strength from inside S, but here the random walk dynamics is not even well defined because there are nodes with no out-links ("dangling nodes").

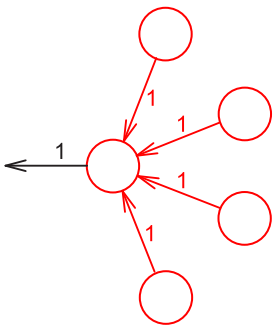

FIG. 4. (Color online) A subnetwork S with zero aggregated stationary probability ($\Sigma_{i\in S}\pi_i = 0$): The persistence probability α_S cannot be defined.

The simplest way to cope with this issue is to adopt the well-known approach used in the PageRank computation [3,37], namely to virtually transform the network into a strongly connected one by introducing a *teleportation* mechanism that, at each time step, with probability $1 - \gamma$ moves the random walker towards a node uniformly chosen. This approach has several drawbacks, however, which can be summarized in a strong sensitivity to γ [37]. Here we adopt a recent proposal, denoted *unrecorded link teleportation* [38], which modifies the standard scheme in two aspects: first, the probability of being teleported to node i is proportional to the out-strength s_i^{out} (this is essentially equivalent to selecting a *link* at random instead of a node); second, only steps along original links, and not teleportations, are recorded in the transitions rate among nodes. As extensively verified on benchmark and real-world networks [38], this computational scheme proves dramatically robust in our framework, namely when the random walker dynamics is at the basis of a community detection method.

The procedure is as follows (see Ref. [38] for detailed motivations and analysis). First, the *preference vector* is defined as $v = (s_1^{\text{out}} s_2^{\text{out}} \dots s_n^{\text{out}})/\sum_i s_i^{\text{out}}$, and possible zero rows of P (corresponding to dangling nodes, i.e., $s_i^{\text{out}} = 0$) are replaced by v, to have a well-defined random walk dynamics. Then, the dynamics with teleportation is defined by a Markov matrix $\tilde{P} = [\tilde{p}_{ij}]$ defined by

$$\tilde{p}_{ij} = \gamma p_{ij} + (1 - \gamma) v^{[i]}, \tag{5}$$

where $v^{[i]}$ is obtained from v by letting $v_i = 0$ (to avoid self-loops) and renormalizing to have sum 1. The stationary distribution π, to be used together with P in computing the persistence probabilities (1), is finally given by $\pi = \tilde{\pi} P$, where $\tilde{\pi} = \tilde{\pi}\tilde{P}$ is the stationary solution of the dynamics with teleportation (5). To summarize, if the original network is strongly connected, we let $\gamma = 1$, leaving it unaltered. Otherwise, we set γ to the standard value 0.85 [37]: As extensively verified in Ref. [38] and confirmed by our experiments, the results of community detection are largely insensitive to rather broad variations of γ in the neighborhood of this value.

C. Examples

The toy network of Fig. 5(a) was proposed by Rosvall and Bergstrom [28] to highlight the different results obtained by their Infomap method (based on random walk dynamics) and by modularity optimization on directed networks [16]. The former approach identifies the four-node "ring" subnetworks (e.g., {1,2,3,4}) as the most significant ones, due to their larger escape time, whereas the latter prefers four-node "inter-ring" subnetworks (e.g., {1,2,7,8}) because of the larger internal weight. Our classification, which is based on random walk dynamics, too, consistently favors the "ring" subnetworks ($\phi_S^{\text{ioc}} = 0.33$) over the "inter-ring" ones ($\phi_S^{\text{ioc}} = 0.42$) if they are assessed as *in-/out-communities*.

If we reverse the direction of half of the "inter-ring" links [Fig. 5(b)], we obtain that two of the "ring" subnetworks ({1,2,3,4} and {9,10,11,12}) have rather large

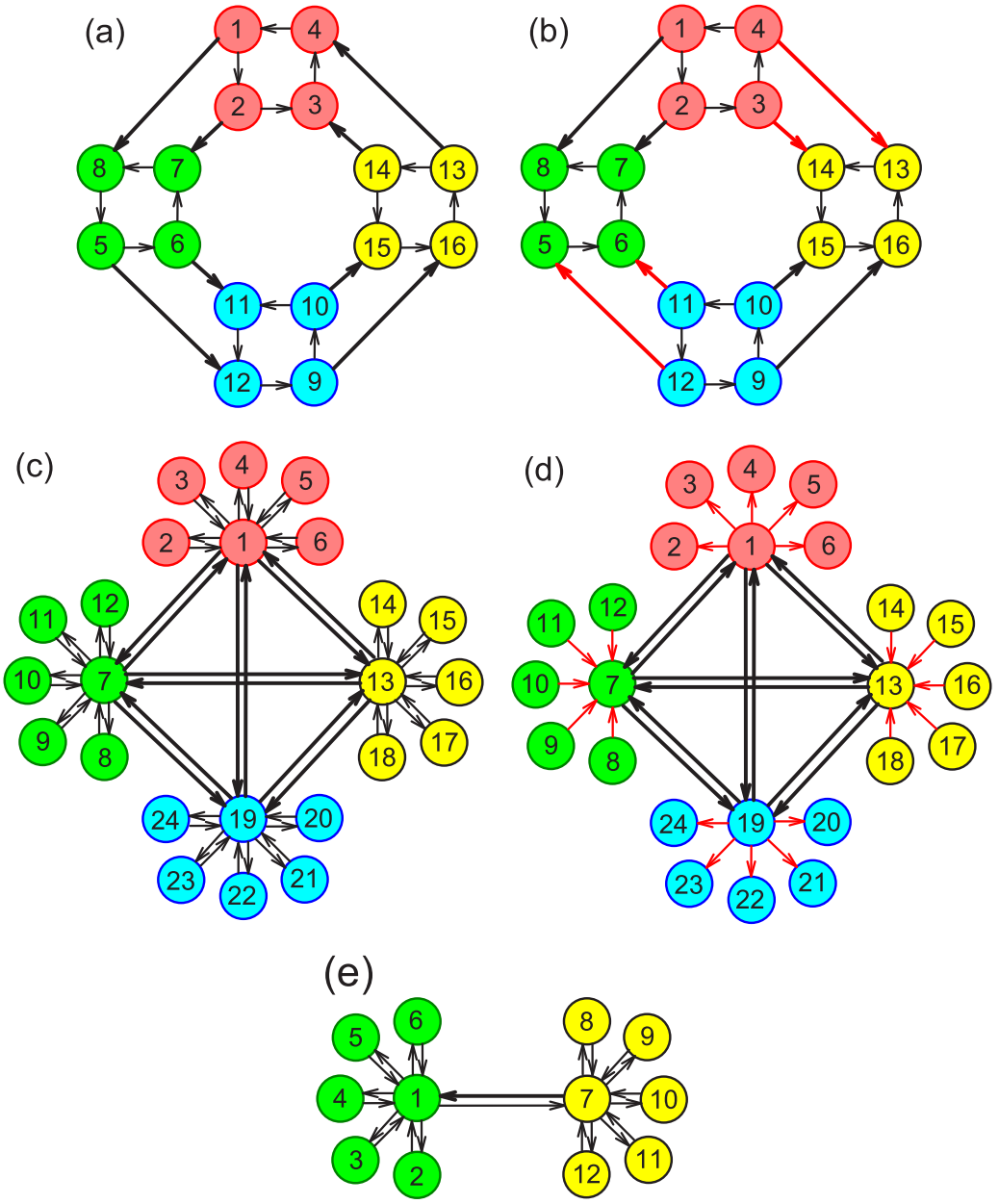

FIG. 5. (Color online) Five networks designed to clarify the definitions and properties of communities and pseudocommunities. In (a) and (b), "inter-ring" links have a weight that is double that of "intraring" links. Network (a) has four *in-/out-communities* (the four-node "ring" subnetworks). Network (b), which has four links reversed with respect to (a), has two *in-communities* ($\{1,2,3,4\}$ and $\{9,10,11,12\}$) and two *out-communities* ($\{5,6,7,8\}$ and $\{13,14,15,16\}$). In (c) and (d), nodes can be classified as "leaders" (nodes 1, 7, 13, 19) and "followers." Leader-leader links have a weight that is 20 times larger than that of leader-follower and follower-leader links. Network (c) has four *in-/out-pseudocommunities* (the four leader-followers subnetworks). Network (d), where leader-followers links have been removed in one direction, has two *in-pseudocommunities* ($\{1,2,\ldots,6\}$ and $\{19,20,\ldots,24\}$) and two *out-pseudocommunities* ($\{7,8,\ldots,12\}$ and $\{13,14,\ldots,18\}$). In (e), link $7 \to 1$ has a weight that is 20 times larger than that of all the others. The network has an *in-community*/*out-pseudocommunity* ($\{7,8,\ldots,12\}$) and an *in-pseudocommunity*/*out-community* ($\{1,2,\ldots,6\}$).

out-connectivity but no in-connectivity, while the other two ($\{5,6,7,8\}$ and $\{13,14,15,16\}$) have opposite properties. As a consequence, the former can be qualified as *in-communities* (they have $\phi_S^{\mathrm{ic}} = 0.33$, while ϕ_S^{oc} and ϕ_S^{ioc} are as large as 0.96 and 0.67, respectively) and the latter as *out-communities*. To find significant *in-/out-communities* in this network, we must glue together two "rings": for example, $\{1,2,\ldots,8\}$ has $\phi_S^{\mathrm{ioc}} = 0.18$ (while, e.g., $\{1,2,3,4\}$ alone has $\phi_S^{\mathrm{ioc}} = 0.67$).

The networks of Figs. 5(c) and 5(d) are instead proposed to highlight the different forms of pseudocommunity. In Fig. 5(c), the four subnetworks of the type "leader + followers" (e.g., node 1 with $\{2,3,4,5,6\}$) can be classified as significant *in-/out-pseudocommunities*. Indeed, they have large average internal strength $\beta_S = \beta'_S = 0.85$ but small persistence probability $\alpha_S = \alpha'_S = 0.14$ (thus $\phi_S^{\mathrm{iop}} = 0.15$), because of the large weight of the link connecting the leader with the rest of the

network. In Fig. 5(d), where the links between the leader and followers are removed in one direction, two of these subnetworks ($\{1,2,\ldots,6\}$ and $\{19,20,\ldots,24\}$) become *in-pseudocommunities*, with $\phi_S^{\text{ip}} = 0.19$, while the others become *out-pseudocommunities*.

Finally, the network of Fig. 5(e) highlights the two hybrid types of structure. Due to the large weight of the link $7 \to 1$ escaping from the right-hand half (nodes $\{7,8,\ldots,12\}$), this latter subnetwork has small $\alpha_S = 0.33$, although β_S is as large as 0.87. It is therefore an out-pseudocommunity but, at the same time, it is strongly cohesive from the in- point of view, since $\alpha'_S = 0.91$, $\beta'_S = 0.97$. Thus $\phi_S^{\text{icop}} = 0.33$, which qualifies the subnetwork as a rather significant *in-community*/*out-pseudocommunity*. The opposite takes place on the left-hand half of the network (nodes $\{1,2,\ldots,6\}$), which has $\phi_S^{\text{ipoc}} = 0.33$ and is therefore worth being qualified as an *in-pseudocommunity*/*out-community*.

D. Undirected networks

Although this paper is essentially focused on *directed* networks, as they potentially display the richest and most interesting structures, the same classifications and methods can be applied to *undirected* networks as well. In this case, obviously, the distinction between in- and out-structures no longer exists. More precisely, the symmetry $W = W^T$ implies $W' = W$ and $P' = P$. As a consequence, for any given subnetwork S we have $\alpha_S = \alpha'_S$ and $\beta_S = \beta'_S$: The subnetwork is simply qualified by the pair (α_S, β_S) rather than by a 4-tuple of indicators. Thus only two types of ε structures are possible (see Fig. 2) as follows:

ε *community*:
$\phi_S^c = \|(\alpha_S,\beta_S) - (1,1)\|_\infty = \max\{1-\alpha_S, 1-\beta_S\} \leqslant \varepsilon$

ε *pseudocommunity*:
$\phi_S^p = \|(\alpha_S,\beta_S) - (0,1)\|_\infty = \max\{\alpha_S, 1-\beta_S\} \leqslant \varepsilon.$

III. FINDING (PSEUDO-)COMMUNITIES

In the previous sections, we described how to classify a subnetwork S on the basis of the associated 4-tuple $(\alpha_S,\beta_S,\alpha'_S,\beta'_S)$, and we defined eight types of nontrivial subnetworks ("structures"). Here we consider the problem of *finding* such structures in a given network.

We propose a *semilocal* searching algorithm to find the structures of a given type—the algorithm can be adapted to search for each different type of structure simply by using the relevant distance function ϕ_S among those previously defined. As we will see shortly, the algorithm is local in the sense that it starts from a given node and considers a set of larger and larger neighborhoods. In this respect, it is similar to several recently published "local" methods of community analysis (e.g., Refs. [30,31]). However, it is not fully local because the assessment of each of these subnetworks requires evaluating the persistence probabilities α_S, α'_S, which depend on the stationary probability π: computing the latter requires knowing, in general, the entire network structure (i.e., the Markov matrix P). We point out that a few methods have been proposed aimed at obtaining, on the basis of local information only, a reliable approximation of the π_is of a given subnetwork (e.g., Refs. [39,40]). These methods could be integrated in our approach to obtain a fully local algorithm. Here we do not explore this aspect, as it falls out of the scope of this paper, and we will assume to know all the π_is that are needed.

Given a node i, we denote by $C_{i,d}$ a d neighborhood of i, namely a subnetwork containing i and at least one node at the shortest-path length d from i but none at a shortest-path length larger than d (here the shortest paths are considered on the symmetrized, binary network, i.e., by disregarding direction and weight of the links). As above defined, the quantity $\phi_{C_{i,d}}$ is associated to the subnetwork: It is the distance of the related 4-tuple from the relevant vertex or edge of the unit hypercube. Let $\Gamma_{i,d} = \{C_{i,d}\}$ be the set of all d neighborhoods of i and let

$$\Phi_{i,d} = \min_{C_{i,d}\in\Gamma_{i,d}} \phi_{C_{i,d}} \tag{6}$$

be the least distance within this set: It is attained by $C^*_{i,d}$, which is therefore the best d neighborhood. The aim of our local optimization is to find

$$d^* = \min d \quad \text{s.t.} \quad \Phi_{i,d+1} > \Phi_{i,d}, \tag{7}$$

namely the smallest neighborhood of i that can only be worsened if enlarged. We denote the best d^* neighborhood by C^*_{i,d^*}.

A word of comment on problem (7). Starting from $d = 0$ (node i alone) and considering larger and larger d neighborhoods, the quantity $\Phi_{i,d}$ starts from $\Phi_{i,0} = 1$ and displays (at least initially) a decreasing trend. Despite such a trend, typically, $\Phi_{i,d}$ is not strictly monotone but may have one or more minima (this is consistent with the behavior of other indicators used in local community analysis [31,32]): problem (7) picks the first of these minima, meaning that we are mostly interested in finding structures which are *small* yet well cohesive (as measured by ϕ) and locally maximal (i.e., they are worsened by any further node inclusion).

The exact solution of problem (7) is computationally unfeasible, in general, since the number of possible d neighborhoods increases very rapidly with d. To get a reasonable suboptimal solution, we use a heuristic, greedy procedure that is similar in spirit to others proposed for local community analysis [30–32]. Starting from node i, we build a sequence of sets $\{i\} = D_{i,1} \subset D_{i,2} \subset \ldots, D_{i,m} \subset \ldots$ by adding one node at a time, m being therefore the number of nodes in the set. We denote by d_m the maximum shortest-path length from i to the nodes of $D_{i,m}$, and we compactly write $\phi_{i,m}$ for $\phi_{D_{i,m}}$. To pass from $D_{i,m}$ to $D_{i,m+1}$, we consider for possible insertion all the nodes in the "external boundary" $BD_{i,m}$ of $D_{i,m}$, i.e., all nodes $\notin D_{i,m}$ but with at least one neighbor in $D_{i,m}$, and we select among them the node which keeps $\phi_{i,m+1}$ as small as possible, namely the node j attaining the minimum [41] in

$$\phi_{i,m+1} = \min_{j\in BD_{i,m}} \phi_{D_{i,m}\cup\{j\}}. \tag{8}$$

We stop at the first minimum $m = \overline{m}$ of $\phi_{i,m}$, namely when

$$\phi_{i,\overline{m}} < \phi_{i,\overline{m}-1} \quad \text{and} \quad \phi_{i,\overline{m}} < \phi_{i,\overline{m}+1}, \tag{9}$$

and we take $d_{\overline{m}}$ and $D_i = D_{i,\overline{m}}$ as our approximations, respectively, of d^* and C^*_{i,d^*} [the optimal solution of problem (7)].

For each one of the eight types of structure, the above algorithm possibly yields a subnetwork D_i for each starting node i, which will be qualified as an ε structure provided $\phi_i =$

$\phi_{D_i} \leqslant \varepsilon$ (see Supplemental Material [42]). Not necessarily, one may want to explore the neighborhood of *all* nodes i, especially if the network is very large—one could only focus on the nodes to which she or he is most interested in, for a specific application, or on those which are most important according to some centrality measure (although, depending on the adopted measure, the most central nodes might tend to lie between communities and not at their core [43,44]). In any case, some of the structures D_i could overlap, meaning that some nodes are at the same time members of a few significant subnetworks. Most notably, some nodes could be at the same time part of structures of *different types*, e.g., an out-community and an in-pseudocommunity, sharing these memberships with different sets of partners.

A. Pruning the set of (pseudo-)communities

The output of the above algorithm is a list of subnetworks $D_1, D_2, \ldots$, each qualified by the corresponding ϕ_i value. In practice, it turns out that many D_is have a D_j which is coincident: This happens, typically, when we start from nodes i, j belonging to the same rather strong (i.e., with small ϕ_i) (pseudo-)community. In other cases, one observes two D_is that differ by just a very small fraction of their nodes: This often happens when the starting nodes are at the periphery of the same (pseudo-)community. In some other cases, two D_is are found that are distinct but have an important, nontrivial overlap.

We complement the searching algorithm with a postprocessing pruning procedure, aimed at simplifying the set of (pseudo-)communities by removing duplications but also, given a user-specified threshold, those D_j which are sufficiently similar to another subnetwork D_i. More specifically, we use the Jaccard index to quantify the similarity between (pseudo-)communities,

$$J_{ij} = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}, \tag{10}$$

and we construct a binary *similarity matrix* M by letting $M_{ij} = 1$ if and only if $J_{ij} \geqslant \nu$, where $0 < \nu \leqslant 1$. We interpret M as the adjacency matrix of an undirected, binary meta-network, of which we detect the maximal cliques (we use Bron-Kerbosch algorithm [45]), corresponding to sets of (pseudo-)communities, of the original network, which are pairwise similar. Then, for each clique, we keep the D_i with minimal ϕ_i (i.e., the most cohesive one) and prune the others. The result is, in general, a shorter list $D_1, D_2, \ldots$, on which we apply recursively the same procedure. We stop when, in a recursion step, no pruning occurs.

Note that setting $\nu = 1$ implies that only duplicate subnetworks are eliminated, whereas, in general, the smaller the ν, the smaller the number of (pseudo-)communities surviving after pruning. We will see in the next sections that, although the final number of (pseudo-)communities is often rather sensitive to $\nu < 1$, the "quality" of the community detection procedure, suitably quantified, turns out to be robust over a fairly wide interval of ν.

IV. TESTS ON BENCHMARK NETWORKS

The above-described method has been tested on a set of benchmark networks with diversified features, having a community structure that is known *a priori*. The aim is to prove that the method is able to correctly recover the existing community structure to a fairly large extent, so to trust its effectiveness in the application to real-world networks (Sec. V) where the "true" structure is obviously unknown. It must be noticed that the most authoritative benchmark networks proposed to date in the literature (we will use LFR benchmarks [46,47]), be they directed or not, are designed to contain in-/out-communities only. We are not aware of benchmark networks considering directionality in communities (although the issue is briefly touched in Ref. [47]), not mentioning the notion of pseudocommunity, so we will limit our tests to discovering in-/out-communities. The issue of generating synthetic benchmark networks containing all the types of structure appears to be far from trivial, and we leave it as a suggestion for future research.

The list $D_1, D_2, \ldots$ resulting from the above-described method might form a partition or cover [6] or, instead, include only a portion of the network. It has to be compared with the benchmark clustering, namely another list $B_1, B_2, \ldots$, which, in turn, might define a partition, a cover, or none of them. Thus the correspondence between the results and the benchmark can be assessed neither by standard indicators that compare partitions, such as the variation of information [48] or the normalized mutual information [49], nor by their generalization to covers [36]. We follow an approach [50,51] which adapts to community analysis the two well-known indicators of *recall* and *precision*, routinely used in information retrieval and classification tasks [52]. We say that the node pair (i, j) is *coupled* by the list $D_1, D_2, \ldots$ (respectively, $B_1, B_2, \ldots$) if (i, j) appears in the same subnetwork D_k (respectively, B_k) for at least one k. Then we denote by *recall* (Rec) the fraction of node pairs coupled by $B_1, B_2, \ldots$ that are also coupled by $D_1, D_2, \ldots$ and by *precision* (Prec) the fraction of node pairs coupled by $D_1, D_2, \ldots$ that are also coupled by $B_1, B_2, \ldots$. These two quantities are typically combined in a single indicator, the so-called *F measure*, ranging from 0 to 1 and taking value 1 if and only if the two lists $D_1, D_2, \ldots$ and $B_1, B_2, \ldots$ are perfectly coincident as follows:

$$F = 2\frac{\text{Rec} \cdot \text{Prec}}{\text{Rec} + \text{Prec}}. \tag{11}$$

A. LFR benchmarks: Undirected networks

LFR benchmarks [46] are a class of synthetically generated networks, purposely designed for testing community detection algorithms. They allow heterogeneity in the distributions of node degrees and community sizes, which are taken as power laws with given exponents τ_1 and τ_2, respectively. In addition, the network is defined by prescribing n, $\langle k \rangle$, and a *mixing parameter* μ such that each node shares a fraction $1 - \mu$ of its links with the other nodes of its own community and a fraction μ with the rest of the network. We first consider undirected, binary networks with $n = 1000$, $\langle k \rangle = 20$, $\tau_1 = 2$, $\tau_2 = 1$, and $\mu = 0.25$ (the latter implying well-separated communities). We produce 10 different network instances: The number of

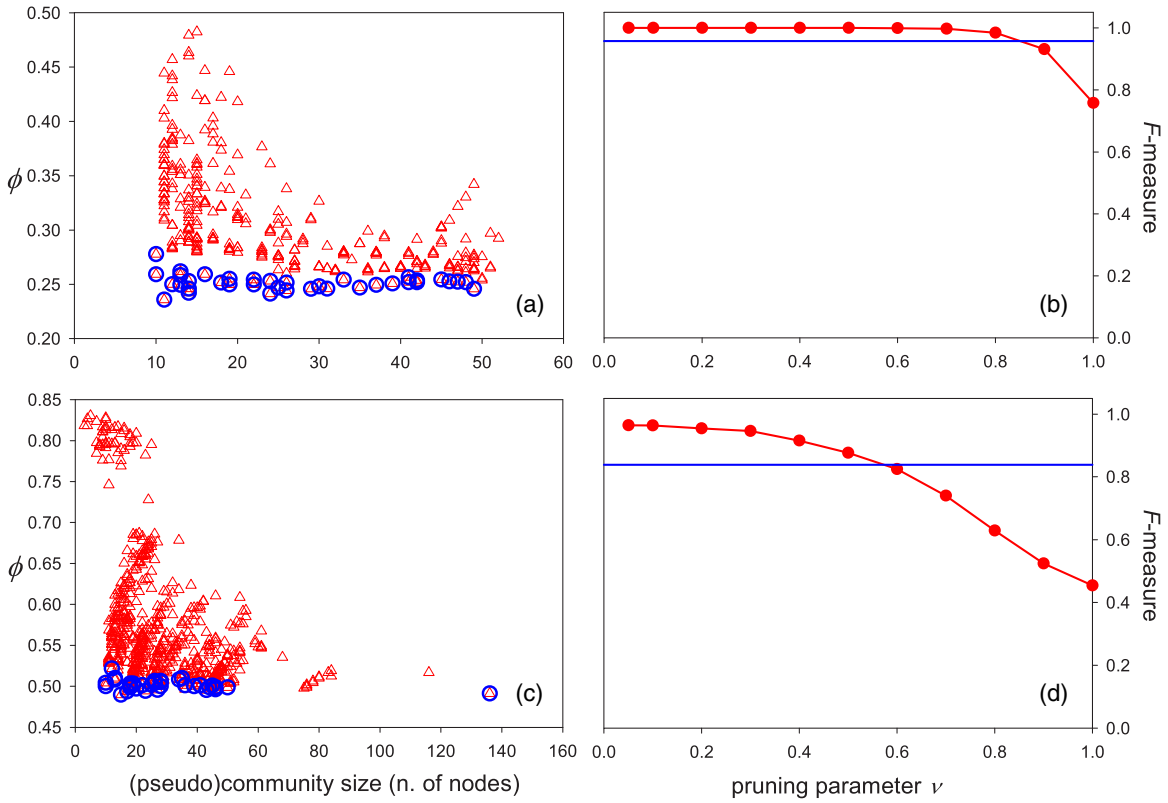

FIG. 6. (Color online) Tests on LFR undirected networks. Above, $\mu = 0.25$. (a) The in-/out-communities found by starting the algorithm from all nodes (red triangles) and those selected by the pruning procedure with $\nu = 0.1$ (blue circles) for one of the network instances. (b) The F measure obtained by comparing the planted partition with the list of in-/out-communities obtained by our method, as a function of the pruning parameter ν (red dots). The horizontal (blue) line is obtained by comparing the planted partition with the one obtained by max-modularity. Both curves are obtained by averaging over 10 network instances. Below (c) and (d): same as in (a) and (b) but with $\mu = 0.5$.

built-in communities turns out to range from 35 to 43, and the size of each community from 10 to 77 nodes.

Figure 6(a) displays, for one of the network instances, the size versus ϕ distribution of the 348 distinct subnetworks found by the algorithm starting from all 1000 nodes. Figure 6(b) shows that the outcome of the pruning procedure is largely insensitive to ν and that the results favorably compare to the partition obtained by max-modularity (we used Louvain algorithm [53]). Specifically, in Fig. 6(a) we also highlight the in-/out-communities selected by pruning with $\nu = 0.1$: The 37 communities of the planted partition are perfectly recovered ($F = 1$). As expected, the ϕ values concentrate around the value of $\mu = 0.25$ (notice that μ prescribes the internal versus external link balance for each single node, while α_S and β_S are related to the same balance but for the whole subnetwork S), while the sizes are spread, by construction, over a fairly large interval.

In Ref. [46] it is discussed how the performance of community detection algorithms deteriorates when μ increases (i.e., communities become less isolated). To analyze this situation, we generate another set of 10 benchmark networks by increasing μ to 0.5: The resulting networks turn out to have from 34 to 44 communities, with sizes ranging from 10 to 73 nodes. Notice that we are generating low-quality clusters, due to the large μ: Actually, they are even on the edge of meeting the requirement of "community in a weak sense" according to Ref. [35]. In other words, the cluster structure of the network is feeble, which is the reason for the rather low performance of community detection tools documented in Ref. [46]. It turns out that, in this difficult situation, the performance of our method is more sensitive to the pruning parameter ν than above [see Fig. 6(d)]. Nonetheless, the results are definitely better than those obtained by max-modularity if sufficiently small ν values are adopted. Furthermore, differently than applying max-modularity, each (pseudo-)community is equipped with its ϕ value that quantifies its (low) quality.

B. LFR benchmarks: Directed networks

We move now to directed networks, which are more properly the topic of this paper. LFR benchmarks of this type can be obtained by an extension of the basic procedure, as described in Ref. [47]. We set the network parameters to (we refer to Ref. [47] for their detailed definition) $n = 1000$, $\langle k \rangle = 25$, $\tau_1 = 2$, and $\tau_2 = 1$. We generated two different sets (each composed of 10 instances) of networks, differentiated by the value of the mixing parameter, which is set to $\mu = 0.3$ and $\mu = 0.6$, respectively. The results of the analysis, summarized in Fig. 7, show that in both cases our procedure is able to recover the planted partitions better than max-modularity and, notably, that the results are practically insensitive to the pruning parameter ν.

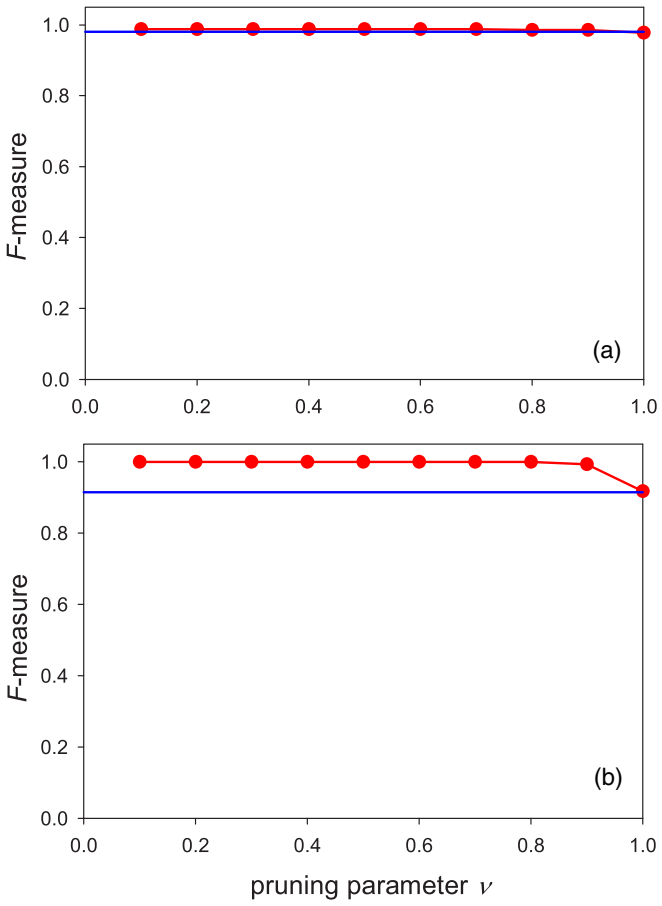

FIG. 7. (Color online) Tests on LFR directed networks. The figure reports the F measure obtained by comparing the planted partition with the list of in-/out-communities obtained by our method as a function of the pruning parameter ν (red dots). The horizontal (blue) line is obtained by comparing the planted partition with the one obtained by max-modularity. Both curves are obtained by averaging over 10 network instances. (a) $\mu = 0.3$; (b) $\mu = 0.6$.

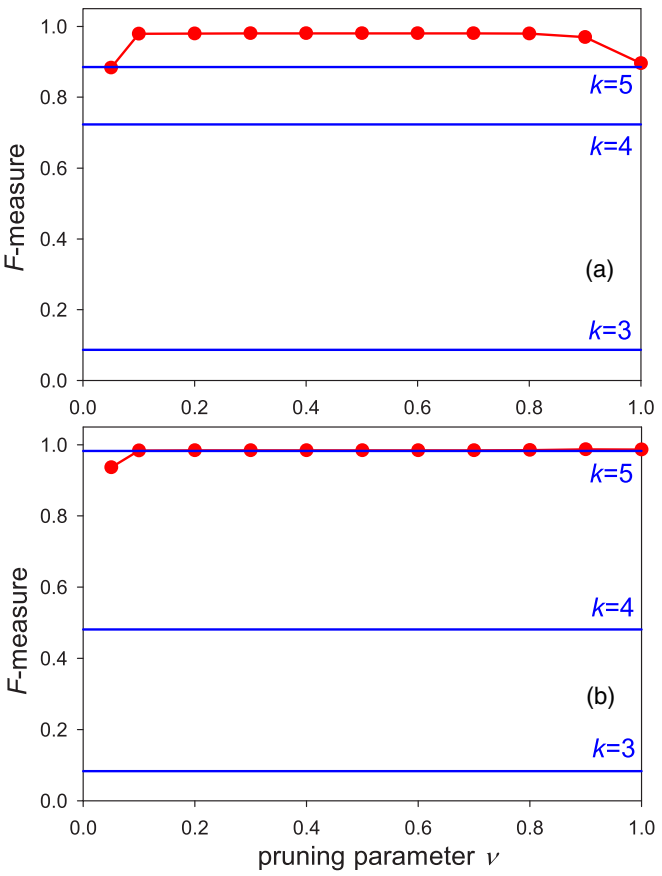

FIG. 8. (Color online) Tests on LFR networks with overlapping communities. The figure reports the F measure obtained by comparing the planted partition with the list of in-/out-communities obtained by our method as a function of the pruning parameter ν (red dots). The horizontal (blue) lines are obtained by comparing the planted partition with those obtained by the clique percolation method [54] for three values of the clique size k (a further increase of k deteriorates the performance). All curves are obtained by averaging over 10 network instances, with 50% of the nodes belonging to two communities. (a) Undirected networks; (b) directed networks.

C. LFR benchmarks: Directed networks with overlaps

A further extension of the LFR benchmarks allows us to create networks with overlapping communities (i.e., some of the nodes belong to two or more communities), a situation where community detection is obviously more difficult [47]. We tested our algorithms both on undirected and directed networks (10 instances for each class) with $n = 1000$, $\langle k \rangle = 20$, $\tau_1 = 2$, $\tau_2 = 1$, and $\mu = 0.1$, with planted partitions (or covers, more precisely) having 50% of the nodes belonging simultaneously to two communities. We compared our results with those obtained with the clique percolation method [54], one of the most popular algorithms to analyze community structure with overlaps. We obtained a very strong agreement between the planted partition and the results of our method (Fig. 8), with a performance superior to clique percolation for undirected networks and essentially equivalent in the directed case. Once again, we point out the strong insensitivity to the pruning parameter ν.

D. Benchmark protein-protein interaction network

We now analyze a real-word network, more precisely, a protein-protein interaction network, which has two distinctive features. First, a biologically based (i.e., not derived from network modeling) benchmark list of trusted communities (protein complexes) is available. Second, this list is neither a partition nor a cover, namely part of the network is not included in any of the benchmark communities. The network implements the interaction database prepared as described in Ref. [55] and postprocessed as in Ref. [56] to preserve only significant interactions. It has 990 nodes and 4687 interactions (undirected weighted links). The benchmark, prepared by MIPS [57], contains 59 communities (if restricted to the postprocessed network), ranging from 6 to 34 nodes (all clusters with smaller size are purposely not considered, and we will do the same in our analysis). Benchmark communities do not overlap, and their union include 628 nodes over 990.

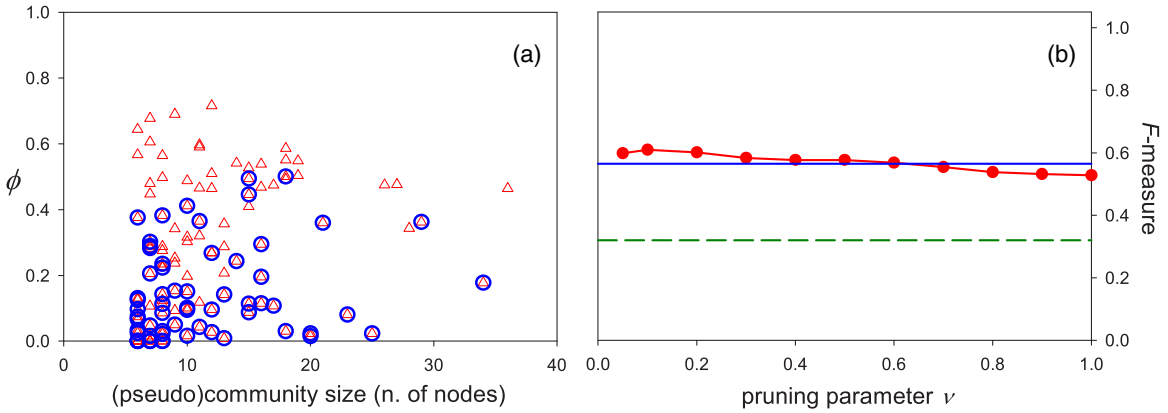

FIG. 9. (Color online) Tests on a protein-protein interaction network. (a) The in-/out-communities found by starting the algorithm from all nodes (red triangles) and those selected by the pruning procedure with $\nu = 0.1$ (blue circles). (b) The F measure obtained by comparing the benchmark communities with the list of in-/out-communities obtained by our method as a function of the pruning parameter ν (red dots). The horizontal solid (blue) line is obtained by comparing the benchmark communities with those obtained by the clique percolation method [54] for clique size $k = 6$ (it is the best performance obtained for k from 3 to 7). The horizontal dashed (green) line is the performance obtained by max-modularity.

As expected, the performance of max-modularity is very poor (Fig. 9), as it is a method that artificially forces a network partition (i.e., all nodes must belong to a community). Local methods, on the contrary, can potentially obtain much better results: Figure 9(b) shows that our method obtains a performance comparable with that of clique percolation. But, in addition, it quantifies the quality of each in-/out-community by means of the ϕ value [Fig. 9(a)]. We note, incidentally, that a few of the benchmark communities are actually small subnetworks disconnected from the giant component (they have $\phi_S = 0$, since $\alpha_S = \beta_S = \alpha'_S = \beta'_S = 1$ for an isolated subnetwork) and that they are perfectly recovered by our method, which is able to manage a network with such features thanks to the teleportation scheme (Sec. II B).

E. Erdős-Rényi networks

In random Erdős-Rényi networks, links should be homogeneously distributed, by definition, and, consequently, there should not be any (pseudo-)communities. It is well known, however, that this is true on average, whereas a specific network instance may display clusters produced by the randomicity in the distribution of links (e.g., Ref. [6]). These clusters, however, are expected to vanish as density increases, i.e., as the network tends, in the limit, to a complete one. In this situation, a community detection method should be able to reveal the low quality of clusters, when existing, and to detect a rapidly vanishing number of them when density increases.

Our method possesses these agreeable properties. We analyzed directed Erdős-Rényi networks with different sizes and densities. The typical outcome is that displayed in

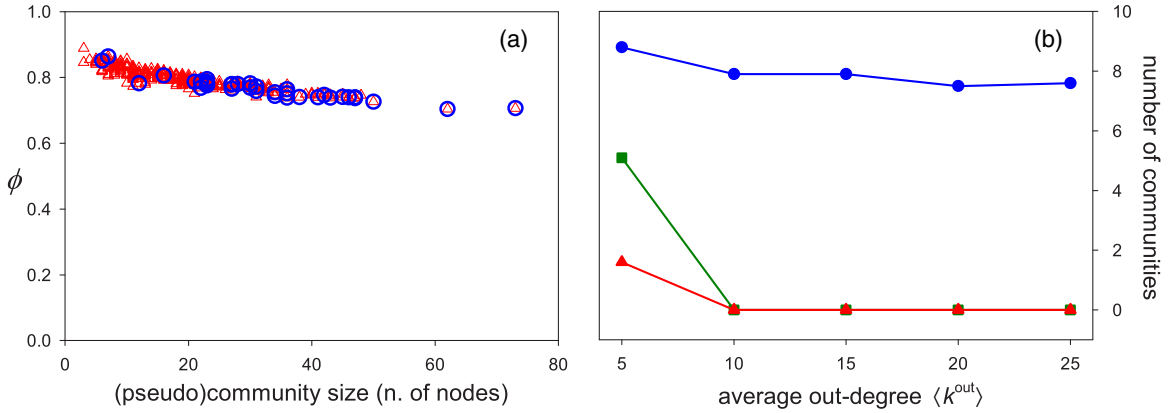

FIG. 10. (Color online) Tests on directed Erdős-Rényi networks. (a) The in-/out-communities found by starting the algorithm from all nodes (red triangles) and those selected by the pruning procedure with $\nu = 0.1$ (blue circles) for a network instance with $n = 1000$, $\langle k^{out} \rangle = 10$. (b) The number of ε in-/out-communities, with $\varepsilon = 0.5$, detected by our method before (green squares) and after pruning ($\nu = 0.1$) (red triangles), and the number of communities detected by max-modularity (blue circles). Each point is an average over 10 network instances with $n = 200$.

Fig. 10(a), where we show the size versus ϕ distribution of the in-/out-communities that are identified on a 1000-node network with rather small density ($\langle k^{\text{out}} \rangle = 10$). After pruning, 36 in-/out-communities are detected. Their ϕ value, however, is always larger than 0.7, revealing a very low cohesiveness (see again Sec. II A). Following the approach of Ref. [49], we systematically analyzed a set of directed Erdős-Rényi networks with increasing density, counting the number of ε in-/out-communities, with $\varepsilon = 0.5$, detected by our method. As summarized in Fig. 10(b), such a number falls rapidly to zero, as it should be for a consistent method [49]. Notice that, on the contrary, max-modularity continues to detect a non-negligible number of clusters, without providing any tool to assess their individual quality.

V. RESULTS ON REAL-WORLD NETWORKS

The above described method (searching algorithm plus pruning procedure) has been applied to a number of real-world networks, which are described in detail (with notes on the data sources) in the Supplemental Material [42]. For each network, the algorithm has been repeatedly used, starting from all nodes $i = 1,2,\ldots,n$, to discover the eight different types of structures. We point out that, in the technical implementation [58], two additional computational parameters are available to the analyst, namely the maximum allowable (pseudo-)community size $n_{\max}$ ($\leqslant n$) and the stopping parameter r ($\geqslant 0$). They are described in detail in the Supplemental Material [42], where we also discuss how they can help the analyst in tuning the algorithm selectivity or in speeding up computations. Here we only point out that, both for the tests on benchmark networks (Sec. IV) and for those on real-world networks (this section), the two above parameters were inactive, i.e., they were set at their default value ($n_{\max} = n, r = 0$).

The results of the analysis are summarized in Tables I and II, where two different values of the quality threshold are used, namely $\varepsilon = 0.5$ and 0.25. We recall that selecting significant structures with $\varepsilon = 0.5$ can be considered as a generalization of the notion of "community in a weak sense" put forward by Radicchi *et al.* [35] and should be regarded as a baseline for significance; $\varepsilon = 0.25$ represents a much stricter quality requirement. Notice that, since ε acts *a posteriori*, the (pseudo-)communities in Table II are trivially a subset of those in Table I: Our aim is simply to show how the number of structures decreases when a more restrictive requirement on cohesiveness is set up.

First, we note that all eight types of structures are actually found in the analyzed pool of networks (although in-pseudocommunities/out-communities and in-communities/out-pseudocommunities, which are comparatively rarer, are not found—with one exception only—in the most restrictive parameter setting). Second, the number of structures that are identified dramatically decreases as the quality threshold becomes more stringent, meaning that the (pseudo-)communities span a broad range of ϕ values (we will discuss some examples in the following): The network analyst should therefore take special care in selecting the value of ε, as it reflects her or his subjective requirements on the subnetwork cohesiveness. In our code implementation, she or he is graphically supported by size versus ϕ plots, which aids in interactively selecting a proper quality threshold. Third, in most networks, in-, out-, and in-/out-communities coexist, proving that their distinction is indeed crucial, although overlooked, in the literature to date. The same holds for the various types of pseudocommunities which, although never studied before, appear to be ubiquitous in directed networks, i.e., they are found in all analyzed cases. Finally, pruning strongly simplifies the set of (pseudo-)communities, as already put in evidence.

In the following, we discuss a few of the structures that have been found in order to highlight the effectiveness of the methodology. We mostly restrict ourselves to the cases of Table II, namely the most significant, and we consider only those (pseudo-)communities that remained after pruning.

The world trade network, which models the international transactions among countries, has been deeply studied in recent years with the tools of network analysis, disclosing a number of interesting features (e.g., Refs. [59–62]). If we consider the "total" network (Wtn), where the flows of all economic sectors are summed up, our method discovers two *in-/out-communities*. They are large clusters (98 and 78 countries, respectively, with $\phi^{\text{ioc}} = 0.21$ and 0.24), which roughly correspond to Europe plus part of Africa, on one side, and America plus Asia plus the other part of Africa, on the other side. The two blocs have a moderate overlap (10 countries of minor economic importance) and, together, cover almost the entire network (more details on this case study, with a remark about the robustness of results, are in the Supplemental Material [42]).

The picture is less trivial if we restrict our analysis to the flows of specific commodities (as classified by the United Nations SITC standard). We consider, for example, the trade flows of "leather, leather manufactures, and dressed furskins" (Wtn61) and "office machines and automatic data-processing machines" (Wtn75). The corresponding networks are much less dense than the total one, and flows are often organized around leading countries. A few examples are in Fig. 11 [63]. In Wtn61 we find, among others, an important *out-pseudocommunity* ($\phi^{\text{op}} = 0.16$, starting node Nigeria) centered on Italy, which imports raw materials from a number of countries and exports (semi-)finished goods to the rest of the network [Fig. 11(a)]. Furthermore, Italy is the only node in common with another structure (although less significant, $\phi^{\text{ipoc}} = 0.31$, and thus not included in Table II, starting node Serbia) which overlaps with the former. It is an *in-pseudocommunity*/*out-community* [Fig. 11(b)] which includes many important European countries (e.g., France and Germany) where products of this sector circulate with low boundary outflow (out-community) and where only a few countries, in practice only Italy, import significantly from the outside (in-pseudocommunity).

Many examples of significant (pseudo-)communities are found in the Wtn75 network, too. The main actors of the 23-node *in-community* ($\phi^{\text{ic}} = 0.21$, starting node Philippines) of Fig. 11(c) are China and Japan, but a few other Asian highly developed and/or developing economies, e.g., South Korea, Malaysia, and Singapore, are included. Being an in-community, this subnetwork (which also contains major countries such as Russia and Australia) proves to be self-sufficient in this commodity sector but able to significantly export worldwide. An example of *in-pseudocommunity* is in

TABLE I. A summary of the results of (pseudo-)community analysis on the pool of real-world networks, with quality threshold $\varepsilon = 0.5$. For each network, n is the number of nodes after isolated nodes have been removed. For each type of structure (OC to ICOP), the table reports the number of ε structures found (distinct but possibly overlapping) and [in brackets] their number after pruning with $\nu = 0.5$. See Supplemental Material [42] for details on network data.

Network	n	OC	IC	IOC	OP	IP	IOP	IPOC	ICOP
Wtn	183	26[4]	40[7]	10[2]	105[32]	109[36]	115[26]	27[10]	30[11]
Wtn61	218	22[8]	59[11]	6[1]	146[27]	134[38]	164[25]	25[13]	70[11]
Wtn75	226	37[4]	12[4]	2[1]	123[22]	165[36]	161[45]	36[4]	12[7]
Ownership	141	18[11]	25[6]	9[5]	30[21]	86[39]	24[23]	20[10]	8[6]
Airports	2939	443[226]	379[171]	583[268]	1103[607]	1320[648]	1714[663]	585[230]	647[286]
Leadership	32	9[5]	12[8]	12[5]	22[14]	12[11]	11[7]	6[4]	14[11]
Prison	67	20[17]	19[18]	17[14]	20[14]	33[24]	21[15]	17[15]	10[7]
Little Rock	183	9[3]	73[2]	5[2]	118[19]	137[11]	41[9]	8[2]	30[6]
St. Marks	49	21[5]	21[4]	2[1]	26[14]	34[9]	16[8]	13[5]	18[4]
St. Martin	45	21[2]	21[2]	1[1]	28[9]	34[13]	17[7]	21[3]	20[11]
Ythan	135	54[2]	49[2]	0[0]	83[14]	102[30]	49[16]	65[2]	53[13]
Neural	297	12[4]	41[21]	7[4]	205[38]	129[57]	60[34]	9[3]	54[24]
E. Coli	418	1[1]	11[8]	3[2]	33[25]	270[69]	87[28]	1[1]	23[7]
S. Cerevisiae	688	81[19]	35[12]	18[8]	125[43]	461[77]	69[17]	37[10]	12[5]
Political blogs	1224	179[5]	81[29]	59[4]	816[105]	764[178]	248[47]	145[17]	20[8]
Japanese	2704	28[11]	114[21]	62[7]	2150[276]	2221[233]	1190[237]	159[16]	190[16]
Netscience	1461	n.a.	n.a.	214[190]	n.a.	n.a.	503[447]	n.a.	n.a.
Ppi	990	n.a.	n.a.	128[94]	n.a.	n.a.	333[213]	n.a.	n.a.

Fig. 11(d) ($\phi^{\text{ip}} = 0.16$, starting node Grenada), in which many of the countries of Central America and the Caribbean get most of their import flow from the US, which, on the other hand, imports mostly from the rest of the world.

In the Ownership network, nodes represent the companies listed in the Italian stock exchange, and the weight w_{ij} is the percentage of the shares of company j owned by i. Typically, ownership networks display a bow-tie structure [64] and, as a consequence, they hardly admit partitions formed by significant communities, as this would require the existence of many subnetworks with strong internal cohesiveness (e.g., large persistence probability). For that, community analysis has mostly been carried out on the undirected (i.e., symmetrized) network, highlighting the relationships between companies regardless of their direction [9]. Yet "who owns whom" is obviously crucial information in many respects. On the directed network, our searching algorithm discovers a rather cohesive *in-community* ($\phi^{\text{ic}} = 0.28$, starting node Data Service) with 22 companies,

TABLE II. A summary of the results of (pseudo-)community analysis on the pool of real-world networks, with quality threshold $\varepsilon = 0.25$. For each network, n is the number of nodes after isolated nodes have been removed. For each type of structure (OC to ICOP), the table reports the number of ε structures found (distinct but possibly overlapping) and [in brackets] their number after pruning with $\nu = 0.5$. See Supplemental Material [42] for details on network data.

Network	n	OC	IC	IOC	OP	IP	IOP	IPOC	ICOP
Wtn	183	0[0]	0[0]	10[2]	0[0]	0[0]	0[0]	0[0]	0[0]
Wtn61	218	0[0]	0[0]	6[1]	92[6]	24[4]	0[0]	0[0]	0[0]
Wtn75	226	0[0]	3[2]	0[0]	10[4]	60[2]	0[0]	5[1]	0[0]
Ownership	141	0[0]	1[1]	0[0]	0[0]	39[8]	0[0]	0[0]	0[0]
Airports	2939	0[0]	0[0]	108[58]	3[3]	10[4]	568[134]	0[0]	0[0]
Leadership	32	0[0]	0[0]	2[2]	0[0]	0[0]	0[0]	0[0]	0[0]
Prison	67	0[0]	0[0]	2[2]	0[0]	0[0]	0[0]	0[0]	0[0]
Little Rock	183	3[2]	27[1]	1[1]	9[1]	89[5]	7[1]	0[0]	0[0]
St. Marks	49	0[0]	4[1]	2[1]	0[0]	2[2]	0[0]	0[0]	0[0]
St. Martin	45	0[0]	2[1]	1[1]	0[0]	2[2]	0[0]	0[0]	0[0]
Ythan	135	0[0]	30[1]	0[0]	17[2]	48[10]	4[1]	0[0]	0[0]
Neural	297	0[0]	3[3]	2[1]	145[23]	2[1]	0[0]	0[0]	0[0]
E. Coli	418	0[0]	0[0]	0[0]	0[0]	165[21]	0[0]	0[0]	0[0]
S. Cerevisiae	688	0[0]	0[0]	0[0]	0[0]	282[28]	0[0]	0[0]	0[0]
Political blogs	1224	3[2]	7[5]	56[2]	152[9]	257[26]	0[0]	0[0]	0[0]
Japanese	2704	0[0]	0[0]	6[1]	1793[108]	1886[90]	419[30]	0[0]	0[0]
Netscience	1461	n.a.	n.a.	151[133]	n.a.	n.a.	5[5]	n.a.	n.a.
Ppi	990	n.a.	n.a.	83[73]	n.a.	n.a.	1[1]	n.a.	n.a.

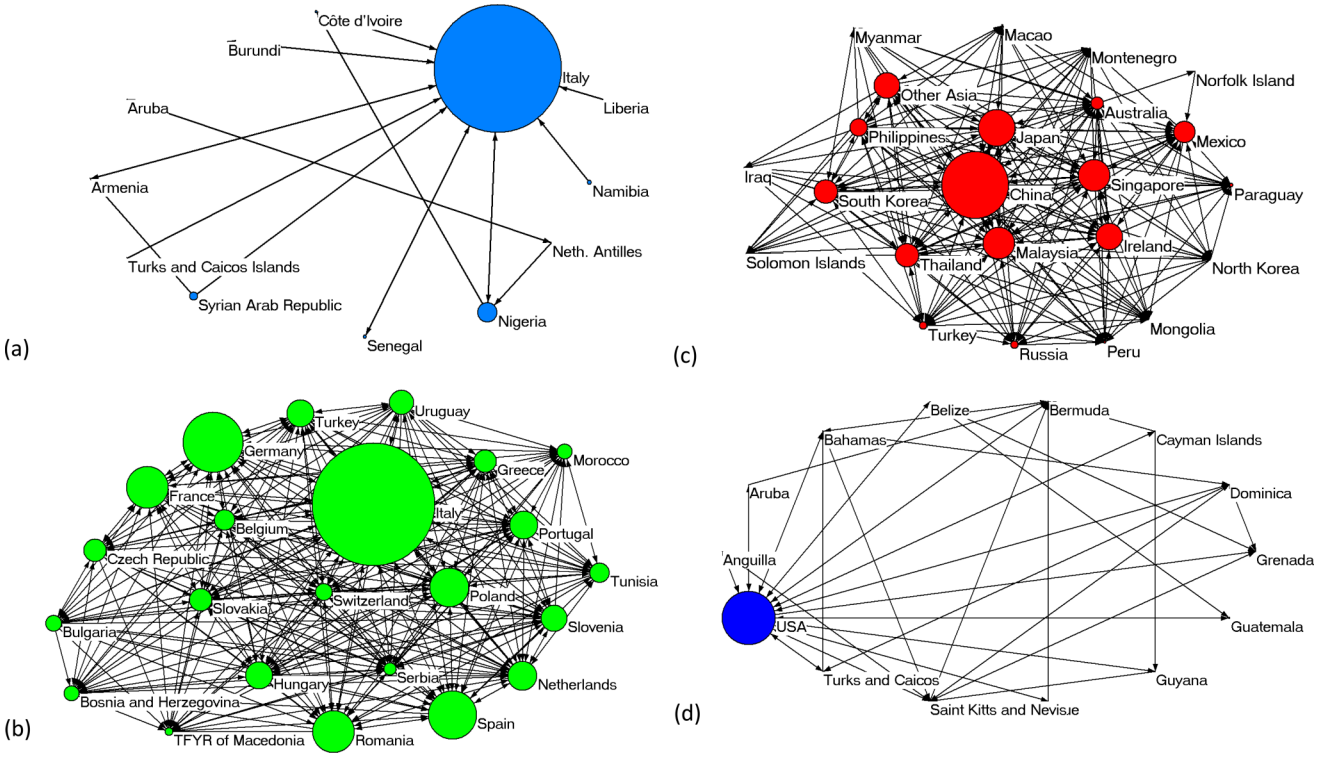

FIG. 11. (Color online) Examples of (pseudo-)communities in the world trade networks. (a) Wtn61: A out-pseudocommunity organized around Italy, which imports raw materials and exports (semi-)finished goods to the rest of the network. (b) Wtn61: An in-pseudocommunity/out-community which overlaps with the former through Italy. (c) Wtn75: An in-community including many Far East highly developing economies. (d) Wtn75: An in-pseudocommunity of Central American and the Caribbean countries importing from the US.

including a few of the most important Italian financial groups [Fig. 12(a)]. It is a coalition of companies which, by means of cross-shareholdings and other forms of alliances (e.g., board interlocks), strategically controls most of the Italian economic system. The characterization of this subnetwork as an in-community reveals that shares of these companies owned by outside firms are negligible. On the contrary, the companies of this set surely own significant shares at the outside: otherwise, this subnetwork would be qualified as out-community, too, whereas α_S turns out to be as small as 0.25. On the same network, a few *in-pseudocommunities* are disclosed, too. A typical example ($\phi^{\text{ip}} = 0.17$, starting node Banco Desio Brianza) is in Fig. 12(b): It contains two leaders, namely IntesaSanPaolo and Assicurazioni Generali (two of the major Italian financial institutions), and a few minor companies for which the leaders are practically the only shareholders among the companies listed at the stock exchange. The whole picture of the in-communities and in-pseudocommunities found in this network is displayed in Fig. 13. The identified structures span a rather broad range both in size and cohesiveness ϕ, with a nontrivial relationship among the two variables. Notice that a few of the (pseudo-)communities are very small (pairs or triads of nodes) and could safely be removed from the total count.

A few types of significant structures are found in the worldwide Airports network, too. As one can expect, the most cohesive communities correspond to regional (i.e., local) transportation systems, namely peripheral subnetworks well connected internally but with just a few routes to and from the

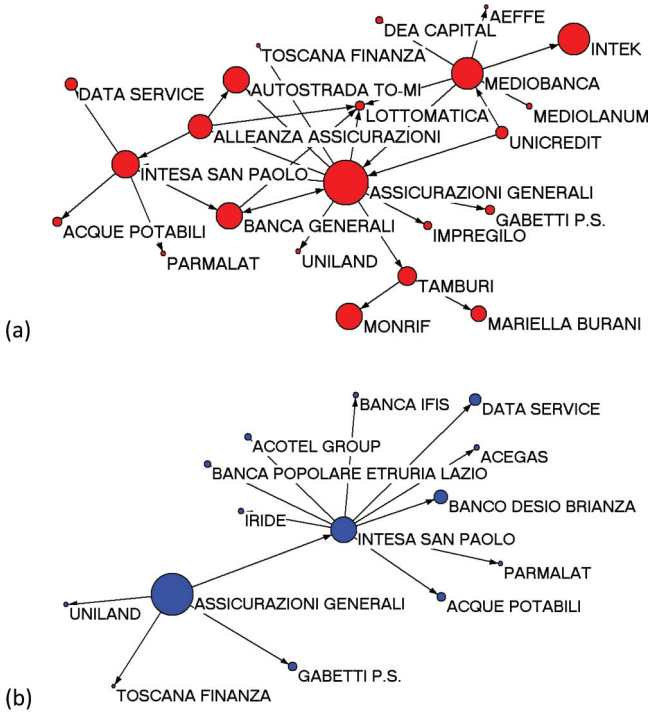

FIG. 12. (Color online) Examples of (pseudo-)communities in the Ownership network. (a) An in-community including the most important Italian financial institutions. (b) An in-pseudocommunity, where the "leader" companies are the only shareholders of the "follower" companies.

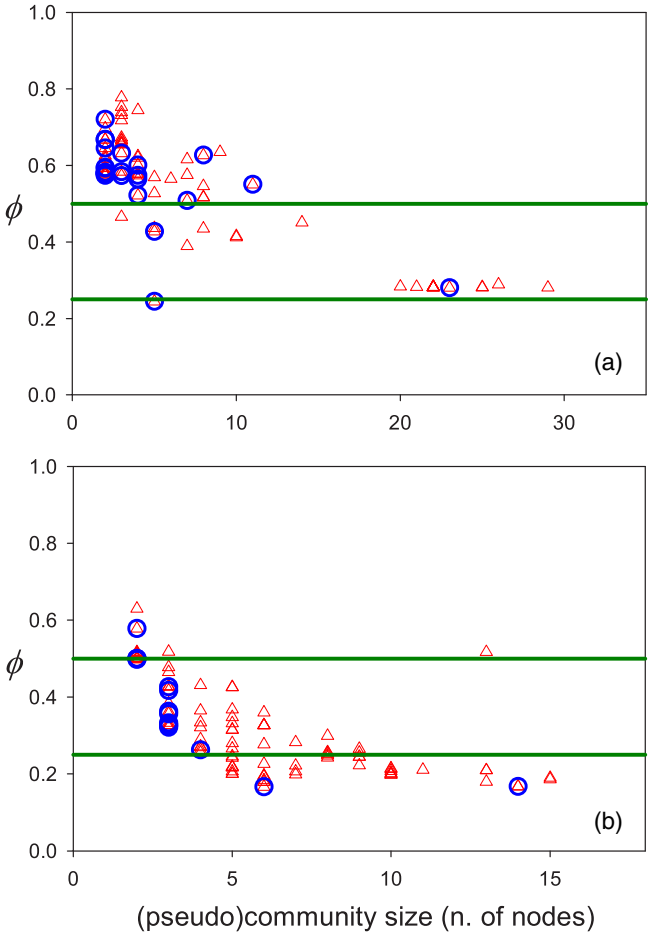

FIG. 13. (Color online) The in-communities (a) and in-pseudocommunities (b) found in the Ownership network by starting the algorithm from all nodes (red triangles) and those selected by the pruning procedure with $\nu = 0.1$ (blue circles). The horizontal (green) lines mark the values $\varepsilon = 0.5$ and 0.25 of the quality threshold used in Tables I and II, respectively.

outside. Figure 14(a) shows one of the *in-/out-communities* with the smallest ϕ^{ioc} value (0.085, starting node Rampart), which refers to Alaska. *In-/out-pseudocommunities* are rather common, too, and they are typically organized as a starlike subnetwork, with a number of minor (peripheral) airports connected to an international hub. A clarifying examples is in Fig. 14(b) ($\phi^{\mathrm{iop}} = 0.087$, starting node Ataturk Istanbul): Twelve Turkish airports are uniquely connected to Istanbul airport, which, on the contrary, has as many as 157 connections with the remaining nodes of the worldwide network.

We close this section with an undirected, weighted network, namely the so-called Netscience, which describes the collaborations (up to 2006) between scholars in network science [65]. Due to its well-pronounced modular structure, this graph has become a standard for community analysis methods. Indeed, our searching algorithm reveals the existence of a number of significant *communities*, with sizes ranging (in the most restrictive parameter setting, see Table II) from 3 to 30 nodes. They correspond to research groups or established cooperations, with a few overlaps between them: An example is displayed in Fig. 15(a), where a 22-node and a 6-node

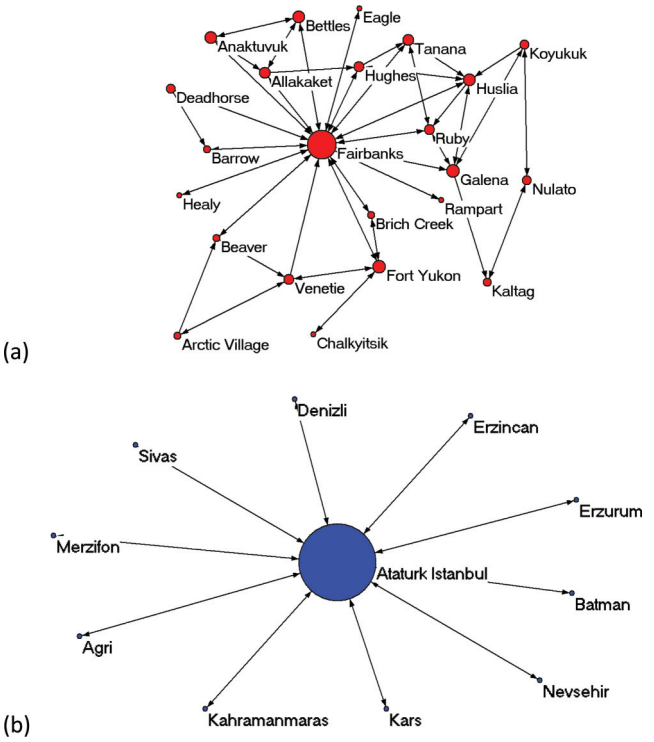

FIG. 14. (Color online) Examples of (pseudo-)communities in the Airports network. (a) An in-/out-community corresponding to a regional transportation system. (b) An in-/out-pseudocommunity, with a number of minor (peripheral) airports connected to an international hub.

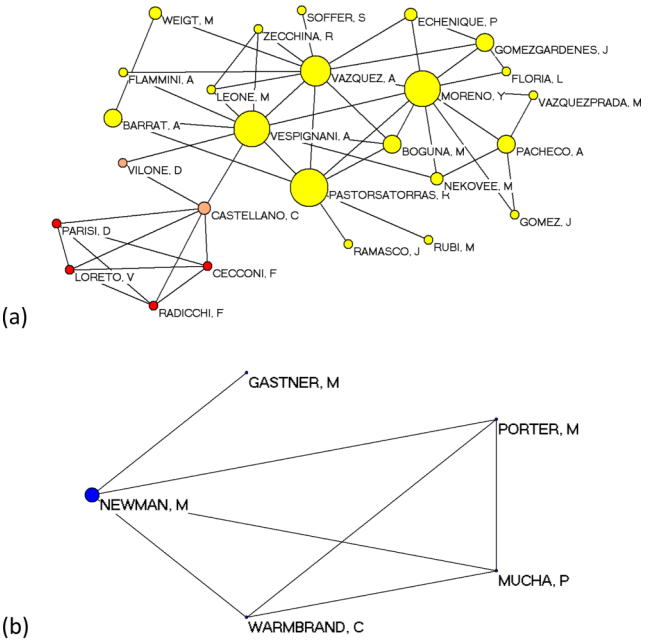

FIG. 15. (Color online) Examples of (pseudo-)communities in the Netscience network. (a) Two overlapping communities (shared nodes Castellano and Vilone are in orange). (b) A pseudocommunity, where a "leader" with many collaborations has only one or few coauthorships with a small group of scholars.

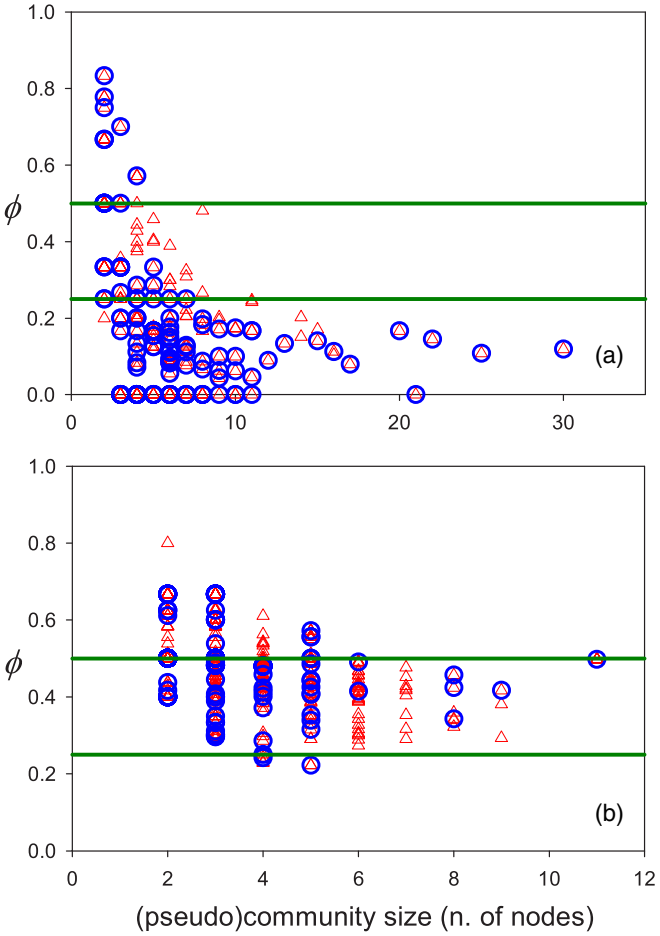

FIG. 16. (Color online) The communities (a) and pseudocommunities (b) found in the Netscience network by starting the algorithm from all nodes (red triangles) and those selected by the pruning procedure with $\nu = 0.1$ (blue circles). The horizontal (green) lines mark the values $\varepsilon = 0.5$ and 0.25 of the quality threshold used in Tables I and II, respectively.

subnetwork (both with $\phi^c = 0.14$, starting nodes Vespignani A. and Parisi D.) have 2 nodes in common. As far as *pseudocommunities* are concerned, in this type of network they typically identify single scholars, or small groups, with just one coauthorship with a leading author, who has instead a large number of collaborations with the rest of the network [see Fig. 15(b) for an example, $\phi^p = 0.22$, starting node Porter M.]. The complete size versus ϕ picture of the identified (pseudo-)communities is in Fig. 16. We report that if we restrict the analysis to the 379-node connected component (as is usually done in the literature), the F measure obtained by comparing our in-/out-communities with those found by max-modularity is as large as $F = 0.95$. Indeed, all node pairs coupled by our method are also coupled by max-modularity. The converse is not true, however, since max-modularity is constrained to get a partition, so a few nodes are unnaturally forced to belong to a community. In our setting, instead, about 7% of nodes are not included in any in-/out-community.

VI. CONCLUSIONS

Directed networks, namely complex interconnected systems with *asymmetric* interactions, are ubiquitous in a number of fields in science and technology. We find them, just to mention a few examples, in economics and finance (trade relationships at the country level or share ownerships among companies), in ecology (prey-predator interactions in a food web), in biology (transcriptional regulatory networks or neural networks), and in information science (WWW or citation networks). As the link directionality strongly affects the structural properties and functioning of the system, it becomes crucial to fully consider its role when communities are sought for. To this aim, we revisited the notion of community—a subnetwork mildly connected to the rest of the system—by distinguishing whether the isolation is related to the in- and/or out-flow. Besides, we introduced the new notion of pseudocommunity to capture notable structures which are often found in applications. The detailed analysis of many real-world networks demonstrated that distinguishing among in-, out-, and in-/out-(pseudo-)communities is crucial to fully interpret the role and function of important subnetworks.

ACKNOWLEDGMENT

The authors are grateful to Isabella Cingolani, Andrea Lancichinetti, and Lucia Tajoli for many helpful suggestions.

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. H. Hwang, Phys. Rep. **424**, 175 (2006).
[2] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, 2008).
[3] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Cambridge, 2010).
[4] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, Cambridge, 2010).
[5] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi, Nature **473**, 167 (2011).
[6] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[7] P. Jonsson, T. Cavanna, D. Zicha, and P. Bates, BMC Bioinform. **7** (2006), doi:10.1186/1471-2105-7-2.
[8] A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor, Nature **426**, 282 (2003).
[9] C. Piccardi, L. Calatroni, and F. Bertoni, Physica A **389**, 5247 (2010).
[10] G. Flake, S. Lawrence, C. Giles, and F. Coetzee, Computer **35**, 66 (2002).
[11] M. A. Fortuna, J. A. Bonachela, and S. A. Levin, Proc. Natl. Acad. Sci. USA **108**, 19985 (2011).
[12] M. Girvan and M. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[13] A. Arenas, L. Danon, A. Diaz-Guilera, P. Gleiser, and R. Guimera, Eur. Phys. J. B **38**, 373 (2004).
[14] R. Guimera, M. Sales-Pardo, and Luis A. N. Amaral, Phys. Rev. E **76**, 036102 (2007).

[15] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, and T. Vicsek, New J. Phys. **9**, 186 (2007).
[16] E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
[17] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, J. Stat. Mech. Theor. Exp. (2009) P03024.
[18] Y. Kim, S.-W. Son, and H. Jeong, Phys. Rev. E **81**, 016103 (2010).
[19] J. C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. USA **107**, 12755 (2010).
[20] A. Arenas, J. Duch, A. Fernandez, and S. Gomez, New J. Phys. **9**, 176 (2007).
[21] O. Popa, E. Hazkani-Covo, G. Landan, W. Martin, and T. Dagan, Genome Res. **21**, 599 (2011).
[22] W. Liao, J. Ding, D. Marinazzo, Q. Xu, Z. Wang, C. Yuan, Z. Zhang, G. Lu, and H. Chen, Neuroimage **54**, 2683 (2011).
[23] A. Yazdani and P. Jeffrey, Water Resour. Res. **48**, W06517 (2012).
[24] F. Bono, E. Gutierrez, and K. Poljansek, Physica A **389**, 5287 (2010).
[25] D. Li, Q. Liu, X. Wang, and Z. Lin, Automatica **49**, 610 (2013).
[26] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, Proc. Natl. Acad. Sci. USA **107**, 10815 (2010).
[27] S.-W. Son, C. Christensen, G. Bizhani, D. V. Foster, P. Grassberger, and M. Paczuski, Phys. Rev. E **86**, 046104 (2012).
[28] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008).
[29] C. Piccardi, PLoS ONE **6**, e27028 (2011).
[30] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).
[31] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[32] J. P. Bagrow, J. Stat. Mech. Theor. Exp. (2008) P05001.
[33] C. Meyer, *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia, PA, 2000).
[34] F. Della Rossa, F. Dercole, and C. Piccardi, Sci. Rep. **3**, 1467 (2013).
[35] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658 (2004).
[36] A. Lancichinetti, S. Fortunato, and J. Kertesz, New J. Phys. **11**, 033015 (2009).
[37] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton University Press, Princeton, NJ, 2006).
[38] R. Lambiotte and M. Rosvall, Phys. Rev. E **85**, 056107 (2012).
[39] Y.-Y. Chen, Q. Gan, and T. Suel, in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04 (ACM, New York, 2004), pp. 381–389.
[40] Z. Bar-Yossef and L.-T. Mashiach, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08 (ACM, New York, 2008), pp. 279–288.
[41] If many nodes j attain the minimum in Eq. (8) we select one of them at random.
[42] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.89.012814 for a detailed description of network data (with their sources) and a discussion on the computational complexing of the algorithm.
[43] R. Guimera and L. Amaral, Nature **433**, 895 (2005).
[44] R. Guimera and L. Amaral, J. Stat. Mech. Theor. Exp. (2005) P02001.
[45] C. Bron and J. Kerbosch, Commun. ACM **16**, 575 (1973).
[46] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[47] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 016118 (2009).
[48] M. Meilă, J. Multivar. Anal. **98**, 873 (2007).
[49] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).
[50] S. Gregory, in *Knowledge Discovery in Databases: PKDD 2007, Proceedings*, Lecture Notes in Artificial Intelligence, Vol. 4702, edited by J. N. Kok J. Koronacki, R. L. DeMantaras, S. Matwin, D. Mladenic, and A. Skowron (Springer-Verlag, Berlin, 2007), pp. 91–102.
[51] S. Gregory, in *Machine Learning and Knowledge Discovery in Databases, Part I, Proceedings*, Lecture Notes in Artificial Intelligence, Vol. 5211, edited by W. Daelemans, B. Goethals, and K. Morik (Springer-Verlag, Berlin, 2008), pp. 408–423.
[52] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval* (Addison-Wesley, Boston, 1999).
[53] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. Theor. Exp. (2008) P10008.
[54] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).
[55] A. Gavin *et al.*, Nature **440**, 631 (2006).
[56] G. Liu, J. Li, and L. Wong, in Genome Informatics 2008, Genome Informatics Series, Vol. 21, edited by J. Arthur and S. K. Ng (World Scientific, Singapore, 2008) pp. 138–149.
[57] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. X. Mayer, V. Stuempflen, and A. Antonov, Nucl. Acids Res. **39**, D220 (2011).
[58] The MATLAB code implementing the searching algorithm and the pruning procedure is available at http://home.deib.polimi.it/piccardi/IOPC.html.
[59] M. A. Serrano and M. Boguñá, Phys. Rev. E **68**, 015101 (2003).
[60] G. Fagiolo, J. Reyez, and S. Schiavo, Physica A **387**, 3868 (2008).
[61] L. De Benedictis and L. Tajoli, World Econ. **34**, 1417 (2011).
[62] C. Piccardi and L. Tajoli, Phys. Rev. E **85**, 066119 (2012).
[63] The network plots in this section were produced with PAJEK [66].
[64] J. B. Glattfelder and S. Battiston, Phys. Rev. E **80**, 036104 (2009).
[65] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[66] V. Batagelj and A. Mrvar, in *Graph Drawing Software*, Mathematics and Visualization, edited by M. Jünger and P. Mutzel (Springer-Verlag, Berlin, 2004), pp. 77–103.