



**HAL**  
open science

# An empirical characterization of community structures in complex networks using a bivariate map of quality metrics

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca

► **To cite this version:**

Vinh-Loc Dao, Cécile Bothorel, Philippe Lenca. An empirical characterization of community structures in complex networks using a bivariate map of quality metrics. *Social Network Analysis and Mining*, 2021, 11 (1), pp.37. 10.1007/s13278-021-00743-1 . hal-01809064

**HAL Id: hal-01809064**

**<https://imt-atlantique.hal.science/hal-01809064v1>**

Submitted on 14 Apr 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# An empirical characterization of community structures in complex networks using a bivariate map of quality metrics

Vinh-Loc Dao, Cécile Bothorel,\* and Philippe Lenca†  
IMT Atlantique, Lab-STICC,  
CNRS UMR 6285, F-29238 Brest, France

(Dated: March 9, 2021)

Community detection emerges as an important task in the discovery of network mesoscopic structures. However, the concept of a “good” community is very context-dependent and it is relatively complicated to deduce community characteristics using available community detection techniques. In reality, the existence of a gap between structural goodness quality metrics and expected topological patterns creates a confusion in evaluating community structures. We thus introduce an empirical multivariate analysis of different structural goodness properties in order to characterize several detectable community topologies. Specifically, we show that a combination of two representative structural dimensions including community transitivity and hub dominance allows to distinguish different topologies such as star-based, clique-based, string-based and grid-based structures. Additionally, these classes of topology disclose structural proximities with those of graphs created by Erdős-Rényi, Watts-Strogatz and Barabási-Albert generative models. We illustrate popular community topologies identified by different detection methods on a large dataset composing many network categories and associate their structures with the most related graph generative model. Interestingly, this conjunctive representation sheds light on fundamental differences between mesoscopic structures in various network categories including: communication, information, biological, technological, social, ecological, synthetic networks and more.

**Keywords:** Complex Networks, Community Structure, Community Characterization, Cluster Description, Graph Model.

## I. INTRODUCTION

The representation of complex networks using graphs opens tremendous possibilities to discover their structural characteristics and latter allows to explain different phenomena related to the system functionality. Hence, an impressive amount of work have been conducted to study large-scale structure in many network categories such as social networks, [28], [47], biochemical networks, [2], [1], [48], computer networks, [12], [44] etc. Among important features, community structure receives an immense attention since it could help to discover network organization and thus allow to explain several mechanisms that affect network evolution in a mesoscopic level; to understand different dynamic processes happening on the network; to study network behaviors associating to functional blocks and so on. Consequently, a variety of approaches to discover different aspects of modular structure have been developed in the last few decades [39], [18], [35], [42]. Several efforts have been also dedicated to evaluate and categorize these methods in a systematic way [31], [13], [8], [43].

Even it is widely accepted that communities are groups of nodes in a network where there are much more edges that connect nodes of the same group than edges that connect nodes in different groups [13], community quality evaluation has been always an controversial issue due

to its ambiguous definition. That is why community detection is still particularly problem-driven, which means the perspective of what is expected as a *good* community varies according to different contexts and there is no formal concept of community that is well mathematically formulated and widely accepted as the same time. For each specific problem, one may have a particular definition of a good community which reflects expected modular structures inside a network of interest. However, plausible solutions are generally difficult to be evaluated or validated without a presence of specialists who understand well the system in question and the discovery mechanisms of algorithms in disposition.

As a consequence, although several community detection methods are available in the literature, the question of which method to be adopted in order to find a specific structure of community remains a challenge to be solved [14]. We thus attempt to answer the question: “*what do structural communities in real networks look like?*” by characterizing systematically communities discovered in many real world network categories. This characterization helps network analysts to discern detectable types of community structures that could be found on their networks. For example, in a recent article, researchers show that immunization and epidemic control are most effective when they exploit the topological properties of communities [16]. The expected structure may exists or not depending on the property of the network under consideration, however a specific characterization of community structures may guide for a good conception of detection mechanism or an appropriate choice of

---

\* cecile.bothorel@imt-atlantique.fr, orcid:0000-0001-6602-1769

† philippe.lenca@imt-atlantique.fr, orcid:0000-0001-5014-2347

community detection technique.

Specifically, we focus on the evaluation of structural communities, which means communities are distinguished based on the interaction between their nodes through edges but not on contextual information neither network meta-data. One could criticize this approach since it is also possible that real communities in networks are not structurally good but yet well cohesive according to a more natural sense of community. However, most of the time, contextual information or metadata are missing and the only way to discover community structure is using network topology. Moreover, a generic analysis using only interaction information enables a comparative approach to contrast communities throughout different network categories, which are not allowed by sophisticated approaches using contextual information.

The paper is organized in the following way: firstly, some close related work are presented in Section II to give an overview about similar existing researches; then quality metrics that are used to characterize structural communities will be introduced in Section III. Then, Section IV describes community detection methods employed to identify and validate communities; subsequently, the network dataset and their uncovered hidden communities are analyzed in Section V using quality metrics presented in Section III. We then present a conceptualization of community structure using a bivariate representation approach which is introduced in Section VI, then based on this approach, we identify in Section VII different structural profiles of communities in various network categories such as: communication, technological, information, biological, social, ecological and synthetic networks. Finally, Section VIII draws some conclusions and envisages some potential perspectives for future work.

## II. RELATED WORK

This paper is not the first effort to bring a comprehensive survey of community detection algorithms in complex networks. Readers can refer to [14], [17], [21], [22], and [3] to get an overview on the different approaches. We also have provided a performance review of a dozen of well-known different methods based on extensive empirical experiments [10], regarding their scalability, community size distribution and co-performance measures comparing them with each other. In this paper, we focus more on a descriptive approach. Our objective is here to understand the topological nature of the communities produced by the algorithms.

Lancichinetti *et al.* characterize community structures of complex networks in different domains by observing the evolution of various qualities such as community scaled density, average shortest path, max internal degree, etc. in large scale networks according to discovered community size [27]. The evolution of these qualities in function of number of nodes in each cluster helps the

authors to deduce and characterize different structures found in many class of networks such as: Internet, communication, information, biological and social networks.

Guimera *et al.* demonstrate that modular networks in real world can be classified into distinct functional classes depending on the composition of connection profiles between their nodes [20]. Specifically, by using two metrics including within-module degree  $z$  and participation ratio  $P$  [19], a node in a community is characterized by seven different roles of hubs and non-hub nodes. Once the role of every node in a network partition is determined, the connectivity profiles of interactions in the network can be analyzed. Specifically, the authors determine two main classes of networks based on the presence of role-to-role connectivity profiles. The first class called *string-periphery* includes metabolic and air transportation networks which are rich in ultra peripheral interactions and hub interactions. The second class called *multi-star* includes protein interactome and Internet networks which are, on the other hand, rich in ultra peripheral-provincial hub interactions.

Leskovec *et al.* investigate the variation of community structure in large scale networks using *conductance* metric [30]. In fact, the authors measure the variation of the lowest community conductance in function of community size. This variation depicts a so-called *network community profile* which helps to characterize community quality over a wide range of size scales. The authors also point out that communities attain the best quality (in terms of conductance) at a characteristic size of around 100 nodes and provide evidences of a high presence of core-periphery community structure in real networks through numerous empirical experiences.

Coscia *et al.* generalize the problem of community detection discovery by reconsidering the question of what can be considered to be a community [8]. The authors then resume popular methods in the literature according different quality aspects such as density-based, vertex similarity-based, action-based or influence propagation-based (for example, [45] introduces Influence-Guided Spherical K-means, an algorithm for community detection based on a new centrality measure, the Shared-Influence-Neighbor similarity). A definition-based classification of community discovery methods according to a large number of community features is then introduced. This classification approach shifts the attention from how communities are detected to what kind of communities to detect and provides another point of view regarding to community detection.

The most common and fundamental point between this paper and the previously mentioned work is the exploratory objective to characterize communities in complex networks by observing qualities using statistical metrics. Concretely, we contribute a methodology to describe community topologies in a systematic and generic way that can be extended to any category of networks. This means one can mechanically apply the same analysis procedure to explore community structures of any network

of interest.

### III. COMMUNITY QUALITY CHARACTERIZING METRICS

The purpose of this work is characterizing structural properties of communities that could be found by using various community detection techniques. Since the idea of community structure varies from one context to another, it is not expected that a finite set goodness features could fit every intuition of what a good community is and the choice of any set of metrics would be adversarial unless a specific context is clearly defined under a constrained circumstance. Meanwhile, many goodness metrics define community qualities based on the conditions where they are found. Consequently, in order to remain the analysis as generic as possible, these metrics are not considered in this paper to characterize communities. Therefore, we restrict our list of quality metrics of interest in the analysis by applying the following criteria from the highest to the lowest priority:

- Since we are characterizing communities in different types of networks, we are only interested in metrics who delineate communities themselves, not in a relative relation with the global structure of networks where they are found (such as Cut ratio [47], Modularity [35] or Description Length [42]) even though their efficacy can not be ignored.
- Potential metrics for the analysis must be relatively uncorrelated from one to another throughout a wide range of networks in order to illustrate different aspects of structural characteristics.
- A metric whose concepts can be represented intuitively and visually in order to describe most distinguishable characteristics is preferable than a metric that reflects statistical ideas which are difficult to be presented by simple topologies.

In the later part, we show that the choice of structural metrics using these criteria helps to distinguish community structures that are found in many categories of networks using different techniques of community detection. But firstly, we introduce some general notations that helps to define metrics commonly used to measure network or community structures.

#### A. General notations

We formulate different quality metrics using the following notations: given an undirected and unweighted graph  $G = (V, E)$  which is composed of a set of  $n = |V|$  nodes and  $m = |E|$  edges where  $E = (u, v) : u, v \in V$ . Each node in a graph is characterized by a degree  $d(u)$  which is the number of connections that it has with the other

nodes in the graph. Given a cluster  $S$  of  $n_S$  nodes, which is a subgraph of  $G$ , a function  $g(S)$  quantifies a structural goodness feature of  $S$  according to a particular expectation of community quality. We denote  $m_S$  as the number of edges inside  $S$ ,  $m_S = |(u, v) \in E : u \in S, v \in S|$ ;  $c_S$  be the number of edges that connect  $S$  to other nodes outside of  $S$ ,  $c_S = |(u, v) \in E : u \in S, v \notin S|$ . The number of connections of a node  $u$  in a subgraph  $S$  with other nodes in  $S$  is called internal degree of node  $u$  and is denoted as  $d_{int}(u)$ . As such, the external degree of node  $u$  can be calculated as  $d_{ext}(u) = d(u) - d_{int}(u)$ . The relation between internal and external degrees of nodes in  $S$  with its number of edges can be resumed as:  $\sum_{u \in S} d_{int}(u) = 2m_S$  and  $\sum_{u \in S} d_{ext}(u) = c_S$ .

Based on the previous preliminary, we present in the following section a list of commonly used goodness metrics which can be classified in four principle families:

#### B. Metrics based on internal edge density

- Density

$$\text{density}(S) = \frac{m_S}{n_S(n_S - 1)/2} \quad (1)$$

This metric captures the idea that nodes in a community must be densely connected wherever possible. It quantifies the fraction of edges inside  $S$  over the total possible edges that could be established.

- Scaled density

$$\text{sc\_den}(S) = \frac{2m_S}{n_S - 1} \quad (2)$$

Scaled density is a kind of normalized density which is defined as  $n_S$  times the density of the community [27], [25]. This normalization is usually applied to palliate an issue due to the fact that the number of edges in a sparse network increases linearly with its size, however the number of possible edges increase quadratically. As a consequence, traditional edge density could not well distinguish large communities and this modification by a multiplication with the number of node is believed to reflect better edge density concept in real world networks.

#### C. Metrics based on centralized or hub structure

- Hub dominance

$$\text{hub\_dom}(S) = \frac{\max_{u \in S} d_{int}(u)}{n_S - 1} \quad (3)$$

Internal edges of a community can be distributed in various ways around its nodes, either concentrating around a few numbers of high centralized nodes or

uniformly divided into every node. The hub dominance metric is designed to identify the level of central organization around well connected nodes. The higher this metric of a community, the more likely it has a hub-like structure [27], [25].

#### D. Metrics based on triadic structure

- Clustering coefficient (CCF)

$$\text{CCF}(S) = \frac{3\Delta_S}{T_S} \quad (4)$$

Where  $\Delta_S$  denotes the number of triangles in community  $S$  and  $T_S$  indicates the number of triples of vertices in  $S$ , which means number of connected subgraphs consisting of 3 vertices. This metric reflects the probability that the adjacent vertices of a vertex are connected. This is a well-known metric which is usually used to evaluate modular structure in networks. It is based on the concept that pairs of nodes with common neighbors are more likely to be connected [5].

- Triangle participation (TPR)

$$\text{TPR}(S) = \frac{\sum_{u \in S} \delta_{uS}}{n_S} \quad (5)$$

Where  $\sum_{u \in S} \delta_{uS} = 1$  if node  $u$  belong to at least one triangle in community  $S$  and  $\sum_{u \in S} \delta_{uS} = 0$  if node  $u$  does not belong to any triangle in community  $S$  [47]. There is a slightly difference between the clustering coefficient and the triangle participation, while the former considers a good community based on the number of possible connections which could be constructed in the community, the latter only cares about whether there are many individuals of the community participate or not in tight connections (cliques).

#### E. Metrics based on external connectivity

- Expansion:

$$\text{expansion}(S) = \frac{c_s}{n_s} \quad (6)$$

The metric measures the number of edges per node that point out side a cluster [47]. It represents the relative out degree of a cluster over its size. The higher the expansion of a community, the stronger the its connection with the rest of the network. Generally, in a common sense, community detection methods try to minimize inter-community connectivity and hence reducing community expansion.

- Conductance

$$\text{conductance}(S) = \frac{c_s}{2m_S + c_s} \quad (7)$$

The conductance represents the fraction of degrees of a community that points outside over the total of its degrees. The conductance reveals how much the direct neighbors of a node in the community belong to neighborhood communities. In other words, the higher the conductance, the more likely that nodes connect to the community belong to another community. Leskovec *et al.* show that finding a configuration in networks that minimizes the conductance of communities helps to identified good network community profile [30].

- Average Out Degree Fraction

$$\text{meanODF}(S) = \frac{1}{n_S} \sum_{u \in S} \frac{|(u,v) \in E : v \notin S|}{d(u)} \quad (8)$$

with  $ODF_S(u) = \frac{|(u,v) \in E : v \notin S|}{d(u)}$  is called the out degree fraction of node  $u$  in subgraph  $S$ . The meanODF value indicates the average of out degree fraction of nodes in a community, a low meanODF implies that nodes in the community connect primarily with other nodes inside the community while a high meanODF means that nodes connect preferably to nodes in other communities rather than to the ones in its own [47].

- Maximum Out Degree Fraction

$$\text{maxODF}(S) = \max_{u \in S} \frac{|(u,v) \in E : v \notin S|}{d(u)} \quad (9)$$

The maxODF reflects the maximum of fraction of edges of a node in community  $S$  that connect outside  $S$ . This metric helps to quantify the interaction of the most active node of community  $S$  with the rest of the network.

Besides, there are many other metrics in this family such as FlakeODF, Cut ratio, Normalized cut, etc. However they expose high correlations with specified metrics in our analysis and are not listed here [47].

## IV. COMMUNITY DETECTION METHODS

We here present community detection methods that have been used in order to study community structure in our network dataset. The choice of analyzing methods should not be considered neither an exhaustive nor a well representative list. Since our objective is to focus on different modular structures that could potentially be identified on real-world networks, we only chose a few numbers of methods whose performances have been

proven in the literature to be reliable and can be well accessed by a large numbers of analysts. The only criterion we take into account is that these methods use different approaches to discover communities. While the edge betweenness method [18] is based on edge centrality detection in order to break networks into several communities; the Louvain method [6] optimizes local modularity by iteratively folding nodes into meta-nodes; the label propagation [39] determines the community of a node by considering the memberships of its neighbors; and the Infomap method [42] relies on finding a configuration that maximizes the compression of a random walks represented by an encoded binary sequence. Of course one could argue that by using only a few numbers of methods, it is likely that some kind of structures are not well covered in the analysis. Although it is a very pertinent requirement, within this study, the authors find that the utilization of some representative methods could already help to reveal substantially many interesting community structures. A summary of community detection methods that have been used to discover networks is presented in Table I.

Method	Approach	Complexity
Edge betweenness [18]	Betweenness centrality	$\mathcal{O}(nm^2)$
Fast greedy [7]	Modularity	$\mathcal{O}(nm \log n)$
Louvain [6]	Local modularity	$\mathcal{O}(\log n)$
Spectral [35]	Modularity	$\mathcal{O}(n(m+n))$
Walktrap [38]	Dynamic distance	$\mathcal{O}(mn^2)$
Label propagation [39]	Topological closeness	$\mathcal{O}(m+n)$
Infomap [42]	Information compression	$\mathcal{O}(n(m+n))$
Spin glass [40]	Energy model	$\mathcal{O}(n(m+n))$

TABLE I. A summary of community detection algorithms used to study community structure in the analysis. They are used to identify communities in the network dataset.

### A. Edge betweenness community detection

Girvan and Newman proposed a method [18] to identify boundaries between communities in network by measuring a factor called *edge betweenness centrality* which reveal the contribution of each edge in the network for constructing shortest path way between two arbitrary nodes. An edge with a high *betweenness centrality* indicator represents an important connection that joints two *compact* groups in a network. It means that it is very likely that the shortest pathway between two nodes of these groups must go through the edge. Thus, removing such a high *edge betweenness* degree connection will probably separate two loosely connected clusters. The authors construct a community detection method based on the idea that if one gradually removes high *betweenness centrality* edges in a network, after some iterations, the network will be disconnected and nodes located in different components can be considered as prospective

communities.

### B. Fast greedy method

This method discover communities in networks using a greedy method to optimize the modularity objective function throughout many iterations. The modularity  $Q$  is defined as the difference between the number of edges within communities and the number of expected number of such edges. It can be written:  $Q = \frac{1}{2m} \sum_{ij} [A_{ij} - P_{ij}] \delta(S_i, S_j)$ , where  $P_{ij}$  represents the probability of having an edge between  $i$  and  $j$ ,  $\delta(S_i, S_j) = 1$  if  $S_i = S_j$  and  $\delta(S_i, S_j) = 0$  otherwise. This is a hierarchical agglomeration algorithm which is well-known for its competitiveness in time complexity  $\mathcal{O}(nm \log n)$  and can be reduced to  $\mathcal{O}(n \log^2 n)$  in sparse networks where  $n$  corresponds to the number of vertices of the network [7]. The idea of the authors to reduce the complexity of the agglomeration process is only that they only consider amalgamations between nodes that share at least one common edge. In that way, they can make use of the data structure of the original graph in order to keep track of the modularity changes  $\Delta Q_{ij}$  when merging two nodes  $i$  and  $j$ .

### C. Multi-level community detection - Louvain method

This heuristic method employs modularity optimization in order to discover hierarchically community structure for networks and is claimed to outperform all other known methods in terms of computation time [6] with a good compromise quality measured by modularity on large networks. The algorithm is executed through two concatenated phases that repeat iteratively. At the first step, each node of the network belongs to its own community and is considered to be merged with its neighbors to establish a new community in a way to gain a maximum increase of modularity. The second step consists of constructing an aggregated network where communities in the original network become new nodes; links between two new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities; links between nodes inside a same community in the first step become self-loop links in the new network. Once the second phase is finished, the first phase of the algorithm is then reapplied to the resulting network. These two phases are iterated until no additional modularity is obtained.

### D. Walktrap method

This method estimates the distances between vertices in a network using a structural similarity measure based on random walks. Such that the distance between two

vertices must be large if they belong to different communities and must be small if they belong to the same community. The approach relies on the intuition that a random walker may have a tendency to be trapped into densely connected parts of networks where nodes have a similar stochastic state. In fact, two vertices are considered to be similar and belong to the same community should “see” the other vertices in the same way. Specifically, the authors proposed to define a similarity distance  $r_{ij}$  between two vertices  $i$  and  $j$  as a function of the difference between the probabilities  $P_i^t$  and  $P_j^t$  to go from  $i$  and  $j$  to other vertices in a short number of  $t$  steps:  $r_{ij}^2 = \sum_{k=1}^n \frac{(P_{ik}^t - P_{jk}^t)^2}{d(k)}$  [38]. Then, a traditional hierarchical clustering algorithm is used to find communities from this dynamic distance.

### E. Label propagation method

This method is based on the idea that nodes in a network have a tendency to participate into the community where the majority of their neighbors are found [39]. The algorithm initializes every node with a unique label and repeats modifying these labels at every iterative step. In each step, each node adopts a new label that most of its neighbors currently have and this process is expected to help identifying densely connected groups of nodes that have unique labels and considered as communities. Ideally, the iterative process should converge when no node in the network changes its label, however it is normally possible that nodes in a network have an equal maximum number of neighbors in two or more communities. In these cases, the algorithm breaks ties randomly among the possible candidates.

### F. Infomap method

With a primal purpose to understand the flow of information on networks, this method is designed in order to decompose a network into communities by optimally compressing description of information flows on the network [42]. There is a conceptual distinction of community notion in this method with the traditional ones. In fact, instead considering density-related elements, the authors consider a community as a group of nodes among which information flows quickly and easily, and so that they can be aggregated and represented by a single well-connected module. In order to do that, it can be imagined that the algorithm employs a random walker to describe information flows on a network of interest and then exploits the regularity of the random walker’s path that have been traced and encoded. The modules are then determined as the configuration that minimizes the amount of necessary codeword length in order to compress the regularity in the path of the random walker.

## G. Modularity spectrum-based method

Spectral method, also known as leading eigenvector, searches for a partition that maximizes the modularity fitness function using a spectral partitioning calculation on a *modularity matrix*  $B$  [33]. This matrix is defined as  $B_{ij} = A_{ij} - P_{ij}$  where  $A_{ij}$  denotes the adjacency matrix of the graph under consideration and  $P_{ij}$  represents the probability  $p_{ij}$  for an edge to fall between the pair of vertices  $i, j$ . Here,  $P_{ij}$  reflects the expectation of the existence of an edge between two arbitrary vertices in an associated graph where node degree sequence keeps unchanged and modular structure is considered not to be presented. Inspired from the spectral partitioning problem to minimize the cut size of a clustering using *Laplacian* matrix, the author find out that the leading eigenvectors that correspond to the positive eigenvalues of the *modularity matrix* helps to find a good partition that maximize the modularity.

### H. Spin glass method

In this method, the problem of community detection is interpreted as finding the ground state of an infinite range spin glass, which is the configuration that minimizes the energy of the system. Determining this state configuration suggests useful information to locate communities being groups of nodes that have the same states. The basic principle of the model is that edges should only connect vertices of the same class, which have the same spin state. Here, the formulation of system energy at the same time rewards internal edges between nodes of the same group and penalizes missing edges between nodes in the same group, penalizes edges between nodes of different groups and rewards non-links between different groups [40].

## V. AN EMPIRICAL ANALYSIS

In this section, we describe some statistical properties of networks that will be included in the following analysis. It is expected that networks in each category are spread in a wide range of structural measures. However, available biological networks that have been published and analyzed widely are relatively small in comparison to the other networks of the other families. Besides, due to the complexity of the analysis process, we limit the domains of interest at 5 categories which are commonly researched and where numerous networks are available. The number of networks considered is 108 which is relatively large in comparison to many studies in the art. Many notable related work where some of these networks are also employed to study community structure could be mentioned for a quick reference: Orman *et al.* use 6 networks to evaluate the structure of communities discovered by several detection techniques [37]; Lancichinetti

*et al.* use 15 networks to characterize structural communities [27]; Darko *et al.* use 16 networks to reveal differences between structural communities and ground truth; Leskovec *et al.* use over 100 networks to analyze network community profile [30] and 230 networks to evaluate the goodness of ground-truth communities in social networks, within this number, 225 samples of the Ning platform’s networks [36] are aggregated [47]. Table II resumes the composition of networks that have been analyzed in this paper.

### A. A description of empirical network dataset

Category	Size	Nodes	Edges	Notable networks
Biological	7	1860	10763	Protein, yeast
Communication	9	39595	195032	Email, forums
Information	25	38358	159812	Citation, Amazon
Social	37	6888	49666	Facebook, Youtube
Technological	19	18431	48494	Internet, P2P
Miscellaneous	11	4298	49033	Ecology, synthetic
Total*	108	1.99M	9.08M	

TABLE II. A summary of network dataset used in this analysis where **Size** is the number of networks analyzed in each category, **Nodes** and **Edges** indicates the average number of nodes and edges of networks in each category respectively. \*The last row shows the total number of networks, nodes and edges in the whole dataset. This dataset is collected from several sources including: <http://networkrepository.com> [41], <http://konect.uni-koblenz.de> [23], <http://snap.stanford.edu> [29]

Some notable structural measures of networks in the dataset are illustrated in Figure 1. It is noticeable that apart from biological networks which are relatively small, the other classes cover quite a wide range of number of nodes, edges, mean degree, clustering coefficient and edge density. Since real world networks are relatively sparse, the number of edges increase linearly in function of the number of nodes and consequently, the edge densities decrease linearly by number of nodes (since the number of possible connection increase quadratically by number of nodes). This sparsity property can be easily noticed from Figure 1(a,d). From Figure 1(b), it can be seen that the mean degrees of the networks in the dataset vary principally between 1 and 100 edges per node except for 2 communication networks. In a global point of view, networks in the dataset have a quite strong modular quality since most of them have relatively high clustering coefficient as shown in Figure 1(c). This quality will be investigated more in community-level in a following section.

### B. Evaluating community structures using quality metrics

In order to characterize structural communities of different types of networks, we apply various community detection methods on the dataset grouped by category of networks. We tested the official implementations of community detection methods introduced in Table I on the dataset collection presented in Table II. The implementations are provided officially by *igraph* (R), the popular network analysis library which can be easily accessed by a large public. The default parameters configured by the implementations remained unchanged during our experiments.

Once communities are produced, the quality metrics presented in Section III are used to evaluate the quality of detected communities. Since many metrics reflect close structural properties, we analyze the correlations between the corresponding qualities on the detected community sets. This analysis allows to select only the most representative structural metrics to delineate community structures.

Figure 2 illustrates the correlation matrices of different structural qualities measured on various community sets identified by community detection algorithms presented in Section IV over 5 classes of networks and the whole dataset. Note that only communities whose sizes are at least 3 nodes are taken into consideration in the figure since many metrics are meaningless for too small communities (which contain one or two nodes). It is important to note that although some statistical metrics are only significant when measuring on large communities, the corresponding correlation matrices for large scale communities resemble globally with those of Figure 2. Specifically, a calculation using only large communities of more than 10 nodes gives quite similar and consistent correlation scores. The employment of representative quality metrics is globally justifiable on the whole range of community size scales. And so that the same metrics can be relatively significant to represent communities on the whole range of community size.

As we can see in Figure 2, there are two groups where metrics are consistently correlated from one to another. The first group includes **maxODF**, **meanODF** and **conductance** who represent community external connection with very high correlation coefficients (except for **maxODF** and **meanODF** in information networks with a relatively weak relation of 0.51). Besides, the **expansion** metric also belong to this group in technological, information and biological networks with high correlation scores and more loosely in the other types of networks. The second group consists in **TPR** and **CCF** who expose triadic tight-knit structures and are observed with very high correlation scores in every case of network category. The lowest correlation score between **TPR** and **CCF** is reported at 0.81 in information networks and approximately around 0.90 in all the other cases. Without losing the generality, in our analysis, these 2 groups of metrics could be reduced



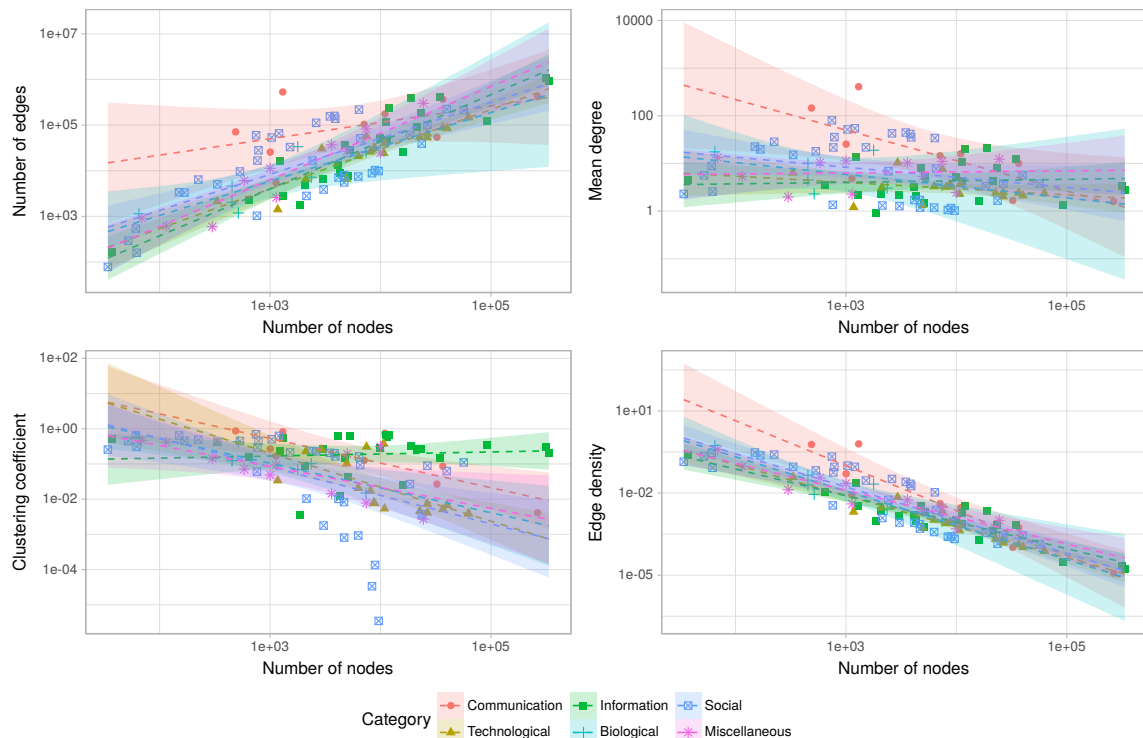


FIG. 1. From left to right, up to bottom, we illustrate structural measures of the 108 networks: (a) Number of edges as a function of the number of nodes, (b) Mean degree  $\langle k \rangle$  as a function of the number of nodes, (c) Clustering coefficient in function of the number of nodes, (d) Edge density in function of the number of nodes. The dark backgrounds represent the 95% confidence intervals of the regression model outputs used to estimate the linear relations for each network category.

to two representative metrics representing two structural properties.

Hub dominance (`hub_dom`) is the only metric who is quite independent of all metrics in the two previous groups in every network category. The highest absolute correlation score between `hub_dom` with these metrics is 0.42 with `maxODF` in social networks, which is still a relative low correlation. This latter, however, is generally correlated with `density` except for the case of communication networks where they are quite orthogonal. In the mean while, scaled density (`sc_den`) shows an inconsistent association throughout the studied network categories. It is close to `CCF` and `TPR` in biological networks but approaches expansion in social networks.

Based on this analysis, the above community quality metrics can be grouped in 6 classes that are presented in Table III according to their correlations over the studied dataset. In other words, these quality metrics are more correlated with ones in the same groups than with the others. Consequently, it is preferable to describe community structure using a cross combination of metrics in these groups. We present in the following section a characterization of internal community structure by a descriptive approach using an association between metrics in 2 different groups. Then we demonstrate by empirical evidences that our approach helps to recognize different community structures in communication, infor-

mation, technological, biological, social, ecological and synthetic networks.

In fact, the previous analysis shows that internal and external structures of communities are generally not correlated. They reflect different facets of community structures. Consequently, the characterization of community structure can be realized separately from 2 distinguishable levels of observation. In this paper, we focus on characterizing internal community structure. Readers who are interested in analyzing community external connectivity can refer to another work presented in [9] where communities are portrayed by two variables: the level of external interaction and the distribution of these interactions over community border nodes.

## VI. A BIVARIATE CHARACTERIZATION OF COMMUNITY STRUCTURE

In this part, we present a categorization of community structure in a descriptive way to illustrate different modular structures. This is an extension of our previous proposition in evaluating communities using a descriptive approach [9] for internal aspect of community structure. We propose a categorization of modular structures using a couple of representative goodness variables to reflect highlight structural characteristics of communities in real

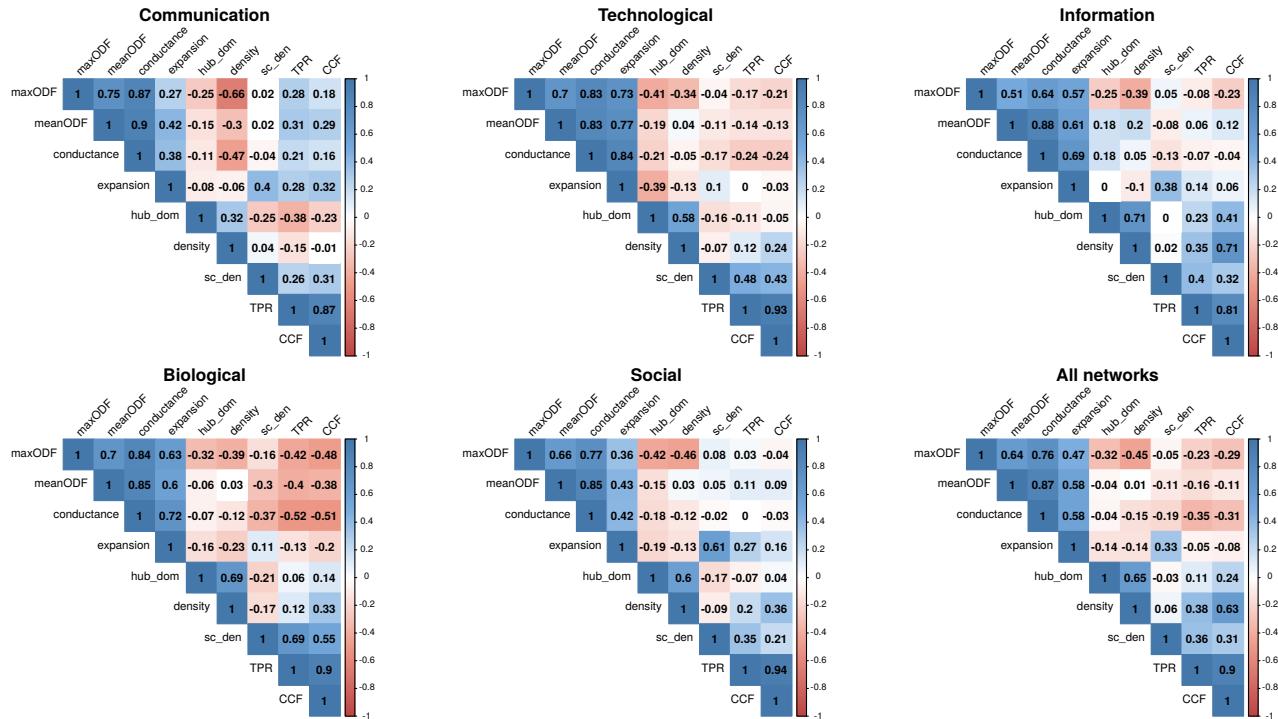


FIG. 2. The Pearson correlations of community metrics measured on the communities detected by the set of community detection methods on the network dataset. These correlation are calculated based on scores of metrics measured on communities that contain at least 3 nodes. Metric correlations are analyzed by group of networks in different domains. Quality metrics are presented in the 6 sub-figures in the same order for a comparative observation. Correlation scores with low estimated significant levels ( $P$ -value  $> 0.01$ ) are reproduced in a blank background.

Metrics	Common concept
maxODF,meanODF,conductance	External activeness
expansion	External connectivity
hub_dom	Centralized connectivity
density	Internal edge density
sc_den	Average internal density
CCF, TPR	Internal triadic closure

TABLE III. Groups of quality metrics that reflect different aspects of community structural property. Two metrics belong to a same category if they show a high correlation over the sets of structural communities. The **Common concept** column precises common structural features that members of each group reflect.

world networks. Here, we focus on internal community structure, i.e. density, sc\_den, CCF, TPR and hub\_dom, will be in the shortlist of interest.

### A. Which metrics fit?

It is well-known that density have a weakness in describing communities of different sizes since in real networks, the number of edges normally increases linearly with its size (real networks are often sparse) but the number of possible connection increases quadratically. As

a consequence, the quality of large communities is usually under evaluated in comparison to small communities. Scaled density (sc\_den) palliates this issue by multiplying the density with the community size, so mathematically its concept is very close with the average degree of a community which is measured by  $\langle k_S \rangle = \frac{2m_S}{n_S}$ . This metric reflects a very important feature of communities and is often used to evaluate community quality in a common sense. However, given a specific value of scaled density, one have several ways to redistribute edges inside a community in a manner that its internal topology changes crucially. In other words, scaled density does not characterize community internal configuration of degrees. This is the reason why we do not use scaled density or traditional density to represent community topology.

The clustering coefficient and the triangle participation ratio (CCF and TPR respectively) are relatively close in their definition and it has been proved to be highly correlated through the previous empirical analysis. They reflect an important topological feature by implying the concept that two arbitrary neighbors of a node in a community should be also connected. This idea is somehow relatively close with the density qualities since a network with high CCF, TPR scores is normally dense; however the opposite way is not always correct, which means a dense network does not necessarily have many triangular connections. Here, we select one metric among CCF

and TPR to describe a common structural property called *transitivity*. Depending on the topology of networks or communities under consideration, one metric will work better than the other. On a same network, CCF score is generally lower than TPR score and hence CCF has a better resolution for networks where triangles are dense. On the the other side, TPR magnifies better topological differences in networks where only a few triangles exist. A further investigation on the dataset shows that there is approximately 90% of networks whose clustering coefficients are larger than 0.01 and this number is around 60% for a coefficient of 0.1 (see Figure 1(d)). This evidence leads to a preference of CCF over TPR to describe the clique dominance characteristic since the networks of interest are quite dense.

Another topological dimension that we employ to describe communities is *hub dominance* which is represented by `hub_dom` metric. Similarly to CCF and TPR, this metric reflect a structural feature of edge organization in a network or community. Specifically, it characterizes whether edges are distributed around one or a few members of their community and make them becoming hubs of connection. We illustrate in the next section that the combination two dimensions quantified by a couple of values (CCF, `sc_den`) reveals distinctive topological structures that could help to get insights on how communities in different networks look like.

### B. Locating community structures in a bi-dimensional space

After choosing two principle characteristics corresponding to two dimensions of community quality space, we describe internal community structures in different locations of this space. In order that the distinction of representative topologies in different coordinates stays clear, we profile them in a coarse-grained description level. Specifically, we considerate 4 fundamental coordinated zones corresponding to 4 underlying topologies which are emphasized in Table IV. These classes of topologies could be explained as following:

Type	Transitivity	Hub dominance	Topology
1	Low	Low	String-based
2	High	Low	Grid-based
3	Low	High	Star-based
4	High	High	Clique-based

TABLE IV. Four distinctive topologies characterized by *Transitivity* (CCF) and *Hub dominance* (`hub_dom`). There is no clear boundary between high and low values in the two dimensions, it is to be specified in accordance with the context. The distinction is more clear for medium and large size communities.

- *String-based topology* of a community is determined by low values of transitivity and hub dominance metrics. The low scores in these two representative

dimensions regulate that there is relatively nearly no presence of clique structure nor hub node. For large communities, there could be one or a few hubs and cliques established, but not enough to dominate the global structure. These communities can be considered as a consequence of a ramification between several sub-strings which generate a few loops and hubs in their intersections. String-based topologies could have a form that looks like chains, braids, rings, etc. as shown in Figure 3(a) depending on the context.

- *Grid-based topology* can be recognized by high values of transitivity and low values of hub dominance metric. The absence of hub nodes in the community organization is probably the most common feature with the string-based topology. Hence there is a homogeneity in the connection pattern between nodes of the grid-based topology. Besides, a high value of transitivity imply that the majority of nodes participate in tight-knit triangular structures which could themselves, at the same time, be attached between one to another to create larger and compacted structures. Grid-based communities generally have large sizes since small ones are usually degenerated into strings, loops or hub structures. In other words, grid-based structures are not recognizable by observing in a small scale or a local scale of communities. Popular topologies of this family consist of lattice topology, partially mesh topology as shown in Figure 3(b).
- *Star-based topology* which sometimes can be considered as tree-based topology is probably one of the most popular structures in networks of many fields. It can be perceived by low values of transitivity and high values of hub dominance. A low transitivity indicates that there is not or very few cliques. On the other hand, a high hub dominance value implies the occurrence of a “key connection” which attracts many edges in its community to become a hub. Some popular topologies which could be found in this class include: flake structure with one central hub and several peripheral hubs; hierarchical tree structure. There is actually a close relation between star-based/tree-based and string-based topology such that in some contexts, a hierarchical tree could be seen as a string and vice versa depending on the point of view. The essential difference of these two topologies which can be observed from our representation space is that the more *edge-attractive* the hub(s) in a community, the more it approaches the *star-based* topology. Note that in graph theory, a tree is an acyclic connected graph. However, in this context, trees accompanied by a few loops are classified in star-based topology unless loops dominate excessively the global community structure. Some representative star-based topologies are shown in Figure 3(c).

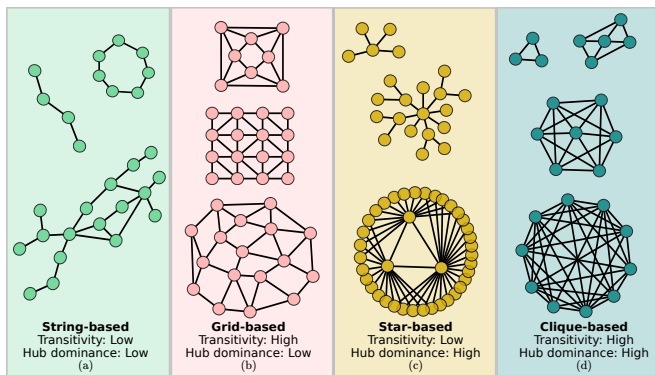


FIG. 3. Topology families, from left hand side to right hand side (a) String-based, (b) Grid-based, (c) Star-based, (d) Clique-based. Depending on the context, one community can belong to different topological families according to specific criteria of analyst reflected by their determination of frontiers between these families.

- *Clique-based topology* is quite common in small and very small communities but very rare in medium and large communities. It is recognized by high scores of transitivity and hub dominance. A simple interpretation of this class of topology is that every node must be connected with every other node of its community in an ideal situation. In a more relaxed context, nodes are not required to connect with all other nodes, but with a majority in order to establish a tight and compact structure. The clique-based topology is quite close to the grid-based topology in many ways. The most notable difference between them is that in a clique-based community, every node must be in the neighborhood of the other nodes of the community (direct connection or by one/two intermediate connections maximum), whether it is not necessary that every node must be close to each other in grid-based topology. Some representative clique-based topologies are shown in Figure 3(d).

A community structure whose transitivity and hub dominance scores are medium needs more investigation to be deduced. Since neither hub, clique nor random structure could dominate the whole community, its topology depends on the distribution of hubs and cliques in the community. It can be composed of a mixture of different component structures presented previously to become a homogeneous and more complex topology. It can also be a simple attachment between various dissimilar structures to establish a heterogeneous unit. In a point of view of dynamic community's evolution, communities in this class might be considered as being in a transition period between elementary structures. Alternatively, it could be a saturated state where communities attain a certain diversity and remain their complex structures. Further extent researches, which will not be mentioned in this paper, are deserved to cover more exhaustive aspects of

this subject

### C. Locating network models in the topological space

Based on the idea that real networks and communities are constructed throughout different mechanisms, their topologies could be in some ways mimicked by using graph generating models. We attempt to locate networks created by popular graph models of the literature in the presented space in order to match them with the most resembling representative topology.

- *Erdős-Rényi model*[11] is among the first models proposed to describe the generation of *random graphs*. In this models, two parameters are required to generate a graph which is a fixed number of vertices  $n$  and a connection probability  $p$  between two arbitrary vertices (alternatively the number of edges  $m$ ). Each pair of vertices is then connected independently of the other pairs with the probability  $p$ , which reflect the randomness property of the resulting graph. The expected number of edges and mean degree of the graph is calculated by  $\langle m \rangle = \frac{pn(n-1)}{2}$  and  $\langle k \rangle = p(n-1)$  respectively. The distribution of degree is binomial or Poisson for large graphs [32]. If we set  $n$  and  $p$  parameters of the model in a way that the model creates a random graph whose average degree approaches real networks:  $\langle k \rangle = p(n-1) = c > 1$ , where  $c$  is a constant and  $c \ll n$ ; the graph will almost surely have a big component containing a large portion of vertices and very small components of less than  $\mathcal{O}(\log(n))$  vertices. This configuration produces vertices that have all around  $c > 1$  connections. In this paper, without any further mentions, we refer to random graphs as ones created by this configuration, whose average node degrees approach those of real networks. Since a random network is constructed from a homogeneous stochastic mechanism, there is normally no hubs nor cliques which means low transitivity and low hub dominance values. A typical random graph constructed with a small value of  $p$  will have its largest component topology resembles the string-based topology as shown in Figure 3(a). In an extreme regime, when the probability of connection  $p$  approaches 1, the associated random graph becomes *nearly complete* as the average degree  $\langle k \rangle$  approaches  $n-1$ , which means every vertex connects with almost every other vertex as illustrated in Figure 3(d). The location of typical random graph's topology in function of two dimensions: transitivity and hub dominance is illustrated in Figure 4 in the bottom left-hand conner which associates to low scores of CCF and `hub_dom`.
- *Watts-Strogatz model* produces networks with *small-world* property, which normally means that

any arbitrary pair of nodes can be connected through a small number of intermediate nodes and the average geodesic distance grows proportionally to the logarithm of the number of nodes  $n$  of the network:  $L \propto \log(n)$ . The model is built to characterize the observation that many real world networks show this property of small path length connectivity and highly clustered like regular lattices which implies a high presence of triadic closures [46]. The generation of a small world network can somehow be considered as an interpolation between regular pattern networks and random networks. From a ring lattice with  $n$  nodes and  $k$  edges per node, each edge is redistributed randomly with a probability  $0 < p < 1$ . The authors find that a small value of  $p$  reduce significantly the path length characteristic of a regular network where nodes are only connected locally. This can be explained as rewired edges create shortcuts between remote areas of the network and hence reduce considerably network characteristic distance. A typical small-world network can be described using an intermediate value of  $p$ , so that the distance of two arbitrary nodes are very small, the clustering coefficient stay high since the random perturbation is not strong enough to break the local structures of nodes in the lattice ring. Besides, the shape of the degree distribution in the network is quite similar to that of a random graph where every node has around  $k$  neighbors and there is normally no hub dominance phenomenon. The topology of a typical small-world network is relatively homogeneous and looks like a grid-based topology from a local observation as shown in Figure 3(b). The location of its topology in function of two dimensions: transitivity and hub dominance is illustrated in Figure 4 in the bottom right-hand corner which associates to high CCF scores and low  $\text{hub\_dom}$  scores.

- *Barabási-Albert (BA) model* [4] is originated from a discovery that the distribution of vertex degrees in many real world networks such as: genetic networks and World Wide Web networks, are quite heterogeneous. Specifically, vertex connectivity follows a *power-law distribution*, which means the probability that a vertex connecting to  $k$  neighbors in its network equals  $p(k) = Ck^{-\alpha}$  where the constant  $C$  is fixed by a normalization requirement and  $\alpha$  is the power-law coefficient. This coefficient varies between 2 and 3 in many networks where the degree sequences are estimated to follow this model. Networks possessing this statistical feature are called *scale-free* by Barabási *et al.* to highlight the scale invariance property. This feature is explained by the authors as a consequence of two principle mechanisms: firstly, networks expand gradually by attracting new vertices to existing ones; secondly, these new vertices have a tendency to attach preferentially to vertices that are already

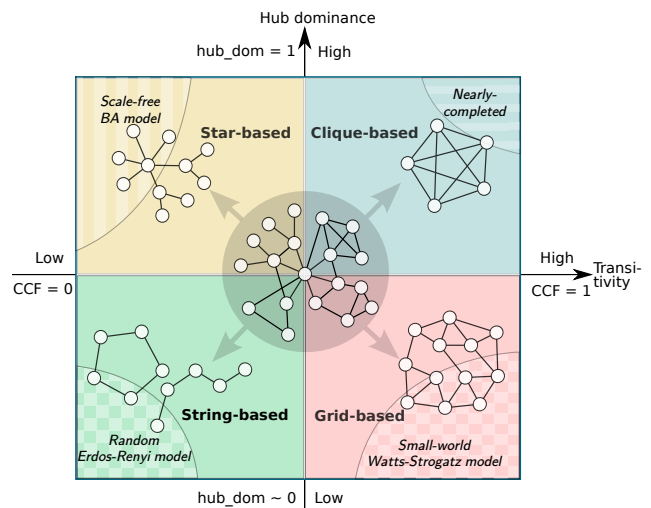


FIG. 4. A categorization of internal community structure according to two structural property dimensions: hub dominance and transitivity represented by  $\text{hub\_dom}$  and CCF respectively. Four representative topological communities are exemplified in 4 coordinating zones according to their corresponding  $(\text{hub\_dom}, \text{CCF})$  scores. The borders between different topologies are usually not clear and can be delineated according to the context. Characteristic community size should be taken into consideration when separating characterized zones since the bigger the community size, the more likely that hubs and cliques become less significant, which means lower thresholds will be more plausible.

well connected. That is why this model is often known as preferential attachment model, implying that the more connected a vertex, the more likely it receives new edges. This mechanism makes scale-free networks hub-profuse since “*richer nodes get richer*”, and hence hub dominance values of scale-free networks are usually high. On the other hand, the associated clustering coefficients are usually low and are decayed quickly in function of network sizes [15, 24], which means low transivities. Consequently, typical scale-free networks have a close structure with that of star-based topologies as depicted in Figure 3(c). The location of scale-free networks in function of two dimensions: transitivity and hub dominance is illustrated in Figure 4 in the top left-hand corner which associates to low CCF and high  $\text{hub\_dom}$  scores.

## VII. IDENTIFYING COMMUNITY PROFILE OF DIFFERENT NETWORK CATEGORIES

In this section, we show empirical evidences to associate structural communities in real world networks with corresponding topologies determined by the bivariate representation. In order to do that, first CCF and  $\text{hub\_dom}$  quality scores are calculated over the whole set

of communities detected on the network dataset by the presented algorithms. Later, these communities are located in the characterized space in function of their couples of values (CCF, hub\_dom) which represent transitivity and hub dominance respectively. The distribution of communities on this two dimensional space helps to match the most corresponding topologies with each set of communities thanks to the topology characterization presented in the previous section. Since it has been noticed that some structural characteristics might differ between small communities called *micro-communities* and large communities called *macro-communities* [27], we proceed to analyze them separately. Figure 5 delineates the distributions of small communities of 10 nodes or less in 6 different network groups including communication, technological, information, biological, social and miscellaneous networks as described in Table II. The homologous distributions for large communities of more than 10 nodes are depicted in Figure 6.

At a first sight, it is easy to remark that there is a much higher diversity of structures at the large scale communities than at the small scale communities as the distributions are much more expanded over the space in the former case. It is reasonable since there are much more possibilities how nodes can be connected in a large community than in a small one. Hence large communities' structures are more distinctive and at the same time more complex. Specifically, most of small communities are found around two axis where  $CCF = 0$  or  $hub\_dom = 1$ , especially at their crosspoint where  $CCF = 0$  and  $hub\_dom = 1$ . It means star-based and hub dominated structures are very well representative for small communities of every network category. On the other hand, grid structure is totally absent at this size scale, which is quite predictable since it requires a large number of nodes for a grid to be formed. Additionally, the heavy-tail degree distribution recognized in many real world networks make grids less likely to be established.

In information and miscellaneous groups, communities are much more rich in structure comparing to the other categories at both scales. Concretely, besides star-like modules, there are also many clique-like communities and mixture structures since clustering coefficient values in these groups stretch across the whole range. Similarly for hub dominance values which are measured approximately from 0.4 to 1 at the small scale and from 0 to 1 at the large scale. Although there are some differences in community structure between various network categories, at a small scale, it not very obvious to distinguish them using the proposed representation. We introduce in the following part a detail inspection, especially for large communities, which would reveal essential distinctions between community structure of each network category. The distribution of communities over the profiled map characterizes the mesoscopic structural identity of networks.

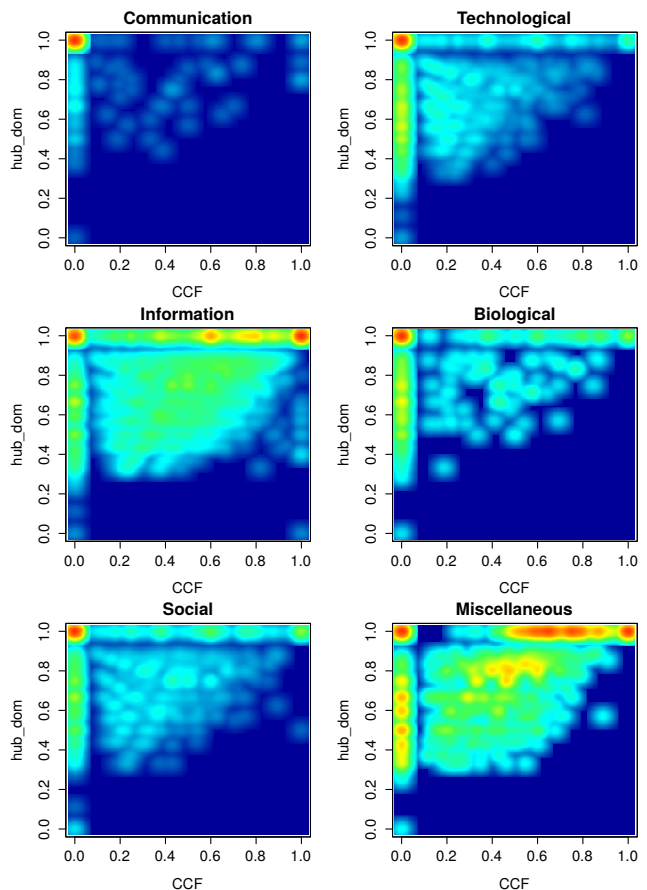


FIG. 5. Heat maps of distributions of small structural communities detected on different categories of networks are presented on a two dimensional space characterized by transitivity (CCF) and hub dominance (hub\_dom). Only communities of **10 nodes or less** are included. From left to right, top to bottom (a) Communication, (b) Technological, (c) Information, (d) Biological, (e) Social, (f) Miscellaneous consists in power networks, ecological networks, artificial networks, etc.

### A. Communication networks

Communication communities consist in subnetworks of message exchange in social networks, email communications, discussions in forums, etc. From the bivariate distributions of communities shown in Figure 6(a) and 5(a), it can be recognized that structural communities are quite homogeneous in terms of topology in both large and small communities. The majority of them have star-based topologies with very strong hubs which connect to almost every other node in their communities and very few number of clique connections. In other words, communication communities are in general very remarkably high centralized and very low transitive. This property is less clear in large communities than in small communities since the larger a community, the more likely non-hub nodes have chances to create interconnections and possibly establish peripheral hubs. This mechanism also gives rise to a few numbers of multi-hub

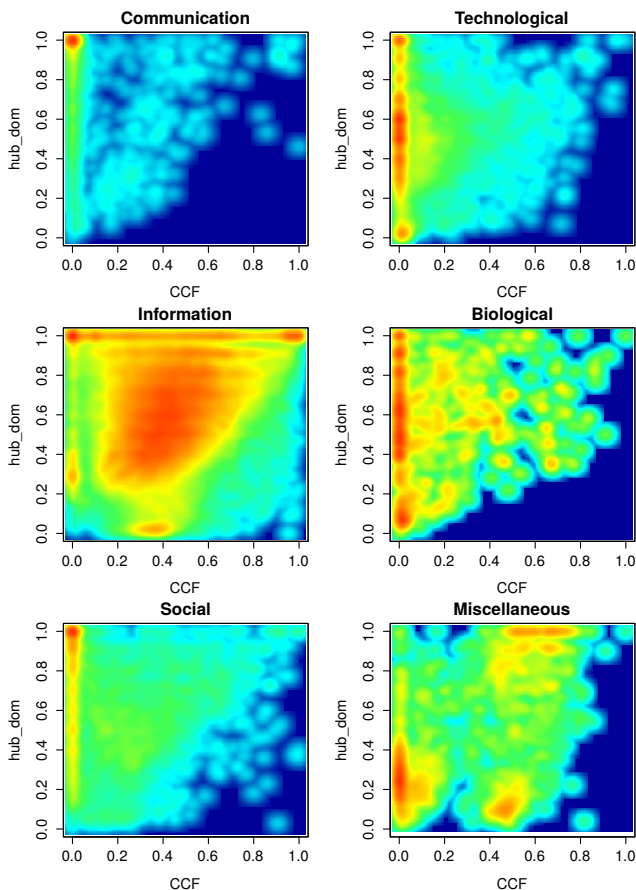


FIG. 6. Heat maps of distributions of large structural communities detected on different categories of networks are presented on a two dimensional space characterized by transitivity (CCF) and hub dominance (hub\_dom). Only communities of **more than 10 nodes** are included. From left to right, top to bottom (a) Communication, (b) Technological, (c) Information, (d) Biological, (e) Social, (f) Miscellaneous consists in power networks, ecological networks, artificial networks, etc.

topologies in large communities. Besides, a small number of hub-absent communities and mesh communities can be discerned. However, they are quite outnumbered by hub structures in this network category. This revelation denotes that exchanges in communication networks often happen around some *central elements* which convey access to their surrounding elements. Figure 7 illustrates some typical structural community topologies that have been identified in the communication network dataset. Among them, star-like topologies with one dominating hub as shown in Figure 7(a),(b) are among the most representative. Besides, there are also communities where hubs are less influential and the presence of a few cliques can be recognized as illustrated in Figure 7(c),(d). However, within the list of network categories that has been analyzed in this study, communication communities show a clearest and strongest hub-periphery connection pattern with more than 80% of communities where there are at least 1 node connected to at least 90% of

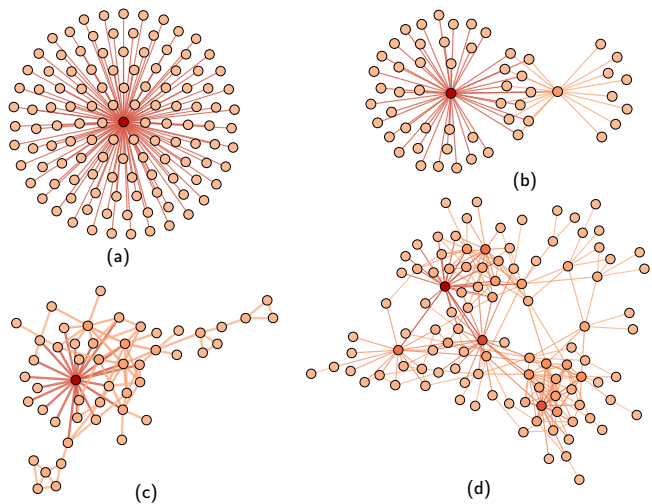


FIG. 7. Some representative topologies detected in *Communication networks* with their corresponding scores (CCF, hub\_dom). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 6(a), 5(a). Hub nodes are darker than peripheral nodes. (a) Email traffic in an European research institution [41] community - (0, 1); (b) Wikipedia adminship vote [29] community - (0.03, 0.87); (c) Email communication Enron network - (0.07, 0.90); (d) Community of email exchange in an university - (0.28, 0.23).

node members in its community and very few periphery-periphery connections. By consequence, communication communities are commonly quite sparse in comparison to other types of networks. Moreover, previous study demonstrated in Figure 4 helps to infer that communities networks reveal strong *scale-free* property. Consequently, a preferential attachment mechanism with an amplified connection probability to hub nodes would efficiently mimic the structure of real world communication networks.

## B. Technological networks

Technological communities include subnetworks in peer-to-peer Gnutella file sharing networks, Internet, highway and airport circulation systems, etc. The most notable similarity between technological communities and communication communities is the high presence of hub-based topologies, especially in small communities as can be seen in Figure 5(b). In large communities, however, technological communities show a quite discernible connection pattern as hubs are less *powerful* in their local as can be interpreted from Figure 6(b). Quantitatively, the majority of hubs in technological networks embrace around 40% to 60% of nodes in their communities. Additionally, the withdraw of super dominating hubs is replaced by the occurrence of more triadic connections in technological communities. It can be explained by the fact that in some infrastructure networks such as

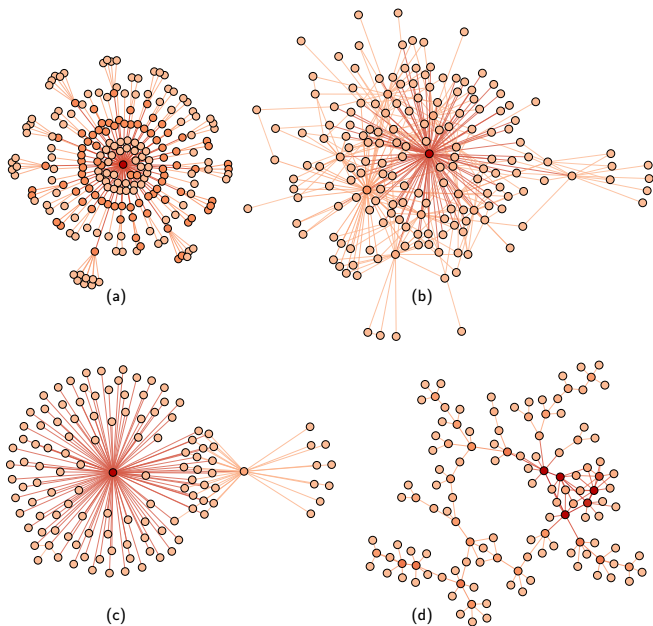


FIG. 8. Some representative topologies detected in *Technological networks* with their corresponding scores ( $CCF, hub\_dom$ ). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 5(b), 6(b). Hub nodes are darker than peripheral nodes. (a) A community of users of the Pretty-Good-Privacy algorithm for secure information interchange - (0.01, 0.48); (b) WHOIS Internet IP community - (0.07, 0.65); (c) A community of AS Caida Internet infrastructure recorded in 2007 - (0.01, 0.92); (d) A Gnutella peer-to peer network community - (0.01, 0.07).

highway networks or the Internet, hubs are often constructed to have a controlled influence and are normally compensated by resilient connections or supplement hubs in order to reduce workload, vulnerability or crucial impact caused by their dysfunctionality. Figure 8 illustrates some community topologies that have been identified in the technological network dataset. Topologies whose hubs connect to around a half of node members as depicted in Figure 8(a),(b) are among the most representative of networks in this class. There is usually a stratification in the connection pattern as many nodes are connected to a central node by intermediate nodes. This phenomenon can be considered as a presence of hierarchical organization frequently found in technological systems. Besides, there is also a considerable number of star-based structures such as those of communication case and string-based structures as shown in Figure 8(c) and 8(d) respectively. In a general view, the scale free property is quite clear although hub attractiveness is relatively reduced comparing to communication networks. A preferentially attachment fitness provided by a model such as Barabási-Albert would allow to imitate well technological structural networks.

### C. Information networks

Information communities contain subnetworks in citation networks, scientific collaboration networks, research engine networks, recommendation networks, etc. Within the studied networks, information networks exhibit the most diverse topological pattern with the bivariate distribution of communities expanded over a wide range of hub dominance axis and transitivity axis as shown in Figure 5(c), 6(c). Globally, information communities are different from communities of the other network categories by their high transitivity. Such that cliques are very well presented in many information networks as depicted in Figure 9. Many information communities can be considered as mixtures of different basic topologies of star-based, string-based, clique-based and grid-based such as the community of collaboration in Arxiv Condensed Matter network shown in Figure 9(h). The presence of hubs in information networks is still high, however they are not anymore the only elements who connect different members of networks. Consequently, information networks are normally much more dense and well connected than other types of networks of the same size scale. This is probably the most representative connectivity feature of information networks. Similar results related to dense and clique structures have been also found by Lancichinetti *et al.* [27]. Figure 9(a-h) depict some representative communities that have been discovered in some information networks. While the structure in Figure 9(d) resembles a star-based topology with a sequence of periphery-periphery connections; the one in Figure 9(e) of Arxiv High Energy Physics collaboration looks like a complete network with some ill-connected nodes. Figure 9(c,g) demonstrating web and recommendation systems reveal a mixture structure where hubs can be well recognized and clique presence is also remarkable at the same time. The hybrid structure is globally more blended in communities of Figure 9(a,b,h) than the others. The diversity in the structure of information networks can be explained by the way we define this category. In fact, a commercial recommendation system could be very unlike a web citation or a collaboration network, even though they are all considered to be information systems in the network science community. Furthermore, their structures are normally exposed to several complex phenomena that regulate network interactions. Hence, simulating information networks merits more investigation on each concrete case to determine principle mechanism that reflects well the mesoscopic organization.

### D. Biological networks

Biological communities comprise subnetworks in brain networks, yeast networks, protein-protein interaction networks, metabolic reaction networks, etc. In some ways, their topologies resemble with technological net-



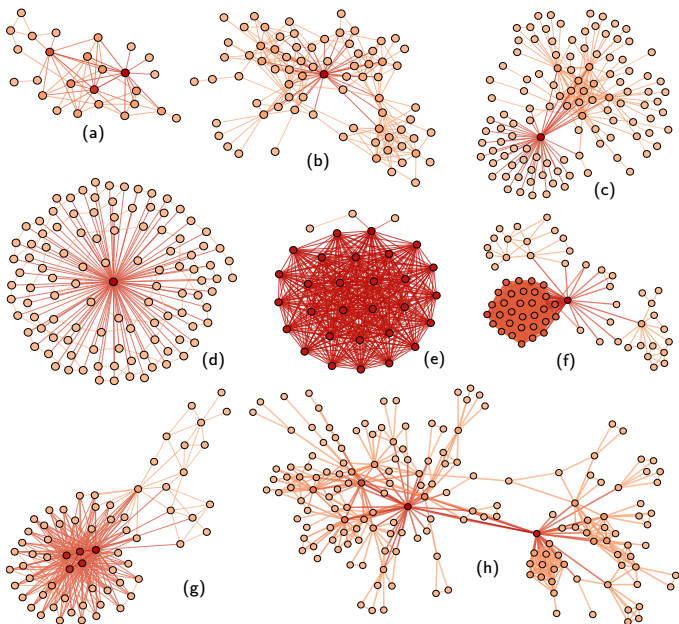


FIG. 9. Some representative topologies detected in *Information networks* with their corresponding scores ( $CCF, hub\_dom$ ). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 5(c), 6(c). Hub nodes are darker than peripheral nodes. (a,b,g) Amazon recommendation groups of products - (0.40, 0.52), (0.33, 0.45) and (0.24, 0.76) respectively; (c) An educational web system cluster - (0.30, 0.43); (d) A group of Indochina websites recorded in 2004 - (0.05, 0.98); (e-f) A community of Arxiv High Energy Physics collaboration - (0.99, 0.97) and (0.95, 0.99); (h) A collaboration community of Arxiv Condensed Matter network - (0.44, 0.36).

works as it can be observed through their distributions in Figure 6(b) and 6(d). The most remarkable discrimination of connection pattern between biological networks with the other ones are their string-based rich structure as can be seen through communities shown in Figure 10(a), (b), (c). The high presence of chains or strings in biological networks has been also found by the other studies using different approaches such as in [27], [20]. This may be caused by the fact that many biological pathways, which are series of molecular interactions, are included in the analysis and contribute to the high presence of strings. Additionally, many biological networks are only constructed partially due to high complexity in construction time and technical constraints in biochemistry [34]. Therefore, we often observe and analyze small fragments of networks where many connections are missing.

Still, there exist biological networks whose topologies are star-based or hybrid as those of communication networks, technological networks or information networks. However, the hub dominance is globally less important as biological communities are normally small and hubs connect to much less number of their surrounding neighbors. A local observation on biological networks probably dis-

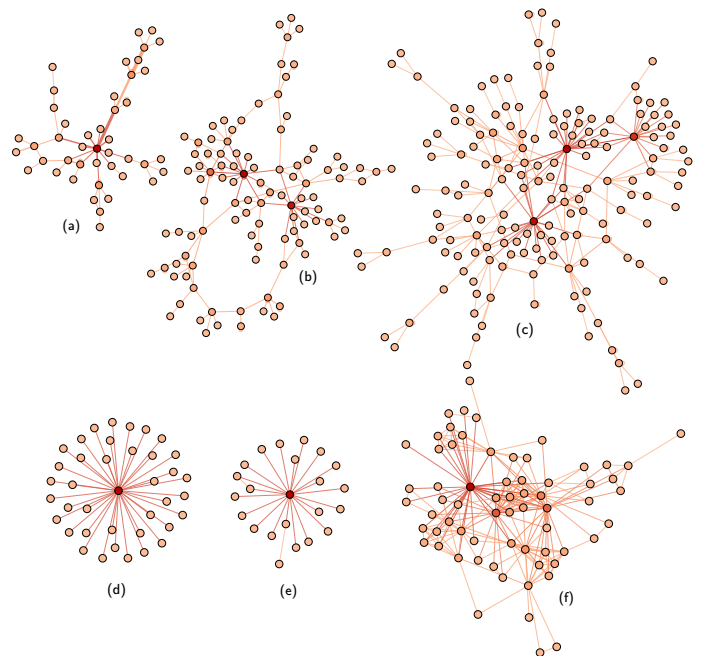


FIG. 10. Some representative topologies detected in *Biological networks* with their corresponding scores ( $CCF, hub\_dom$ ). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 5(d), 6(d). Hub nodes are darker than peripheral nodes. (a) A circuit of medulla of drosophila fly brain - (0.06, 0.44); (b-c) A protein-protein interaction network of yeast - (0.03, 0.16) and (0.05, 0.16) respectively; (d-e) protein interactions of drosophila melanogaster (0, 1) and (0.01, 0.95); (f) A cluster of human disease network (0.47, 0.51).

closes random structures in many parts of the networks although hubs are still well widespread. This emergence of random structures could be the most typical characteristic that differs biological networks from the others. Finally, popular properties such as *scale-free*, *small-world* are less significant in biological class than in information or technological class.

### E. Social networks

Social communities involve subnetworks of friendship networks, share or re-tweet networks, followings in Google Plus, Facebook, Twitter, Youtube, etc. Our analysis shows a high similarity in the distribution of large communities in the social networks and communication networks as depicted by Figure 6(a), Figure 6(e). For small communities, social networks are closer to technological networks and biological networks as shown in Figure 5(b), 5(d), 5(e). A reasonable explanation for the popularity of the star-based topology in social network is that there are many well-known users who are followed or subscribed by a large number of peoples and are becoming mega-connected hubs. Additionally, many samples of social networks that are studied consist of ego networks

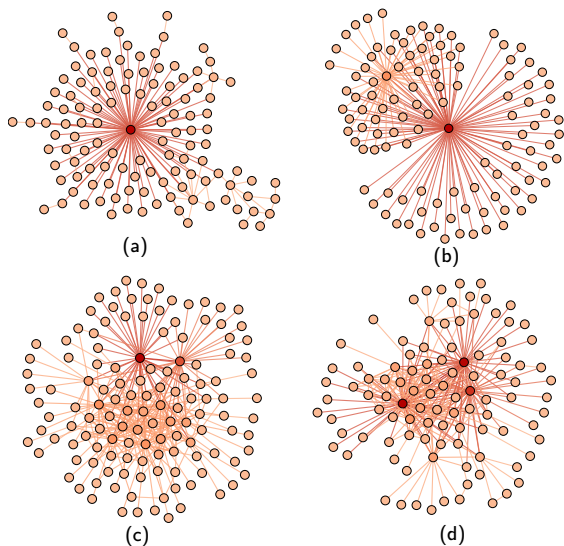


FIG. 11. Some representative topologies detected in *Social networks* with their corresponding scores ( $CCF$ ,  $hub\_dom$ ). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 5(e), 6(e). Hub nodes are darker than peripheral nodes. (a) A structural community in Youtube video sharing friendship network - (0.01, 0.81); (b) A community in Google Plus network - (0.02, 0.95); (c) A political re-tweet network in Twitter - (0.12, 0.60); (d) A subnetwork of location-based social networking Brightkite - (0.27, 0.51).

of celebrities in social media, which makes them intrinsically high centralized around some mega-hub nodes. The only difference with communication communities that has been found in this study is that there are generally more connections between peripheral nodes in social communities. This can be interpreted that friendship or following interactions are generally more frequent than communication interactions. Although different networks of social and communication have been used in this analysis, it makes sense to explain that many users are connected in a social media without or very few communicating interactions in the same channel. For example, two users could be connected on Facebook as friends, but they never exchange any message on the Facebook conversation platform which makes the number of social connections exceeds the number of communications. Figure 11 demonstrates some popular topologies of communities in social networks. Note that these topologies are not chosen to argue the differences between various social networks and it is not the objective of this study. They are listed to illustrate some typical and representative structural communities that we discover in the network dataset. Social networks show a clear *scale-free* property as in communication and technological networks, however they are less affected by mega-hubs and are partially occupied by clique-based structures and many random connections like that of *small-world* phenomenon.

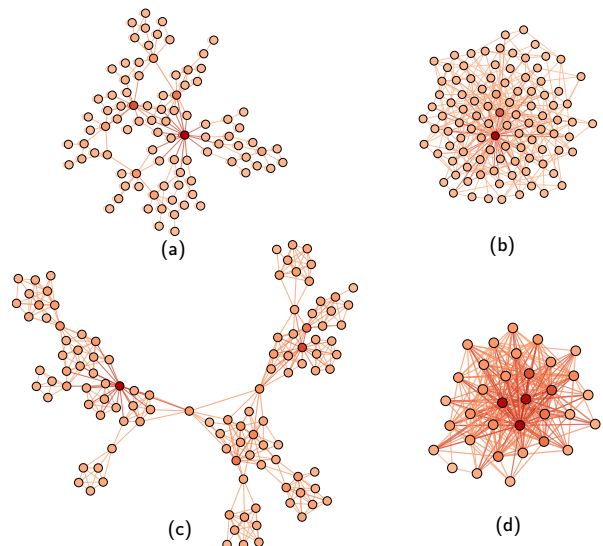


FIG. 12. Some representative topologies detected in miscellaneous group with their corresponding scores ( $CCF$ ,  $hub\_dom$ ). Topologies are ordered from the most famous to the less famous in their network category as shown in Figure 5(f), 6(f). Hub nodes are darker than peripheral nodes. (a) A cluster of a power network system - (0.07, 0.21); (b) A quadratic sieve of a factorization of a 130 bit number - (0.08, 0.39); (c) A cluster of a Lancichinetti-Fortunato-Radicchi (LFR) synthetic network [26] - (0.56, 0.18); (d) A cluster in an ecological network - (0.51, 0.94).

## F. Miscellaneous networks

Miscellaneous communities cover subnetworks in ecological networks, some power system networks, sport competition networks, synthetic networks, etc. Here, we find many structures, especially in Lancichinetti-Fortunato-Radicchi (LFR) synthetic networks [26], that are not very popular in the previously studied networks. Specifically, except for information networks, structural communities in the other types of networks are usually very hub-centralized and relatively low in transitivity. On the contrary, in LFR networks, cliques are quite popular and normally aggregated to produce compacted structures as illustrated in Figure 12(c), which makes the communities highly transitive. Additionally, although structures of LFR networks are regulated by many configuration parameters, their hubs generally have less impact in their neighborhoods than those of real world networks such as in social or communication. This is one property that makes a huge difference between LFR benchmarking networks and real world networks. Some other discovered structural communities are illustrated in Figure 12. In a general view, community detection methods identified well compacted sub-graphs in most of the cases.

## VIII. DISCUSSION AND PERSPECTIVES

As a whole, we provide a novel analysis process to categorize mesoscopic organization of networks into four essential topological groups which show different node organizations. Each representative group is then associated to the corresponding graph generative model that produces a high similarity in connection patterns. Surprisingly, our empirical study uncovers that networks across different categories including communication, technological, information, biological and social networks might have different community structures and can be described by distinguishable characterized topologies.

The difference of modular topology between networks in various categories could help to construct network profiles or network signatures by domain of study, and hence open a possibility for creating adapted network generative models, network class prediction algorithms, dynamical processes simulation and analysis, etc. Specifically, since networks in each domain reveal some particular modular structures, the mechanisms which are responsible for their creations, evolutions, degradations are also discernible. Hence, different simulation or analysis strategies will generate different impacts on the networks in a predictable way if their structures are well understood. In other words, the network structure profiling assists to achieve suitable network analysis processes and

to interpret obtained results without requiring expensive brute force analysis.

We have also some network categories (e.g. information networks) which have more diverse topological patterns, which means the bivariate distribution of network subcategories might differ from each other, even though they belong to the same category. Describing how big the difference between subcategories in the same category is would be an interesting perspective. It would require a dedicated in-depth analysis, involving the test of several types of bivariate maps, and probably resort to business experts to better understand the particularities.

In this study, we include many state-of-the-art community detection methods whose approaches are quite distinct in order to exploit different facets of modular structure that could be detected on the networks. Nevertheless, the impact of these methods on the revealed structures merits to be examined in more details. An interesting perspective could be the relation between different topologies and different mechanisms that are responsible for identifying or transforming networks/subnetworks from one type to another. Understanding these mechanisms could help us to explain the effects of community detection algorithms on the partitioning of networks and also how different dynamical processes influence evolving networks.

- 
- [1] Albert, R.: Scale-free networks in cell biology. *Journal of Cell Science* **118**(21), 4947–4957 (2005). doi:10.1242/jcs.02714
- [2] Alm, E., Arkin, A.P.: Biological networks. *Current Opinion in Structural Biology* **13**(2), 193 – 202 (2003). doi:https://doi.org/10.1016/S0959-440X(03)00031-9
- [3] Azaouzi, M., Rhouma, D., Romdhane, L.B.: Community detection in large-scale social networks: state-of-the-art and future directions. *Social Network Analysis and Mining* **9**(1), 23 (2019)
- [4] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999). doi:10.1126/science.286.5439.509
- [5] Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences* **101**(11), 3747–3752 (2004). doi:10.1073/pnas.0400087101
- [6] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008 (2008). doi:10.1088/1742-5468/2008/10/p10008
- [7] Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004)
- [8] Coscia, M., Giannotti, F., Pedreschi, D.: A Classification for Community Discovery Methods in Complex Networks. *Statistical Analysis and Data Mining* (2012)
- [9] Dao, V.L., Bothorel, C., Lenca, P.: Community structures evaluation in complex networks: A descriptive approach. 3rd International Winter School and Conference on Network Science pp. 11–19 (2017). doi:10.1007/978-3-319-55471-6\_2
- [10] Dao, V.L., Bothorel, C., Lenca, P.: Community structure: A comparative evaluation of community detection methods. *Network Science* **8**(1), 1–41 (2020). doi:10.1017/nws.2019.59
- [11] Erdős, P., Rényi, A.: On random graphs, I. *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959)
- [12] Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.* **29**(4), 251–262 (1999). doi:10.1145/316194.316229
- [13] Fortunato, S.: Community detection in graphs. *Physics Reports* **486**(3-5), 75–174 (2010). doi:10.1016/j.physrep.2009.11.002
- [14] Fortunato, S., Hric, D.: Community detection in networks: A user guide. *Physics Reports* **659**, 1–44 (2016). doi:10.1016/j.physrep.2016.09.002
- [15] Fronczak, A., Fronczak, P., Hołyst, J.A.: Mean-field theory for clustering coefficients in barabási-albert networks. *Physical Review E* **68**(4) (2003). doi:10.1103/physreve.68.046126
- [16] Ghalmane, Z., El Hassouni, M., Cherifi, H.: Immunization of networks with non-overlapping community structure. *Social Network Analysis and Mining* **9**(1), 45 (2019)
- [17] Ghasemian, A., Hosseinmardi, H., Clauset, A.: Evaluating overfit and underfit in models of network community structure. *ArXiv e-prints* (2018)

- [18] Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**(12), 7821–7826 (2002). doi:10.1073/pnas.122653799
- [19] Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* **70**(2), 025101 (2004). doi:10.1103/PhysRevE.70.025101
- [20] Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Classes of complex networks defined by role-to-role connectivity profiles. *ArXiv Physics e-prints* (2007)
- [21] Jebabli, M., Cherifi, H., Cherifi, C., Hamouda, A.: Overlapping community detection versus ground-truth in amazon co-purchasing network. In: 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 328–336. IEEE (2015)
- [22] Jebabli, M., Cherifi, H., Cherifi, C., Hamouda, A.: Community detection algorithm evaluation with ground-truth data. *Physica A: Statistical Mechanics and its Applications* **492**, 651–706 (2018). doi:https://doi.org/10.1016/j.physa.2017.10.018
- [23] Jerome, K.: The koblenz network collection. In: *Proceedings Conference on World Wide Web Companion*, pp. 1343–1350 (2013). URL <http://konect.uni-koblenz.de>
- [24] Klemm, K., Eguíluz, V.M.: Growing scale-free networks with small-world behavior. *Physical Review E* **65**(5) (2002). doi:10.1103/physreve.65.057102
- [25] Labatut, V., Orman, G.K.: Community structure characterization. In: *Encyclopedia of Social Network Analysis and Mining*, pp. 1–13. Springer New York (2017). doi:10.1007/978-1-4614-7163-9\_110151-1
- [26] Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**(4), 046110 (2008). doi:10.1103/PhysRevE.78.046110
- [27] Lancichinetti, A., Kivela, M., Saramaki, J., Fortunato, S.: Characterizing the community structure of complex networks. *PLoS ONE* **5**(8), e11976 (2010). doi:10.1371/journal.pone.0011976
- [28] Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, pp. 462–470. ACM, New York, NY, USA (2008). doi:10.1145/1401890.1401948
- [29] Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection (2014). URL <http://snap.stanford.edu/data>
- [30] Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. *Proceeding of the 17th international conference on World Wide Web - WWW 08* (2008). doi:10.1145/1367497.1367591
- [31] Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical Comparison of Algorithms for Network Community Detection. *ACM WWW International Conference on World Wide Web* (2010)
- [32] Newman, M.E.J.: The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98**(2), 404–409 (2001). doi:10.1073/pnas.98.2.404
- [33] Newman, M.E.J.: Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**(23), 8577–8582 (2006). doi:10.1073/pnas.0601602103
- [34] Newman, M.E.J.: *Networks: An Introduction*. Oxford University Press (2010). doi:10.1093/acprof:oso/9780199206650.001.0001
- [35] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Physical Review E* **69**(2) (2004). doi:10.1103/physreve.69.026113
- [36] <https://www.ning.com/>
- [37] Orman, G.K., Labatut, V., Cherifi, H.: Comparative evaluation of community detection algorithms: a topological approach. *Journal of Statistical Mechanics: Theory and Experiment* **2012**(08), P08001 (2012). doi:10.1088/1742-5468/2012/08/p08001
- [38] Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: p. Yolum, T. Güngör, F. Gürgeç, C. Özturan (eds.) *Computer and Information Sciences - ISCIS 2005*, pp. 284–293. Springer Berlin Heidelberg (2005)
- [39] Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* **76**(3) (2007). doi:10.1103/physreve.76.036106
- [40] Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. *Physical Review E* **74**(1) (2006). doi:10.1103/physreve.74.016110
- [41] Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015). URL <http://networkrepository.com>
- [42] Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105**(4), 1118–1123 (2008). doi:10.1073/pnas.0706851105
- [43] Rosvall, M., Delvenne, J.C., Schaub, M.T., Lambiotte, R.: Different approaches to community detection. *ArXiv e-prints* (2017)
- [44] Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-scale topological and dynamical properties of the internet. *Phys. Rev. E* **65**, 066130 (2002). doi:10.1103/PhysRevE.65.066130
- [45] Wang, W., Street, W.N.: Modeling influence diffusion to uncover influence centrality and community structure in social networks. *Social Network Analysis and Mining* **5**(1), 15 (2015)
- [46] Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440–442 (1998). doi:10.1038/30918
- [47] Yang, J., Leskovec, J.: Defining and Evaluating Network Communities based on Ground-truth. *Knowledge and Information Systems* **42** (2015)
- [48] Zhu, X., Gerstein, M., Snyder, M.: Getting connected: analysis and principles of biological networks. *Genes and development* **21**(9), 1010–1024 (2007)