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

Vinh-Loc Dao1 · Cécile Bothorel1 · Philippe Lenca1

Received: 12 June 2018 / Revised: 1 March 2021 / Accepted: 16 March 2021 / Published online: 10 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract
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

1 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 (Leskovec et al. 2008; Yang and Leskovec 2015), biochemical networks, (Alm and Arkin 2003; Albert 2005; Zhu et al. 2007), computer networks (Faloutsos et al. 1999; Vázquez et al. 2002), 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 (Raghavan et al. 2007; Girvan and Newman 2002; Newman and Girvan 2004; Rosvall and Bergstrom 2008). Several efforts have been also dedicated to evaluate and categorize these methods in a systematic way (Leskovec et al. 2010; Fortunato 2010; Coscia et al. 2012; Rosvall et al. 2017).

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 (Fortunato 2010), community quality evaluation has been always a 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.

Cécile Bothorel
cecile.bothorel@imt-atlantique.fr

Philippe Lenca
philippe.lenca@imt-atlantique.fr

1 Lab-STICCC, CNRS UMR, 6285, IMT Atlantique, F-29238 Brest, France
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 (Fortunato and Hric 2016). 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 (Ghalmane et al. 2019). The expected structure may exist 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 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 is presented in Sect. 2 to give an overview about similar existing researches; then, quality metrics that are used to characterize structural communities will be introduced in Sect. 3. Then, Sect. 4 describes community detection methods employed to identify and validate communities; subsequently, the network dataset and their uncovered hidden communities are analyzed in Sect. 5 using quality metrics presented in Sect. 3. We then present a conceptualization of community structure using a bivariate representation approach which is introduced in Sect. 6, and then, based on this approach, we identify in Sect. 7 different structural profiles of communities in various network categories such as communication, technological, information, biological, social, ecological and synthetic networks. Finally, Sect. 8 draws some conclusions and envisages some potential perspectives for future work.

## 2 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 (Fortunato and Hric 2016; Ghase minibian et al. 2018; Jebabli ewt al. 2015; Jebabli et al. 2018), and (Azaouzi et al. 2019) 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 (Dao et al. 2020), 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 and max internal degree, in large-scale networks according to discovered community size (Lancichinetti et al. 2010). 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 (Guimera et al. 2007). Specifically, by using two metrics including within-module degree z and participation ratio P (Guimerà et al. 2004), 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 (Leskovec et al. 2008). 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 (Coscia et al. 2012). The authors then resume popular methods in the literature according to different quality aspects such as density-based, vertex similarity-based, action-based or influence propagation-based [for example, Wang and Street 2015 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 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.

3 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 (Yang and Leskovec 2015), modularity (Newman and Girvan 2004) or description length (Rosvall and Bergstrom 2008)] even though their efficacy cannot 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.

3.1 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 $d_{in}(u)$. As such, the external degree of node $u$ can be calculated as $d_{ext}(u) = d(u) - d_{in}(u)$. The relation between internal and external degrees of nodes in $S$ with its number of edges can be resumed as: $\sum_{u \in S} d_{in}(u) = 2m_S$ 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:
3.2 Metrics based on internal edge density

- **Density**
  \[
  \text{density}(S) = \frac{m_S}{n_S(n_S-1)/2}
  \]  
  (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**
  \[
  \text{sc\_den}(S) = \frac{2m_S}{n_S-1}
  \]  
  (2)

  Scaled density is a kind of normalized density, which is defined as \( n_S \) times the density of the community (Lancichinetti et al. 2010; Labatut and Orman 2017). 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 increases 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.

3.3 Metrics based on centralized or hub structure

- **Hub dominance**
  \[
  \text{hub\_dom}(S) = \frac{\max_{u \in S} d_{\text{int}}(u)}{n_S-1}
  \]  
  (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 (Lancichinetti et al. 2010; Labatut and Orman 2017).

3.4 Metrics based on triadic structure

- **Clustering coefficient (CCF)**
  \[
  \text{CCF}(S) = \frac{3 \Delta_S}{T_S}
  \]  
  (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 three 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 (Barrat et al. 2004).

- **Triangle participation (TPR)**
  \[
  \text{TPR}(S) = \frac{\sum_{u \in S} \delta_{us}}{n_S}
  \]  
  (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 \( S \) (Yang and Leskovec 2015). 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).

3.5 Metrics based on external connectivity

- **Expansion:**
  \[
  \text{expansion}(S) = \frac{c_j}{n_j}
  \]  
  (6)

  The metric measures the number of edges per node that point out side a cluster (Yang and Leskovec 2015). 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**
  \[
  \text{conductance}(S) = \frac{c_j}{2m_S + c_S}
  \]  
  (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 identify good network community profile (Leskovec et al. (2008)).
The out degree fraction of node $u$ in subgraph $S$ is defined as:

$$\text{ODF}_S(u) = \frac{\sum_{v \in S} |\{(u, v) \in E : v \notin S\}|}{d(u)}$$

with $\text{meanODF}(S) = \frac{1}{n_S} \sum_{u \in S} \text{ODF}_S(u)$ 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 (Yang and Leskovec 2015).

### Maximum Out Degree Fraction

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

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 (Yang and Leskovec 2015).

### 4 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 (Girvan and Newman 2002) is based on edge centrality detection in order to break networks into several communities; the Louvain method (Blondel et al. 2008) optimizes local modularity by iteratively folding nodes into meta-nodes; the label propagation (Raghavan et al. 2007) determines the community of a node by considering the memberships of its neighbors; the Infomap method (Rosvall and Bergstrom 2008) 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 1.

### 4.1 Edge betweenness community detection

Girvan and Newman proposed a method (Girvan and Newman 2002) 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.

### 4.2 Fast greedy method

This method discovers 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 $O(nm \log n)$ and can be reduced to $O(n \log^2 n)$ in sparse networks where $n$ corresponds to the number of vertices of the network (Clauset et al. 2004). 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 $j$.

### 4.3 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 (Blondel et al. 2008) 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 in 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.

### 4.4 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 $P_i$ and $P_j$, to go from $i$ and $j$ to other vertices in a short number of $t$ steps: $r_{ij}^2 = \sum_{k=1}^{n} \frac{(P_{i,k} - P_{j,k})^2}{d(k)}$ (Pons and Latapy 2005). Then, a traditional hierarchical clustering algorithm is used to find communities from this dynamic distance.

### 4.5 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 (Raghavan et al. 2007). 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.

### 4.6 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 (Rosvall and Bergstrom 2008). 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, 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.

### 4.7 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 \( B \) (Newman 2006). 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, the author finds out that the leading eigenvectors that correspond to the positive eigenvalues of the modularity matrix help to find a good partition that maximize the modularity.

### 4.8 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 (Reichardt and Bornholdt 2006).

### 5 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 with 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 with 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 six networks to evaluate the structure of communities discovered by several detection techniques (Orman et al. 2012); Lancichinetti et al. use 15 networks to characterize structural communities (Lancichinetti et al. 2010); 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 (Leskovec et al. 2008) 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 \(^1\) are aggregated (Yang and Leskovec 2015). Table 2 resumes the composition of networks that have been analyzed in this paper.

| 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 2 A summary of network dataset used in this analysis where Size is the number of networks analyzed in each category, Nodes and Edges indicate 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 (Rossi and Ahmed 2015), http://konect.uni-koblenz.de (Jerome 2013), http://snap.stanford.edu (Leskovec and Krevl 2014).

---

\(^1\) https://www.ning.com/
5.2 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 1 on the dataset collection presented in Table 2. 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 experimentations.

Once communities are produced, the quality metrics presented in Sect. 3 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 Sect. 4 over 5 classes of networks and the whole dataset.

[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.]

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 Fig. 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 Fig. 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 belongs 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 two groups of metrics could be reduced 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 meanwhile, 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 3 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.

| 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 3 Groups of quality metrics that reflect different aspects of community structural property

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.
information, 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 Dao et al. (2017) where communities are portrayed by two variables: the level of external interaction and the distribution of these interactions over community border nodes.

6 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 (Dao et al. 2017) 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-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.

6.1 Which metrics fit?

It is well known that density has 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 with 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 \( \langle k \rangle = \frac{2m}{n} \). 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 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 Fig. 1d). 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 reflects 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 makes 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.

6.2 Locating community structures in a bidimensional 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 four fundamental coordinated zones corresponding to 4 underlying topologies, which are emphasized in Table 4. These classes of topologies could be explained as follows:
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 Fig. 3a depending on the context.

- **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 Fig. 3a 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 implies 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 Fig. 3b.

- **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 Fig. 3c.

- **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 Fig. 3d.

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.

### 6.3 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** Erdős and Rényi (1959) is among the first models proposed to describe the generation of random graphs. In this model, 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 are 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 (Newman 2001). 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 Fig. 3a. In an extreme regime, when the probability of connection \( p \) approaches 1, the associated random graph becomes nearly complete as the average degree \( < k > \) approaches \( n - 1 \), which means every vertex connects with almost every other vertex as illustrated in Fig. 3d. 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 corner 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 (Watts and Strogatz 1998). 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 probabil-
ity \(0 < p < 1\). The authors find that a small value of \(p\) reduces 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 is very small, the clustering coefficient stays 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 Fig. 3b. The location of its topology in function of two dimensions: transitivity and hub dominance are illustrated in Fig. 4 in the bottom right-hand corner which associates to high CCF scores and low hub dom scores.

- Barabási–Albert (BA) model Barabási and Albert (1999) is originating from a discovery that the distribution of vertex degrees in many real-world networks, such as genetic networks and World Wide Web networks, is 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 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 (Klemm and Eguíluz 2002; Fronczak et al. 2003), which means low transitivities. Consequently, typical scale-free networks have a close structure with that of star-based topologies as depicted in Fig. 3c. The location of scale-free networks in function of two dimensions: transitivity and hub dominance, is illustrated in Fig. 4 in the top left-hand corner, which associates to low CCF and high hub dom scores.

7 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 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 (Lancichinetti et al. 2010), 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 2.

The homologous distributions for large communities of more than 10 nodes are depicted in Fig. 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 is 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.

### 7.1 Communication networks

Communication communities consist in sub-networks of message exchange in social networks, email communications, discussions in forums, etc. From the bivariate distributions of communities shown in Figs. 6a and 5a, 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 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

---

**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.

**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.
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 Fig. 7a, 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 Fig. 7c, 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 nodes in their community and very few periphery-periphery connections. Consequently, a preferential attachment mechanism with an amplified connection probability to hub nodes would efficiently mimic the structure of real-world communication networks.

7.2 Technological networks

Technological communities include sub-networks 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 Fig. 5b. 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 Fig. 6b. Quantitatively, the majority of hubs in technological networks embrace around 40–60% 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 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 Fig. 8a, 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 Fig. 8c and 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.

### 7.3 Information networks

Information communities contain sub-networks 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 Fig. 5c, 6c. 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 Fig. 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 Fig. 9h. 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. 2010. Figure 9a–h depicts some representative communities that have been discovered in some information networks. While the structure in Fig. 9d resembles a star-based topology with a sequence of periphery–periphery connections, the one in Fig. 9e of Arxiv High Energy Physics collaboration looks like a complete network with some ill-connected nodes. Figure 9c,g demonstrating web and recommendation systems reveals 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 Fig. 9a,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.

### 7.4 Biological networks

Biological communities comprise sub-networks in brain networks, yeast networks, protein–protein interaction networks, metabolic reaction networks, etc. In some ways, their topologies resemble with technological networks as it can be observed through their distributions in Fig. 6b and d. The most remarkable discrimination of connection pattern between biological networks with the other ones is their string-based rich structure as can be seen through communities shown in Fig. 10a–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 Lancichinetti et al. (2010) and Guimera et al. 2007. 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 (Newman 2010). 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 discloses 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.

### 7.5 Social networks

Social communities involve sub-networks 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 Fig. 6a, e. For small communities, social networks are closer to technological networks and biological networks as shown in Fig. 5b, d, 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 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 exceed the number of communications. Figure 11 demonstrates some popular topologies of communities in social networks. Note that these topologies are not chosen to argument the differences between various social networks, and it is not the objective of this study. They are...
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.

### 7.6 Miscellaneous networks

Miscellaneous communities cover sub-networks 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 (Lancichinetti et al. 2008), 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 Fig. 12c, 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 Fig. 12. In a general view, community detection methods identified well compacted sub-graphs in most of the cases.

### 8 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 with 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 that 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 detail. An interesting perspective could be the relation between different topologies and different mechanisms that are responsible for identifying or

---

**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 Figs. 5f, 6f. 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 (Lancichinetti et al. 2008) - (0.56, 0.18); **(d)** A cluster in an ecological network - (0.51, 0.94)
transforming networks/sub-networks 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.

References

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(21):4947–4957. https://doi.org/10.1242/jcs.02714
Alm E, Arkin AP (2003) Biological networks. Curr Opin Struct Biol 13(2):193–202. https://doi.org/10.1016/S0959-440X(03)00031-9
Azaouzi M, Rhouma D, Roudhane LB (2019) Community detection in large-scale social networks: state-of-the-art and future directions. Soc Net Anal Min 9(1):23
Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509
Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci 101(11):3747–3752. https://doi.org/10.1073/pnas.0400087101
Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008
Clauset A, Newman M, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111
Coscia M, Giannotti F, Pedreschi D (2012) A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining
Dao VL, Bothorel C, Lenca P (2017) Community structures evaluation in complex networks: A descriptive approach. In: 3rd International Winter School and Conference on Network Science
Dao VL, Bothorel C, Lenca P (2020) Community structure: a comparative evaluation of community detection methods. Netw Sci 8(1):1–41. https://doi.org/10.1017/nws.2019.59
Erdős P, Rényi A (1959) On random graphs I. Publicationes Mathematicae. 286. 5439. 509
Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. SIGCOMM Comput Commun Rev 29(4):251–262. https://doi.org/10.1145/316194.316229
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44. https://doi.org/10.1016/j.physrep.2016.09.002
Fronczak A, Fronczak P, Holyst JA (2003) Mean-field theory for clustering coefficients in barabási-albert networks. Phys Rev E. https://doi.org/10.1103/PhysRevE.68.046126
Ghalmane Z, El Hassouni M, Cherifi H (2019) Immunization of networks with non-overlapping community structure. Soc Netw Anal Min 9(1):45
Ghaseimian A, Hosseinnardi H, Clauset A (2018) Evaluating overfit and underfit in models of network community structure. ArXiv e-prints
Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70(2):025101. https://doi.org/10.1103/PhysRevE.70.025101
Jebabli M, Cherifi H, Cherifi C, Hamouda A (2015) 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
Jebabli M, Cherifi H, Cherifi C, Hamouda A (2018) Community detection algorithm evaluation with ground-truth data. Phys A Stat Mech Appl 492:651–706. https://doi.org/10.1016/j.physa.2017.10.018
Jerome K (2013) The koblenz network collection. In: Proceedings Conference on World Wide Web Companion, pp. 1343–1350 http://konect.uni-koblenz.de
Klemm K, Eguíluz VM (2002) Growing scale-free networks with small-world behavior. Phys Rev E. https://doi.org/10.1103/physrev.e.65.057102
Labatut V, Orman GK (2017) Community structure characterization. In: Rokne J, Alhajj R (eds) Encyclopedia of social network analysis and mining. Springer, New York. pp 1–13. https://doi.org/10.1007/978-1-4614-7163-9_110151-1
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110. https://doi.org/10.1103/PhysRevE.78.046110
Lancichinetti A, Kivelä M, Saramäki J, Fortunato S (2010) Characterizing the community structure of complex networks. PLoS ONE 5(8):e11976. https://doi.org/10.1371/journal.pone.0011976
Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) 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. https://doi.org/10.1145/1401890.1401948
Leskovec J, Krevl A (2014) SNAP Datasets: Stanford large network collection (2014). http://snap.stanford.edu/data
Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Proceeding of the 17th international conference on World Wide Web - WWW 08 (2008). https://doi.org/10.1145/1367497.1367591
Leskovec J, Lang KJ, Mahoney MW (2010) Empirical Comparison of Algorithms for Network Community Detection. In: ACM WWW International Conference on World Wide Web
Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409. https://doi.org/10.1073/pnas.98.2.404
Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
Newman MEJ (2010) Networks: an introduction. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001
Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E. https://doi.org/10.1103/physreve.69.026113
Orman GK, Labatut V, Cherifi H (2012) Comparative evaluation of community detection algorithms: a topological approach. J Stat Mech Theory Exp 2012(08):P08001. https://doi.org/10.1088/1742-5468/2012/08/p08001
Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: Yolum P, Güngör T, Gürgen F, Özturcan C (eds) Computer and Information Sciences-ISCIS. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 284–293

### Springer
Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E. https://doi.org/10.1103/physreve.76.036106

Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E. https://doi.org/10.1103/physreve.74.016110

Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence http://networkrepository.com

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123. https://doi.org/10.1073/pnas.0706851105

Rosvall M, DeLvenile JC, Schaub MT, Lambiotte R (2017) Different approaches to community detection. ArXiv e-prints

Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the internet. Phys Rev E 65:066130. https://doi.org/10.1103/PhysRevE.65.066130

Wang W, Street WN (2015) Modeling influence diffusion to uncover influence centrality and community structure in social networks. Soc Netw Anal Min 5(1):15

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393(6684):440–442. https://doi.org/10.1038/30918

Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213

Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21(9):1010–1024