Detection and Interpretation of Communities in Complex Networks: Methods and Practical Application

Vincent Labatut, Jean-Michel Balasque

Detection and Interpretation of Communities in Complex Networks: Practical Methods and Application

Vincent Labatut and Jean-Michel Balasque

Summary Community detection, an important part of network analysis, has become a very popular field of research. This activity resulted in a profusion of community detection algorithms, all different in some not always clearly defined sense. This makes it very difficult to select an appropriate tool when facing the concrete task of having to identify and interpret groups of nodes, relatively to a system of interest. In this chapter, we tackle this problem in a very practical way, from the user's point of view. We first review community detection algorithms and characterize them in terms of the nature of the communities they detect. We then focus on the methodological tools one can use to analyze the obtained community structure, both in terms of topological features and nodal attributes. To be as concrete as possible, we use a real-world social network to illustrate the application of the presented tools, and give examples of interpretation of their results from a Business Science perspective.

1 Introduction

Network modeling has been used for years in many application fields: biological, social, technological, communication, information (see [1] for a very comprehensive review of applied studies). The necessity to focus on some subparts has appeared quite soon for instance in sociology [2], and was initially performed manually, with a qualitative approach. However this type of analysis changed radically during the last decades, with the coming of the information age. Technology provided scientists with means to store, access and take advantage of very large amount of data (databases, internet, computing power). The analysis of very large networks became possible, provided appropriate techniques were used. Network

Vincent Labatut
Galatasaray University, Computer Science Department, Çırağan Cad. No:36, 34357 Ortaköy/İstanbul, Turkey
e-mail: vlabatut@gsu.edu.tr

Jean-Michel Balasque
Galatasaray University, Business Science & Marketing Department, Çırağan Cad. No:36, 34357 Ortaköy/İstanbul, Turkey
e-mail: jmbalasque@gsu.edu.tr

analysis took a quantitative turn, which initiated a very creative phase, leading to the development of powerful tools.

Large real-world networks are characterized by a heterogeneous structure, which leads to particular properties. Various subfields of network analysis focus on different properties: efficiency of information propagation, robustness, stability, synchronization, etc. [1]. In particular, an heterogeneous distribution of links often leads to a so-called community structure [3]. A community roughly corresponds to a group of nodes more densely interconnected, relatively to the rest of the network [4]. Note this concept has been translated into different more formal definitions, which we will review later in this document. The way such a structure can be interpreted is obviously dependent on the modeled system. However, independently from the nature of this system, the study of communities constitutes a mesoscopic analysis, complementary to the microscopic (node-wise) and macroscopic (network-wise) approaches one can also adopt. Because of this intermediary position, the community structure conveys some very important information, necessary to the good understanding of the system [5]. Consequently, detecting communities is an essential part of modern network analysis.

In this chapter, we focus on this task with a very practical and operational approach, and adopt the user's point of view. To our opinion, someone willing to perform community detection on his data needs to answer three important questions: Which algorithms should I apply? How will I compare their results? How will I interpret the obtained communities? As stated before, networks are used in many application fields. However, modern community detection tools have not significantly penetrated certain research areas yet. We believe one of the reasons for this is the profusion of tools and the lack of information regarding their similarities and differences, which underlines the importance of our first question. Most articles present new community detection algorithms and compare them to existing ones, using real-world and artificially generated data. However, the algorithms are generally compared only in a quantitative way, thanks to some performance measures [6]. Yet, algorithms rely on different formal definitions of what a community is. It therefore seems incomplete, or even unfair, to compare algorithms which do not actually try to detect the same objects. Moreover, once communities have been identified, one wants to give them a meaning relative to the studied system, and this task is largely dependent on the selected algorithm.

We aim at offering the user the information he needs to determine which algorithms are adapted to his data, apply and compare them, and interpret their result in meaningful terms, relatively to the applicative context. As an illustration, we will apply the described methods to some data describing a sample of 552 university students. These data were gathered during a survey performed in the Galatasaray University at Istanbul, Turkey [7]. Its goal was to retrieve the information needed to extract a network representing the students' social interactions, and perform an analysis of their purchasing behavior. Thus, besides the social network itself, the data includes a whole set of nodal attributes describing *factual* (age, gender, clubs membership, etc.), *behavioral* (perceived actions in terms of human interaction and purchasing behavior) and *sentimental* (personal thoughts and feelings relative to university, friends, desires, favorite brands, etc.) information. In this chapter, however, we do not mean to conduct an exhaustive analysis of these data, but simply to use them as a practical example (cf. [7] for the details regarding the survey and this analysis).

The rest of this chapter is organized as follows. Section two is dedicated to community detection algorithms: we describe their properties, how to compare their results, and how to select the most relevant community structure. In the third section, we show different types of analysis oriented towards the interpretation of the community structure. We focus on different methods allowing to characterize communities, based on both topological information and nodal attributes. Finally, we conclude by mentioning alternative methods which we could not describe in details.

2 Community Detection Process

Our goal in this section is first to review the existing community detection methods from the user's perspective. Usually, these algorithms are presented from the author's perspective, with emphasis on process, performance and computational cost [6]. However, the community detection problem is known to be ill-defined [3,8,9,5], which is why so many different algorithms exist: they do not define the concept of community in the same formal way. They consequently do not necessarily detect the same communities. Under these conditions, comparing raw performances obtained from different algorithms seems very little relevant.

We think the final user is basically interested in three properties. First, the type of information the algorithm is able to process. Indeed, there are various ways of describing a network and one can embed different sorts of data: link attributes (weights, directions), node attributes, different classes of links (multiplex networks) or nodes (n-mode or multipartite networks), temporal information, etc. The user may want to select a method able to take advantage of all the available data. In this chapter, we decided to focus on plain networks, with simple links.

Second, the kind of community structure the algorithm produces. One generally distinguishes *partitions* and *covers*, i.e. mutually exclusive and overlapping communities. We decided to focus on the former, because only a few algorithms are able to identify covers already. Most algorithms output a single partition, but some of them are able to produce a collection of community structures estimated for different granularities. In the case of hierarchical algorithms, communities belonging to neighboring granularities are hierarchically related. In a given level, communities may correspond to the merging of several lower level communities, while being a part themselves of larger communities in the upper level. *Multiresolution* methods also estimate the community structure at different granularities, but without looking specifically for hierarchical relationships between them. They either scan automatically various scales or allow to specify them parametrically [10].

Third, the nature of the communities the algorithm is able to identify. As stated before, there are many ways to define formally what a community is. Yet this concept is at the center of the analysis, and is therefore of utmost importance. The user should select his tool mainly based on this feature.

In order to give the user all the information he needs, we reviewed community detection methods according to the three properties we mentioned. Note excellent reviews exist, which describe in great details the points we chose to ignore here [3,8,9,11]. The rest of the section is more practical. We present a list of publicly available tools and summarize their features in the previously mentioned terms. We then consider the very common case where one could estimate several community structures for a network of interest. We present various ways to tackle the problem of selecting the most appropriate community structure depending on the user's criteria and objectives.

2.1 Concept of Community

A very widespread informal definition of the community concept considers it as a group of nodes densely interconnected compared to the rest of the network [3,8,9,11]. In other terms, a community is a cohesive subset clearly separated from the rest of the network. Formal interpretations try to formalize and combine both these aspects of cohesion and separation. Note this definition is not always explicit: procedural approaches exist, in which the notion of community is implicitly defined as the result of the processing. Although it is not always straightforward to categorize the definitions, we regroup them in four classes: density-, pattern-, node similarity- and link centrality-based approaches. The last subsection is dedicated to methods which did not fit in the previous definitions.

2.1.1 Density

A whole family of formalizations is based on a direct translation of the informal community definition given above. The general approach consists first of specifying two distinct measures to assess separately cohesion and separation, and then in defining a global measure by considering their difference or ratio. For instance, Mancoridis *et. al* [12] defined their intra-connectivity and inter-connectivity to measure the cohesion and separation of a community, respectively. The former is simply the regular density processed when considering only the links located inside a community, i.e. connecting two nodes belonging to the community. The latter is the density processed when considering only the links between a pair of communities. Let us note n_C the number of nodes in community C, and m_{CD} the number of links between communities C and D. Then, for an undirected network, the intra-connectivity of community C is:

$$A_C = \frac{m_{CC}}{n_C\,(n_C - 1)/2} \tag{1}$$

and communities C and D inter-connectivity ($C \neq D$) is:

$$B_{CD} = \frac{m_{CD}}{n_C n_D} \tag{2}$$

Mancoridis *et al*. proposed to quantify the quality of a whole community structure by considering the difference between these measures averaged over the network ($i \neq j$):

$$MQ = \langle A_i \rangle - \langle B_{ij} \rangle \tag{3}$$

Fortunato gives a different definition of the inter-connectivity in his review [3], by considering the links between the community of interest and the rest of the network:

$$B'_C = \frac{\sum_{i \neq C} m_{Ci}}{n_C(n - n_C)} \tag{4}$$

where n is the total number of nodes in the network.

Instead of using the density measure, some authors represent cohesion and separation in terms of internal and external degrees, respectively. The former corresponds to the number of links a node has with other nodes from its community, whereas the latter concerns the nodes located out of the community. If we note k_i the number of links a node has with some community i and if we consider a node belonging to community C, then its internal degree is k_C and its external degree is $\sum_{i \neq C} k_i$. This led to the notions of *weak* and *strong* communities [13]. The former is characterized by the fact all of its nodes have a greater internal than external degree, whereas the latter applies the same constraint to internal and external degrees of the community as a whole. Certain algorithms are based, sometimes implicitly, on the notion of strong community (or on a related definition), like for instance the Label Propagation method [14].

Alternatively, in place of deciding what is and what is not a community, it is possible to use these degrees to quantify how good a community is. The *conductance* Φ_C of a community C is the ratio of its external degree to the minimum between its total degree and that of the rest of the network [15]. In the case of a community much smaller than the network, it is therefore its proportion of external links:

$$\Phi_C = \frac{\sum_{i \neq C} m_{Ci}}{\sum_i m_{Ci}} \tag{5}$$

Although not explicitly, many algorithms optimize this quantity or one of its variants [3], via spectral analysis of matrices derived from the adjacency matrix [16], use of certain random walk-based distances, simulation of synchronization processes, etc. Lancichinetti *et al.* defined a similar measure at the level of the node: their *embeddedness* e corresponds to the ratio of the node internal degree k_C to its total degree k [5]:

$$e = k_C / k \tag{6}$$

It can be averaged over the node community or the whole network, to assess the quality of the community or the community structure, respectively. In the latter case, the obtained measure is close to the *coverage* measure, which is the ratio of the intra-community links to the total number of links in the whole network [3]:

$$CV = \sum_i \frac{m_{ii}}{m} \tag{7}$$

where m is the total number of links in the network.

Newman's original *modularity* [4] can be viewed as a chance-corrected version of the coverage considered at the level of the communities. Let us note $q_{ij} = m_{ij}/m$ and $q_{i+} = \sum_j q_{ij}$ then the modularity is:

$$Q = \sum_i (q_{ii} - q_{i+}^2) \tag{8}$$

First the proportion of internal links in the whole network is processed over all communities ($\sum_i q_{ii} = CV$), and then the corresponding proportion estimated for a comparable random network is subtracted ($\sum_i q_{i+}^2$). The null model used by Newman is a randomly rewired version of the network of interest, with preserved size (numbers of nodes and links) and degree distribution. Such a network is not supposed to have any community structure because of the uniformly random rewiring. Therefore, in order to have a significant community structure, the network of interest is required to have a much greater proportion of internal links. The modularity is certainly the most popular measure to assess the quality of a partition community structure. Many algorithms were designed to optimize it, explicitly or not: spectral approaches [17], random walk-based distances [18], genetic algorithms, greedy approaches [4,19-22], simulated annealing [23], mathematical programming [24], extremal optimization, spin glass model [10], etc.

However, the modularity is known to have at least two important limitations. First, its maximal value is not constant, and depends on the considered network structure, which makes it impossible to compare modularity values between different networks. It could be normalized using the maximal modularity of the associated null model, but this value is itself difficult to process [3]. Second, it has a resolution limit, meaning it cannot detect perfectly valuable communities if they are smaller than a critical size depending on the network itself [25]. Several extensions such as [10] were developed to solve this problem.

Most density-based definitions have been extended for weighted and directed networks (conductance [26], modularity [27,28]). The extension is generally straightforward, by considering strength instead of degree for weighted links, and by distinguishing in- and out- degree for directed ones. The adaptation of the algorithms is not always as simple though, for instance spectral approaches are more difficult to apply when the adjacency matrix is asymmetric, which is generally the case when dealing with directed networks [26].

2.1.2 Pattern

Another way to define cohesion and separation consists of identifying maximal subsets composed of small specific interconnection patterns, e.g. cliques. One can consider a community to be either the largest identified pattern, or a set of patterns with common nodes [29,3]. This approach can be seen as more qualitative than the density-based one, because it does not rely only on numeric values to formalize these concepts. Separation is represented by the fact one is looking for maximal subsets, which implies these are separated from the rest of the network. The nature of this separation and the notion of cohesion both depend on the selected interconnection pattern.

The most basic pattern one can use is the clique, a set of completely interconnected nodes. Luce & Perry present a clique as a group of mutual friends [30]. The connectivity is complete and direct, i.e. for a set of n nodes, each node is connected to the $n-1$ other nodes from the clique, and is consequently at a distance 1 of anyone of them. However, a clique structure represents a strong constraint, especially for real-world networks [3,29]. For this reason, many partially relaxed variants exist, which focus either on the complete or the direct aspects of clique connectivity. The patterns called k-plex and k-core belong to the first kind. For the latter, k represents the minimal number of neighbors a certain node must have in the pattern [31]. On the contrary, for the former, it is the maximal number of non-neighbors [32].

The concept of n-clique relaxes the direct aspect of clique connectivity: it does not require all nodes to be connected by a direct link, but at least by a path whose length is at most n [30]. However, this pattern is too relaxed and allows paths to go through nodes located out of the n-clique, possibly leading to an n-clique made of disjoint subsets of nodes. This, of course, is not compatible with the intuitive notion of community, which implies connectedness. For this reason, another pattern called n-clan was defined by adding a constraint on the diameter of an n-clique, stating it should not be greater than n [33].

The approach can be extended to consider directed or weighted links. For instance, an f-group is a maximal subset of weakly and strongly transitive triads. A triad is a set of three nodes, and it is considered as transitive if it is completely connected (i.e. a 3-node clique, or triangle). According to Hanneman, it is strongly transitive if all three links have the same weight, and weakly transitive if the link with the smallest weight is at least above a certain threshold [29]. Palla *et al.* presented a clique-based method to process both directed and weighted networks [34]. However, to our knowledge, no extension was designed to deal with individual information (i.e. nodal attributes).

Most pattern-based algorithms are computationally demanding [3]. Although this is a drawback in the context of complex networks analysis, due to their size, the pattern-based approach still has an interesting advantage: it allows specifying more precisely the internal structure of the communities. If any *a priori* knowledge of the studied system is available, it is possible to use it to constraint the community identification process. Of course, the pattern has to be chosen thoroughly: some networks do not exhibit certain patterns. For instance, technological networks and certain social networks do not contain many cliques [3].

2.1.3 Node Similarity

By using an appropriate similarity function, the topological notions of cohesion and separation can be translated in terms of intra-community similarity and inter-community dissimilarity. In other terms: a community is viewed as a group of nodes which are similar to each other, but dissimilar from the rest of the network. Once all node-to-node distances are known, detecting a community structure can be performed by applying a distance-based classic cluster analysis algorithm [35]. Such a tool is designed to minimize the internal and maximize the cluster-to-cluster distances. Depending on the desired output (overlapping vs. mutually exclusive community, hierarchy of community), different clustering algorithms can be applied [36,37].

The strong point of this approach is the possibility to include any information in the definition of the similarity function. Purely topological functions exist, such as those based on structural and regular equivalence, which state two nodes are similar if they share the same connection pattern to the same neighbors, or possibly different neighbors, respectively. Structural equivalence can be quantified using, for instance, Jaccard's coefficient [38] (ratio of the intersection cardinality to the union cardinality of two sets of interest) applied to both nodes sets of neighbors (other methods exist, see [3]). If they are structurally similar, two nodes are supposed to be close (and hence to belong to the same community) even if they are not directly linked, because they are likely to be indirectly connected through their neighbors. Note strict equivalence is sometimes too restrictive, and relaxed versions exist (cf. the appendix of [39]).

Other topological functions rely on paths instead of direct connections. One can consider the number of paths, or distinct paths (i.e. the same node or link does not appear twice), or shortest paths between two nodes to assess their similarity: the highest this number, the more similar the nodes. Some authors rather adopt a probabilistic approach, considering random walks. The expected path lengths can be processed, for instance the *first passage time* is the expected number of steps a random walker would need to go from the source node to the target one [40], while the *commute time* additionally considers the return time [41]. An alternative is to consider instead the probability value itself: probability to visit the target node in a given number of steps [42], probability to reach it before coming back to the source [43], etc.

Finally, similarity can also be defined using both topological and individual information. In [44], Handcock *et al.* make the assumption the nodes of a network can be characterized by their location in an unobserved so-called *social space*. This location depends on topological information and nodal attributes. Communities are identified by clustering nodes depending on their distance in this social space.

2.1.4 Link Centrality

The concept of community can also be defined in terms of link centrality. There are several definitions for this notion, but link centrality is basically related to two properties: the number of pairs of nodes the link is connecting (directly or not) and how likely these connections are to be used. Under these terms, links located between communities are supposed to be very central, since they allow to connect the nodes from one community to those from the other one, and there are only few of them (by definition inter-community links are sparse) so they are very likely to be used. On the contrary, the links located inside communities connect comparatively few nodes (mainly those from the same community), and the community is supposed to be densely connected, so many different path exist to connect two nodes, making it less likely for a link to be used. In other words, the high centrality of inter-community links and the low centrality of intra-community links relate to separation and cohesion, respectively.

Tyler *et al*. explicitly defined a community as a set of nodes whose links centrality must not be greater than a certain threshold [45]. They consider the most isolated node a community can contain is a leaf (degree 1 node), whose only link has the maximal centrality in this community. They consequently define their threshold as the centrality exhibited by this link. The fact some node set contains a link more central than this threshold means this link connects two subsets both larger than one node. These subsets could be separated, leading to two communities.

Various edge centrality measures were defined using principles not unlike those employed for path-based node centrality measures. Some of them are not adapted to this case though: number of paths (generally infinite), distinct paths (inefficient on degree 1 nodes). Girvan and Newman defined their *edgebetweenness* measure by considering the total number of shortest paths going through a link [46]. They also used the non-deterministic approach and defined a random walk centrality based on the probability a link has to be passed by the walker, averaged over all pairs of source and target nodes. The extension to directed links is straightforward (one consider only directed paths). Newman proposed extensions of both measures for weighted links [27], by normalizing edgebetweenness with the considered link weight, and by using weights to process the random walker transition probabilities. Although not explicitly stated, the approach described in [47] is related to link centrality, this time defined in terms of currents flow. The network is view as a resistor network and inter-community links are characterized by significant voltage differences.

Radicchi *et al*. proposed an alternative link centrality called *edge clustering* [13]. It corresponds to the ratio of the number of existing cycles containing the link of interest, to the number of possible cycles given the existing links. Therefore, unlike betweenness centrality, a high value means here the link is likely to be inside a community, since cycles are much more likely to happen there. The measure was extended to weighted links similarly to what was done for the edgebetweenness, i.e. using a normalization based on the weight of the considered link [48].

2.1.5 Others

Certain definitions of the concept of community do not fit the classes we described in the previous subsections. We present here only two of them, because they are used in some of the publicly available algorithms we present in the following section. The reader should notice other specific approaches exist, though (see [3]).

To define the concept of community, Rosvall & Bergstrom [49] do not use an approach based on cohesion and separation like all the previous community definitions. They adopt a data compression perspective and consider the community structure as a set of regularities in the network topology, which can be used to represent the whole network in a more compact way. The best community structure is therefore the one maximizing compactness while minimizing information loss. They implement this definition through the use of the mutual information measure applied to different representations of the network based on the adjacency matrix [49] and on a node nomenclature [50]. Ziv *et al*. adopted a comparable approach, but used instead a diffusion process to represent the network [51].

Van Dongen proposed to simulate another kind of diffusion process in the network to detect communities [52]. This approach relies on the transfer matrix of the network, which describes the transition probabilities for a random walker evolving in this network. Two specific transformations are iteratively applied on this matrix. First, it is raised to some specified power, in order to get a transfer matrix containing probabilities for longer paths. Second, each element in the matrix is raised to some specified power, in order to favor the higher probability values, which correspond to nodes presumably belonging to the same community. The resulting matrix is then normalized to get a new transfer matrix. Both steps are repeated until convergence. The resulting matrix can be interpreted as the adjacency matrix of a network with disconnected components. These correspond to communities in the original network.

2.2 Publicly Available Tools

In this section, we present publicly available implementations of community detection algorithms. Table 1 shows them in order of publication, with their main features. A large part of these algorithms are dedicated to modularity optimization. The first is *Fast Greedy*, a C implementation of a greedy approach by Newman & Clauset [4,19] (http://cs.unm.edu/~aaron/research/fastmodularity.htm). It is able to process large networks, however it suffers from a bias toward large communities. Several variants were defined to correct this: *Wakita-Tsurumi* [22] (Java implementation at http://ken-wakita.net/research/en/software), *Multistep Greedy* [20] (C++ implementation at http://www.biochem-caflisch.uzh.ch/public/5). The *Louvain* algorithm [21] (C++ code at http://sites.google.com/site/findcommunities) implements a different greedy approach designed for very large networks. Newman also proposed his *Leading Eigenvector* algorithm [17] to optimize modularity by applying a spectral approach on a specific matrix. The *NetCarto* algorithm [23] (C code available on demand to its authors) implements a simulated annealing approach, which allows it to get very close to the actual optimum, but makes it in turn very slow. Reichardt and Bornholdt reformulated the modularity optimization problem using a *Spin Glass* model [10]. Their approach actually generalizes modularity in order to overcome its resolution limit, and let the user specify a resolution parameter. With *TimeScale* [18] (C++ source code available at http://www.lambiotte.be/codes.html) Lambiotte *et al*. proposed to apply a related extension of the modularity on their Louvain algorithm. Finally, the version of Agarwal & Kempe [24] (C++ and Java codes at http://www-scf.usc.edu/~gaurava) adopts a mathematical programming approach to the same modularity optimization problem.

Besides the modularity, other density-based definitions of the community concept are used. *CommFind* adopts a spectral approach to optimize a partition quality measure related to the conductance [16,53] (C code at http://wdb.ugr.es/~donetti). *VBmod* [54] (Matlab code at http://www.columbia.edu/~chw2) relies on a Bayesian approach whose probabilistic model is related to the embeddedness measure. *Label Propagation* [14] simulates the spread of values in the network until convergence, and identifies communities as sets of nodes associated to the same value. At the end of the process, the value associated to a node is the majority one amongst its neighbors, so this can be seen as a relaxed version of the strong community concept [13].

Node similarity-based approaches are also fairly represented. *WalkTrap* [42] is based on a random walk distance which considers the probability to go from one node to another in a given number of steps. This parameter affects the resolution of the resulting communities, so the tool can be considered as multiresolution. Zhou also used a random walk based distance, but this time considering the excepted number of steps to from one node to another [40] (Fortran implementation at http://www.mpikg-golm.mpg.de/th/people/zhou). *Jerarca* [55] uses an original distance definition based on the detection of patterns (C++ implementation at http://jerarca.sourceforge.net). Three distinct distance functions with different computational complexities are defined based on different patterns.

The *EdgeBetweenness* algorithm [46] was the first link centrality-based tool. Radicchi *et al*. proposed a variant relying on their edge clustering measure [13,48] (C code at http://filrad.homelinux.org). Hu & Huberman used a different approach based on currents flow [47].

Several approaches are based on a compression view of the community structure (cf. section 2.1.5): *InfoMod* [49] and *InfoMap* [50] C++ implementations are available at http://www.tp.umu.se/~rosvall/code.html (the latter was recently extended to output dendrograms), whereas the Matlab code for *ITmod* [51] can be downloaded at http://www.columbia.edu/~chw2. Finally, the diffusion-based approach implemented in *MarkovCluster* [52] can be found at http://www.micans.org/mcl (C code). An inflation parameter allows setting the granularity of the search, making the approach multiresolution.

Table 1. List of publicly available community detection tools, and their main features. The inputs are simple (S), weighted (W) or directed (D) networks. The outputs can be a simple partition (P), or a collection of partitions hierarchically ordered (H) or not (multiresolution, M). Only the class of community definition is indicated, see the text for more details. The complexities are expressed for sparse networks, i.e. the number of links is of the order of the number of nodes (Param. dep. stands for parameter dependent). Implementations can be author-made (A, see the text for details and URL) or belong to the igraph library (I) [56] (R and Python languages), the Jung package (J) [57] (Java) or the Gephi software (G) [58] (Java).

Name	Input	Output	Community	Complexity	Impl.
Edge Betweenness [46]	S, D	H	Link centrality	$O(n^3)$	I, J
Zhou [40]	S, W	H	Node similarity	$O(n^3)$	A
Radetal [13,48]	S, W	H	Link centrality	$O(n^2)$	A
Fast Greedy [4,19]	S, W	H	Density	$O(n\log^2 n)$	A
CommFind [16]	S	H	Density	$O(n^3)$	A
NetCarto [23]	S	P	Density	Param. dep.	A
Wu-Huberman [47]	S, W, D	P	Link centrality	$O(n \log n)$	J
WalkTrap [42]	S, W	H , M	Node similarity	$O(n^2 \log n)$	A, I
ITmod [51]	S, W	H	Compression	-	A
Leading Eigenvector [17]	S	H	Density	$O(n^2 \log n)$	I
SpinGlass [10]	S, W	M	Density	Param. dep.	I
Label Propagation [14]	S, W	P	Density	$O(n)$	I
InfoMod [49]	S	P	Compression	-	A
Wakita-Tsurumi [22]	S	H	Density	$O(n\log^2 n)$	A
Agarwal-Kempe [24]	S	P	Density	$\Theta(n^2)$	A
Louvain [21]	S, W	H	Density	$O(n)$	A, I
MarkovCluster [52]	S, W, D	M	Diffusion	$O(n^3)$	A, G
VBmod [54]	S	P	Density	$O(n^2)$	A
InfoMap [50,59]	S, W, D	H	Compression	-	A
Multistep Greedy [20]	S, W, D	H	Density	$O(n\log^2 n)$	A
TimeScale [18]	S, W	H, M	Density	-	A
Jerarca [55]	S	H	Node similarity	$O(n \log n)$	A

Pattern-based implementations are mainly used to detect cover and not partitions (e.g. *Cfinder* [60]), which is why they are not represented here. Note some of these algorithms are also very conveniently implemented in libraries dedicated to network analysis, such as *igraph* [56] and *Jung* [57] (see Table 1), which gives the user a uniform access to their functionalities.

Most of these algorithms were individually tested on both real-world and randomly generated networks, and several review articles directly compared some of them [6]. However, these performance assessments have to be considered with caution. Concerning real-world networks, the reference communities have to be manually defined, and are therefore subjective. On the contrary, in the case of generated networks, they are objective because they are a part of the generative process. However, this process itself is biased in direction of one definition of the community concept (e.g. embeddedness for [6]), and the resulting benchmark therefore favors algorithms based on the same definition. The only relevant comparison concerns algorithms all based on the same community definition, like for instance the various ways of optimizing the modularity.

2.3 Comparing Partitions

Thanks to the information provided in the previous sections, the user should be able to choose an appropriate tool based on the data to process, the desired kind of community structure, and most of all a relevant definition of the community concept. However, various situations can lead to results taking the form of several partitions, when one is generally interested in a single one. First, given the profusion of algorithms, several of them might be adapted to a given study, probably resulting in several different partitions. Second, even if a single algorithm is used, one can obtain a collection of community structures if this algorithm has a hierarchical or multiresolution output. In both cases, the user has to make a choice in order to select the community structure he is going to interpret. In this section, we present methods to make this choice.

2.3.1 Different Algorithms

In the case where one has several partitions coming from different algorithms, the simplest way seems to be comparing the quality of the partitions through the use of a quality measure, and ultimately selecting the partition with the highest quality. However, different problems can arise. First, if the algorithms rely on different community definitions, the quality measure, which has itself to implement such a definition, will be biased towards certain algorithms. Second, even when comparing algorithms using the same definition, e.g. modularity optimization methods, the quality measure may present limitations. For instance, the modularity is known to have a resolution limit, which means it will disadvantage partitions displaying communities below this limit, even if these are the actual communities.

A complementary approach consists of comparing the partitions themselves instead of their qualities. The goal is then to assess how much algorithms agree rather than to identify the best partition. This is particularly relevant in the context of an exploratory analysis where one could not choose a community definition adapted to his data and decided to use several algorithms based on various definitions. The fact these algorithms identify similar partitions is a sign of the stability of the community structure, whereas if they are very different, one should question his results.

We propose to use the *adjusted Rand index* (ARI), which is rather popular in cluster analysis. The original Rand index (RI) [61] is defined as:

$$RI = \frac{a + d}{a + b + c + d} \tag{9}$$

where a (resp. d) corresponds to the number of pairs of nodes belonging to the same (resp. different) community in both partitions, and b (resp. c) to the number of pairs whose nodes belong to the same community in the first (resp. second) partition, whereas they belong to different communities in the second (resp. first) one. The adjusted version [62] is defined as:

$$ARI = \frac{RI - E}{1 - E} \tag{10}$$

where E is the amount of similarity expected to be due to chance, estimated by considering the products of marginals: $E = (a + b)(a + c)/n^2 + (b + d)(c + d)/n^2$. The upper limit of this measure is 1 (the two partitions are exactly the same). The value 0 indicates a partial overlap, equivalent to what would be observed if both partitions were random (i.e. $RI = E$). Negative values indicate a strong divergence between the partitions. Note there are other measures one can use to assess the similarity of two partitions [36,3]. We can also mention the normalized mutual information, which has been used in recent community detection works [6].

Table 2. Agreement measured by the ARI for a selection of community detection algorithms.

Algorithm	Fast Greedy	SpinGlass	Label Prop.	InfoMod	MarkovCluster
Fast Greedy	-	0.80	0.52	0.30	0.36
SpinGlass	-	-	0.57	0.26	0.40
Label Prop.	0.57	0.57	-	0.14	0.68
InfoMod	0.26	0.26	0.14	-	0.09
MarkovCluster	0.40	0.40	0.68	0.09	-

As an example, we applied several community detection algorithm to our social network of university students. Table 2 gives the ARI values for some of these results. One can notice the maximal agreement is reached for the two modularity-based algorithms (Fast Greedy and SpinGlass). Moreover, their ARI values when compared to the other algorithms are very close, so we can conclude both partitions are certainly highly similar. The other algorithms differ in the definition of community they rely on, and this shows through the ARI values: InfoMod, with its information theory-based approach, is isolated and largely disagrees with the others. Although they do not use the same approach at all, Label Propagation and MarkovCluster partially agree. Their partitions are nevertheless significantly different from those estimated by the modularity-based approach.

2.3.2 Different Granularities

Consider now the case where one wants to compare several partitions corresponding to different granularities output by the same algorithm. If the algorithm is hierarchical, the agreement approach is not relevant, because agreement measures take the hierarchical aspect into account, i.e. two partitions corresponding to different levels in the same hierarchy will necessarily be very similar. The approach can be applied to multiresolution outputs though, in order to check if the partitions obtained at different granularities are really different. If they are similar, on the contrary, one can conclude they are related by a partial hierarchical order.

In both the hierarchical and multiresolution cases, partitions can be compared through their quality, like in the previous subsection. Moreover, here only one algorithm is involved, so it makes sense to rely on the quality measure it optimizes. However, not all algorithms use such a measure, in which case one has to select a measure which would be compatible in terms of community definition. For instance, using the modularity to select the best cut in a dendrogram produced by the EdgeBetweenness algorithm seems rather inappropriate, because the algorithm was not designed to maximize it. But there are not so many quality measures, and in practice the modularity is used most of the time.

The partition quality is important, but is not necessarily the only criterion to take into account. Indeed, one generally wants to identify a community structure in order to subsequently interpret it. He will therefore be interested in the number of communities and in their size: too large or too small values are likely to prevent any meaningful interpretation. Alternatively, some knowledge concerning the studied system might allow for the definition of preferences regarding these quantities. Under these conditions, the selection of the most appropriate partition should result from a compromise between the measured quality and the nature of the community structure.

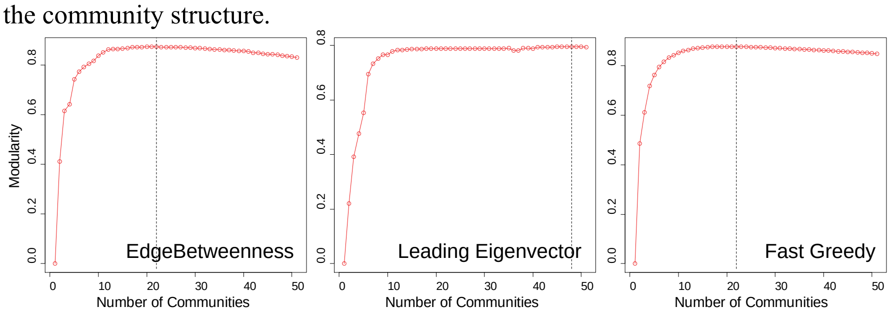


Fig. 1. Modularity values obtained for three hierarchical algorithms when applied to our data. Only the higher levels of the hierarchy are represented. The dotted lines indicate the partitions of maximal modularity.

The partition quality measured over the dendrogram output by a hierarchical algorithm often follows the evolution displayed in Fig. 1 for three hierarchical algorithms we applied on our data. In particular, one may notice the partitions surrounding the partition of maximal quality (dotted line) have very similar quality themselves. This situation is favorable to the compromise we mentioned, because it supports the selection of a neighboring partition without losing too much quality. Suppose we want to select a partition containing fewer communities than the optimal one, i.e. a partition located a few merges away. We have to consider candidates relatively to two criteria: the loss of quality compared to the maximal quality partition, and the number of nodes concerned by the merges. This optimization problem is extremely context-dependent, and it is therefore difficult to propose a general method. A reasonable approach consists of defining two limits based on the modeled system and the user's objectives: first the maximal acceptable loss in quality, and second the maximal size allowed for a merged community. The user can then select the partition with fewest communities respecting both constraints. Let us consider the hierarchy estimated by Fast Greedy on our data. The best partition has 22 communities, with a modularity of 0.8780. Suppose we allow a quality loss of 0.01 and the merge of communities representing up to 5% of the network nodes. Then we could select the 13-community partition, with a modularity of 0.8696 (loss 0.0084), the largest community merged containing 4.2% of the nodes.

3 Interpretation of the Communities

Community detection is not an end in itself: once communities have been identified, one wants to understand what they mean. Two kinds of analysis can be performed for this matter. First, it is necessary to study the topology of the community structure. This allows assessing the structural significance and quality of the community structure, but also starting the interpretation process, by discussing the similarities and differences observed between the communities, and by identifying nodes with specific roles. The second phase of the analysis relies on the exploitation of nodal attributes. It is guided by the structures identified during the first phase thanks to the topology of the network (communities, nodes of interest). It consists of characterizing and discussing these structures in terms of the numeric or nominal data specific to the considered system and application domain. In this section, we present consensual tools allowing to perform these analysis. We illustrate their use on our data, commenting from a Business Science perspective the communities identified by Fast Greedy, which are represented on Fig. 2. In this field, detecting communities is a very valuable task with huge implications, especially if these communities can be characterized in terms of specific purchase behaviors.

3.1 Topological Properties

Classical network analysis can be performed both at a macroscopic and microscopic levels, i.e. by considering respectively topological properties of the network as a whole, or of some specific nodes taken individually. Networks can be characterized by a whole set of measures such as density, transitivity (a.k.a. clustering coefficient), degree distribution, etc. (see [63] for a very comprehensive review). However, in this chapter, we rather focus on the community structure, which adds an intermediary level. It allows not only a mesoscopic analysis, but also brings a new point of view regarding individual nodes: one can consider their position in their respective communities or in the community structure (by opposition to their position in the whole network). In this section, we first introduce tools allowing to assess the quality of the communities collectively and separately, and then we consider the characterization of nodes relatively to the community structure. When not indicated differently, we used the igraph library [56], which also contains several community detection algorithms (cf. section 2.2), to process the topological properties. Figures have been produced using igraph and the Java open source software Gephi [58].

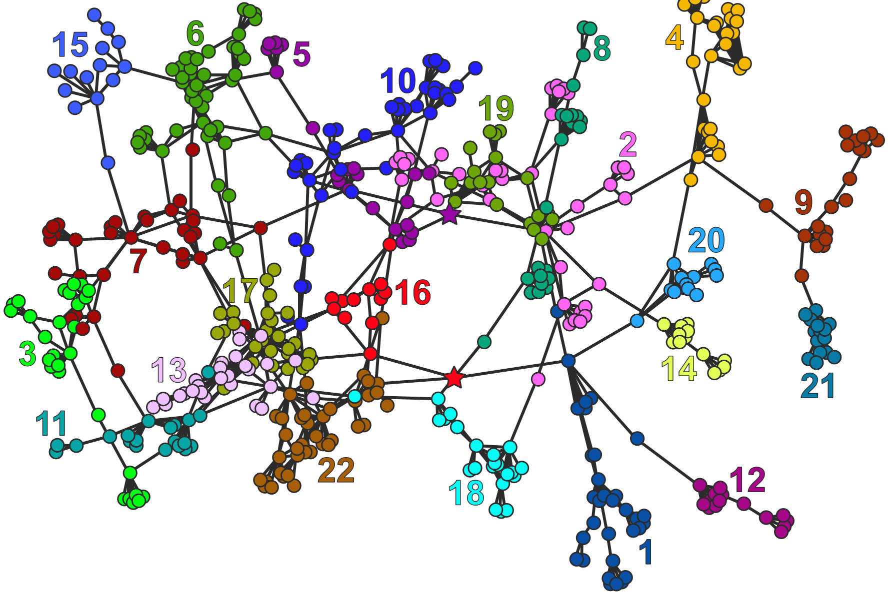


Fig. 2. Community structure obtained with Fast Greedy (maximum modularity cut). Each one of the 22 communities is represented by a different color. The two stars stand for the nodes with minimal embeddedness ($e = 0.33$).

3.1.1 Communities

Before starting the analysis of the community structure, it is important to evaluate its significance. Various methods were rather recently proposed for this purpose [3], but the one described in [64] has the advantage of being independent of the modularity measure, and to allows evaluating the communities separately (instead of the whole distribution). The C++ implementation is available at http://filrad.homelinux.org. This method relies on a null model similar to the one used in the modularity measure (cf. section 2.1.1). The authors propose two measures to quantify the community significance. The first one is called the C-score and corresponds, for a given community, to the probability of appearance of a community with similar topological features in the null model. It is based on the so-called worst node of the community, i.e. the node with lowest internal degree. The C-score is estimated by considering the probability for its counterpart in the random network to have an equal or larger internal degree. The second measure, called B-score, extends the C-score by considering several nodes instead of a single one. The resulting measure is supposedly more relevant, but also computationally more demanding [64]. We applied it to our data, and Table 3 shows 21 communities out of the 22 identified by Fast Greedy are significant ($B < 0.05$), the only exception being the 16th ($B = 0.089$). Note the significance of the community structure can be considered as an additional criterion in the community structure selection problem introduced in section 2.3.2.

The first step in the analysis of the community structure is generally to characterize the *distribution of community sizes* (expressed in nodes), which is supposed to follow a power-law in many real-world networks [19,3]. In our case, the number of communities is too small for this distribution to be statistically tested. It can be noticed (cf. Table 3) it is right-skewed though, with a single large community and many small ones. However, the difference between the smallest and largest communities is not comparable to what can be observed in other networks [5]. Consequently, we can conclude our community structure is relatively homogenous regarding the community sizes.

One of the most important aspects of the identified communities is their quality in terms of cohesion and separation. Several properties can be used for this assessment. In terms of cohesion, one can consider the *density* of each community, when considered separately from the rest of the network. By definition, communities are denser subgraphs, so their density is supposed to be much larger than for the whole network. Table 3 shows this is very much the case for our data, with a network density of 0.01 when most communities are 10 times denser. We remind the reader real-world networks are generally sparse, which explains the low density observed on our data. Moreover, sparsity is actually a prerequisite for the existence of a community structure [3]. The density varies much between our communities. It is strongly correlated to their size ($r = -0.72$), which indicates the smaller the communities, the denser they are.

A small *average distance* between nodes of the same community is also a sign of good cohesion. In our data, the average distance of a community is much smaller than its size. Of course, it is also much smaller than the distance averaged over the whole graph, due to its community structure and sparsity. Communities are supposedly small-world, which means the average distance increases logarithmically with the community size [5]. In our case, the distances are highly correlated with the logarithm of the community sizes ($r = -0.77$), however we could not perform a significant test due to the small number of communities.

Table 3. Topological properties of the network and its communities: n is the number of nodes, d the density, $\langle e \rangle$ the average embeddedness, ℓ the average distance, k_{max} the maximal degree, h the hub dominance and B the B-score.

Community	n	d	$\langle e \rangle$	ℓ	k_{max}	h	B
1	32	0.07	0.96	3.57	10	0.32	0.012
2	39	0.06	0.93	3.99	10	0.26	0.018
3	28	0.08	0.99	3.20	10	0.37	0.001
4	30	0.11	0.98	2.99	10	0.34	0.001
5	23	0.09	0.94	3.45	8	0.36	0.014
6	46	0.07	0.97	3.28	11	0.24	0.001
7	34	0.08	0.93	3.05	11	0.33	0.002
8	23	0.09	0.97	3.14	10	0.45	0.002
9	20	0.11	0.95	3.36	9	0.47	0.001
10	39	0.07	0.96	3.43	10	0.26	0.013
11	20	0.12	0.93	2.64	9	0.47	0.034
12	15	0.13	0.99	2.59	9	0.64	0.002
13	28	0.11	0.96	2.53	12	0.44	0.001
14	13	0.15	0.99	2.21	8	0.67	0.003
15	14	0.16	0.95	2.44	9	0.69	0.038
16	13	0.19	0.86	2.15	7	0.58	0.089
17	28	0.11	0.96	2.63	10	0.37	0.002
18	22	0.15	0.97	2.53	10	0.48	0.005
19	20	0.16	0.97	2.60	8	0.42	0.006
20	12	0.35	0.97	1.74	9	0.82	0.012
21	15	0.24	0.99	1.90	10	0.71	0.001
22	38	0.09	0.97	2.91	12	0.32	0.000
Network	552	0.01	0.96	8.48	12	-	-

A small average distance can be explained by a high density and/or the presence of hubs, i.e. nodes connected to most of the other nodes belonging to the same community [5]. Hub dominance can be assessed using the following ratio:

$$h = \max_{C}(k)/(n_C - 1) \tag{11}$$

where $\max_C(k)$ and n_C represent the maximal degree and number of nodes in community C, respectively. When at least one node is connected to its whole community, it reaches unity. Table 3 shows only a few communities have a dominant hub (ratio greater than 0.5), and these are the smallest. Indeed, the correlation between community size and hub dominance is very strong ($r = -0.9$). This is due to the fact the maximal degree a node can reach is biased by construction of the network. Indeed, a student can cite a maximum of 10 friends, which makes it rather easy to get a degree of 10. But to get past this value, the student must be cited by persons he did not cite himself, which proved to be rather rare. Consequently, the maximum degree in a community is always very close to 10, independently from its size. The fact small communities are dominated by hubs while the large ones are not is a common feature of social networks [5].

Community separation can be measured by considering the proportion of links laying in-between them. In our case, only 52 out of the 791 links (6%) connect nodes of different communities. In other terms, the average number of links between two communities is only 0.23. This affects the embeddedness, as seen on Table 3. The values are averages over each community (and over the network, for the last one). The fact they are all very close to 1, including the network value, indicates nodes are very dominantly connected to other nodes from the same communities. This remark holds for all communities, independently from their size ($r = -0.02$). It is worth noticing the only non-significant community in terms of B-score (16^{th}) exhibits the lowest maximal degree and embeddedness. The embeddedness distribution is also interesting, because unlike what is generally observed in social networks [5], it is not uniform at all. Instead, most nodes are very strongly embedded in their community: only 4% of them have an embeddedness of 0.5 or less. We suppose this is due to the size of our network, which is much smaller than those studied in [5].

3.1.2 Nodes

Weakly embedded nodes are remarkable because they are generally located in-between communities: their small embeddedness reflects the fact there is no clear dominance among the communities of their direct neighbors. For example, Fig. 2 shows the two nodes with smallest embeddedness ($e = 0.33$), under the form of stars. Both are clearly lying at the interface of several communities. How these nodes can be used depends largely on the modeled system, but they generally constitute very valuable information. For instance, in the context of Business Science there are two main uses for them. First, these in-between nodes can be used as a base for certain communication strategies [65], consisting of making these persons as active as possible, in order to have them propagating messages to their contacts [66]. In this diffusion process, they can be considered as bridges between communities, and can play the role of accelerators. Second, these people can often be characterized by specific purchase behaviors, constrained by the fact they try to improve part of their social image in order to increase their membership to a group [67].

Other methods exist to characterize the position of a node relatively to the community structure. Guimerà & Amaral defined two measures for this purpose

[39]: the first concerns the node community whereas the second focuses on the rest of the network. The *within-community degree* z has more or less the same interpretation than the embeddedness: it quantifies how well a node is connected to the rest of its community. Its expression is different though, since it is defined as the z-score of the node internal degree relatively to its community C:

$$z = (k_C - \langle k_C \rangle)/\sigma \tag{12}$$

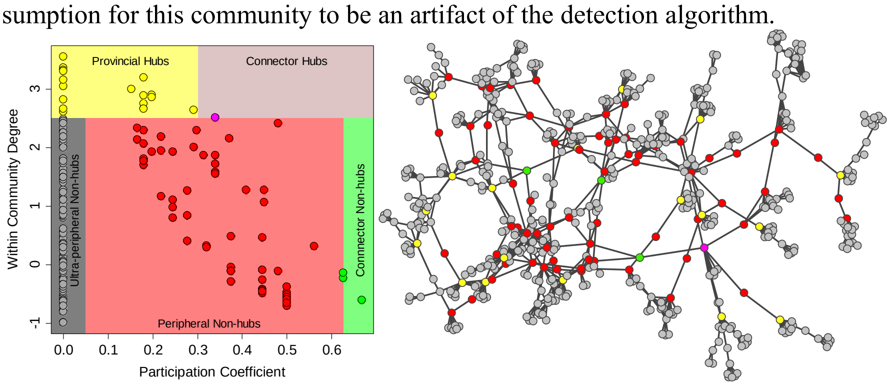
where $\langle k_C \rangle$ is the internal degree averaged over all nodes in community C, and σ is the corresponding standard deviation. A large within community degree means the node has many more links inside its community than most other nodes belonging to this community. The second measure is the *participation coefficient* P, which is defined as:

$$P = 1 - \sum_i (k_i/k)^2 \tag{13}$$

where k is the node total degree, k_i is its number of links with some community i (possibly its own community) and the sum is processed over all communities. It quantifies how much the node of interest is connected to multiple communities, and gets close to unity when it is evenly connected to all of them. On the contrary, when all the neighbors are in the same community ($k_C = k$), the participation coefficient is zero.

Guimerà & Amaral use both measures to characterize a node, and distinguish seven different roles depending on the observed combination of values, and to a set of thresholds. The choice of these thresholds is arbitrary [3] and we present here those determined empirically in [39]. First, nodes with a within-community degree smaller than 2.5 are considered as *hubs*, whereas the remaining ones are *non-hubs*. Finer roles are then defined by applying different thresholds on the participation coefficient. Hubs can be *provincial* (almost all neighbors in the same community, $P \leq 0.3$), *connector* (a majority of neighbors in the same community, $P \leq 0.75$), or *kinless* (less than half the neighbors in the same community, $P > 0.75$). The first can be considered as having an important local role for the cohesion of the community, the second allows connecting communities, and the third does not clearly belong to the community it was assigned to. Non-hubs can be: *ultra-peripheral* (all neighbors in the same community, $P \leq 0.05$), *peripheral* (a large majority of neighbors in the same community, $P \leq 0.62$), *connectors* (approximately half the neighbors in the same community, $P \leq 0.80$), *kinless* (a large majority of neighbors in other communities, $P > 0.80$).

If we consider our data, we get the distribution represented in Fig. 3, which is rather similar to the results obtained by Guimerà & Amaral on metabolic networks (appendix of [39]). A large majority of nodes have a zero participation coefficient, which means all their neighbors belong to their community. This is of course related to the fact only 4% of the nodes have an embeddedness smaller than 0.5. These nodes only differ in their within community degree, and only a few of them are hubs. Consequently, most of the nodes in our network are ultra-peripheral (84%) or peripheral (12%). Three nodes are non-hub connectors, only one is a connector hub, and we have no kinless hub. The rest (3%) are provincial hubs. This is consistent with the community structure of our network, since non-modular networks exhibits many kinless and very few (ultra-)peripheral nodes [39]. However, it is interesting to notice the hub distribution is not completely compatible with the hub dominance measure. For instance, on the one hand community 20 has the maximal hub dominance, however it does not contain any according to the role approach. On the other hand, community 1 has very low hub dominance, when it contains two hubs, including the only connector of the network. In both cases, the hub dominance might be fooled by the community sizes (very small for the first, much larger for the second). Besides these cases, roles and hub dependence agree on most communities. However, this highlights the fact that, when several alternative tools are available, one should confront their results. Another interesting point is the fact community 16 not only contains one of the two minimal embeddedness nodes, but also one of the three connector non-hubs. This seems to confirm our assumption for this community to be an artifact of the detection algorithm.



(a) (b)

Fig. 3. Distribution of roles (a) in terms of within-community degree z and participation coefficient P; and (b) in the network. The colors are the same than in [39]: grey, red and green for ultra-peripheral, peripheral, and connector non-hubs; yellow and pink for provincial and connector hubs, respectively.

3.2 Attribute-based Interpretation

After having described and analyzed the community structure, one is generally interested in giving a context-dependent interpretation, allowing for instance to explain why or how this structure appeared, or to perform some prediction regarding some data not available at the moment of the study. For this matter, in many situations, one has to focus solely on the topological properties described in the previous section. However, it is sometimes possible to associate tabular data to the studied network, defining various attributes for each node. This is particularly true for domains in which the objects composing the networks are complex enough to need being described according to several informative dimensions (e.g. social sciences). When such information is available, one can discuss the topological properties in terms of nodal attributes, which can help a lot understanding the system. In this section, we present both descriptive and inferential tools adapted to this purpose. Note most of them are implemented in statistical softwares such as SPSS or R, and even Microsoft Excel for the descriptive methods.

3.2.1 Description

The formation of communities, especially in social networks, can sometimes be explained by homophilic relationships, i.e. a tendency for nodes to connect with other nodes more or less similar to them, relatively to some criteria of interest. Let us consider the sequence of all links present in the network: the values of some attribute for the corresponding source and target nodes can be viewed as two distinct series. The homophily can be measured as the level of association between these two series. For instance, Newman proposed to use the Cohen's Kappa statistic and Pearson's correlation coefficient for nominal and numeric attributes, respectively [68]. It is generally processed over the whole network, but in our case it can also be used to characterize the communities: there is no reason for them to exhibit the same homophily. Table 4 shows some results for the gender (G) and class (C) attributes. Most communities have close to zero homophily for gender, except for a few ones for which it reaches a value close to 0.5 (10, 13, 17). This means students do not bond depending on their gender, except for these communities. Homophily values are more contrasted for the class, with values either very close to 0 (3, 9, 15, 17…) or to 1 (8, 20).

Another approach consists of considering the community structure instead of the links as the relevant topological information. Under this assumption, communities are simply groups of nodes one wants to characterize relatively to their attributes. This problem is much more general than network analysis, since it also occurs in classic cluster analysis [69]. As an example, we present in Table 4 some of the most characteristic attributes of our data. Of course, all communities are not characterized by the same attributes, which is why we selected different types of data: factual (class and department), behavioral (hobbies, mobile phones, digital players) and sentimental (best friend consideration and loan inclination).

For space matters, we focus our comments only on a few communities. Let us consider first community 7. It contains only students of 3rd and 4th year of License, but this holds for other communities too (3, 17), so this property alone is not sufficient to characterize it. However, unlike community 17, its dominating department is Business Science. Communities 3 and 7 can be distinguished by considering the former has no dominant hobby, and their dominant mobile phone brands are different. Students from community 15 are more inclined to take a loan, they have the highest average score for that question (LI). They will certainly be the most receptive to commercial pressure. Detecting such a community can have quite huge implications in the Business field.

Table 4. Description of the network and its communities in terms of attributes. The G and C columns represent the homophily for the gender and class attributes, respectively. The Class (resp. Dept.) columns describe the two most represented classes (resp. department) in each community: the left column is the number of concerned students and the right one is the class (resp. department) name. The represented classes are Preparatory (P1-2) and License (L1-4), the departments are Business Science (BS), Computer Science (CS), Economics (Ec), Industrial Engineering (IE), International Relations (IR), Law (La), Literature (Li), Mathematics (Ma), Philosophy (Ph), Sociology (So). H is the most popular hobby: music (M), cinema (C), sport (S), photography (P), reading (R), theater (T). The next two columns are the most widespread brands of mobile phones (MP) and digital players (DP): Nokia (No), Samsung (Sa), Sony-Ericsson (SE), Apple (Ap), Creative (Cr), Sony (So). The two last columns indicate if a student thinks he has his best friends in the university (BF) and his inclination to take a loan (LI), respectively. Both answers are expressed on a scale ranging from 1 (clear no) to 5 (clear yes).

Com.	n	G	C	Class 1		Class 2		Dept.1		H	MP	DP	BF	LI
1	32	0.24	0.30	25	P2	6	L1	7	BS	M	No	Ap	3.80	2.60
2	39	0.36	0.54	15	L2	12	L3	10	IE	C	No	-	3.00	1.78
3	28	0.12	0.00	25	L4	3	L3	12	BS	-	No	-	3.25	2.50
4	30	0.11	–0.04	26	P1	3	P2	6	So	M	No	Cr	2.78	1.78
5	23	0.19	–0.05	17	L2	-	-	15	RI	S	Sa	Ap	2.75	2.25
6	46	0.01	0.65	19	L2	18	L3	14	So	R	Sa	Ap	3.43	1.54
7	34	0.25	0.19	25	L3	9	L4	24	BS	C	Sa	Ap	3.78	2.67
8	23	0.17	0.74	12	P1	10	L1	9	IE	S	Sa	Ap	3.00	3.00
9	20	0.19	0.00	18	P1	2	P2	5	BS	-	No	Ap	2.17	1.67
10	39	0.51	0.55	31	L2	6	L1	17	BS	C	No	Ap	2.92	1.92
11	20	–0.15	0.16	17	L4	2	L3	11	IE	M	No	-	3.80	2.00
12	15	0.11	0.61	7	L1	7	L2	7	La	-	SE	Ap	4.00	1.50
13	28	0.46	0.60	14	L1	13	L2	12	Ma	M	No	So	3.64	1.64
14	13	0.00	0.56	6	L3	5	L4	7	CS	-	No	-	3.50	2.00
15	14	–0.12	0.00	13	L3	-	-	8	RI	-	-	-	2.67	3.50
16	13	–0.10	–0.03	11	L1	-	-	10	Ph	P	-	-	3.67	1.33
17	28	0.48	0.00	25	L4	3	L3	14	IE	C	No	Ap	3.11	2.22
18	22	–0.09	0.54	12	L3	-	-	-	-	-	No	-	3.71	1.17
19	20	–0.06	0.00	19	L1	-	-	9	Ma	-	No	-	3.71	1.71
20	12	–0.14	1.00	12	P1	-	-	2	-	R	No	Ap	2.14	1.71
21	15	–0.16	0.00	13	P2	2	L1	11	La	T	No	-	4.00	1.67
22	38	–0.05	–0.02	34	L1	2	P2	22	BS	C	Sa	Ap	4.00	2.00
Net.	552	0.25	0.78	124	L1	107	L2	96	BS	S	No	Ap	3.29	1.94

Community 16 contains almost exclusively first year Licenses from the philosophy department, which is already discriminant when considering the other communities. Moreover, from an application point of view, it is interesting to notice the dominant hobby is photography and there is no dominant brand for electronic devices. Community 20 is very interesting because its students tend to think their best friends are not in the university (BF column): they have the lowest average score for the corresponding question. Nevertheless, this community is quite similar to others regarding hobbies and brands. This may be due to the fact those students are in first year, often in a new city, far away from their family and high-school friends. A similar observation can be on the communities containing a majority of first year students (e.g. 9), and the effect tends to disappear for the communities of older students (12, 21, 23).

As we shown, the visual inspection of the community composition allows to detect attributes of interest. This inspection can be enhanced by a graphical representation of the network. Fig. 4 gives an example based on the distribution of the class attribute in the network of communities. This figure includes, among other, the columns Class 1 & 2 from Table 4. It confirms our remarks regarding the relatively discriminant power of the class attribute, and the fact it is not enough to uniquely characterize all communities. However, these somewhat subjective observations must be confirmed objectively in order to be relevant and useful. In other terms, one has to assess statistically the significance of the differences observed between the communities. For this matter, the selection of an adapted statistical tool depends on the nature of the attribute of interest.

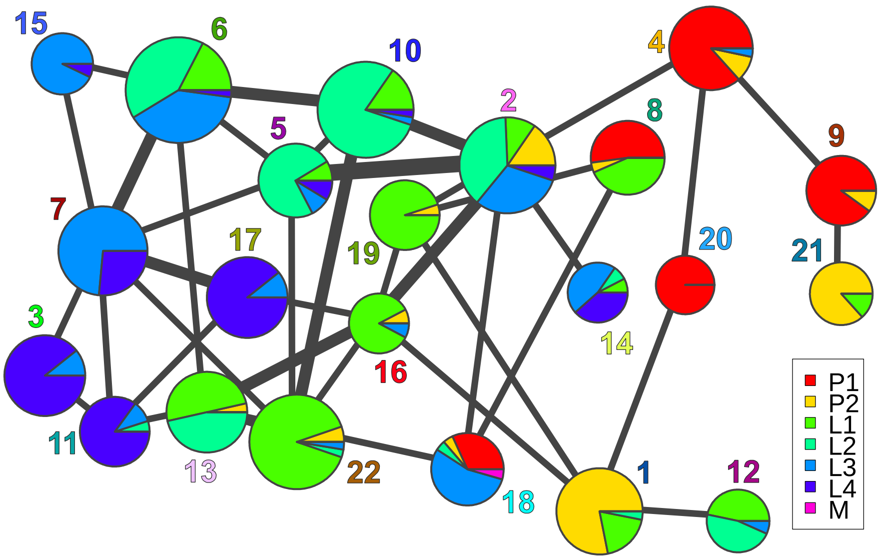


Fig. 4. Class distribution in the community network. Each node represents a community from Fig. 1, with matching number values and colors. Node diameters and link widths are proportional to community sizes (expressed in number of students) and to number of inter-community links, respectively. Each pie chart represents the class attribute distribution in a community. Possible classes are Preparatory (P1-2), License (L1-4) and Master (M).

First, suppose we want to determine if community membership depends on some nominal attribute. In other terms, we want to assess the significance of the association between two nominal variables: the community and the attribute [70]. In this case, the most popular test is the well-known Pearson's chi-square test. Note that extensions exist for tables of higher dimension, allowing to test for association using several attributes. Also, association measures derived from the χ^2 statistic (Pearson's Φ, Cramér's V, etc.) allow quantifying the strength of the association, by opposition to its simple existence. They have been questioned though, and alternatives exist, such as the λ coefficient [71], which has the advantage of being asymmetrical. In our example, the associations between class and department on one side, and community membership on the other side, are very significant ($p < 0.001$), which means those attributes are generally good to characterize our communities.

In the case of a quantitative attribute, one can perform a classic Anova to test whether its means are significantly different across communities [69], under the assumptions of independence, normality and homoscedasticity (variance homogeneous across communities) [70]. Note if several attributes have to be considered at once, an extension called factorial Anova must be used instead. As an example, we performed an Anova on the sentimental attributes (best friend consideration and loan inclination). We first tested for homoscedasticity using Levene's test and got low p-values (respectively 0.068 and 0.085), but not enough to reject the homoscedasticity assumption for $\alpha = 0.05$. For the Anova itself, on the contrary, the p-values were small enough to reject the hypothesis of uniform mean (0.032 and 0.049, respectively). In other words, significant differences exist between communities for both attributes. To identify precisely which communities differ, one has to perform a post-hoc test such as Tuckey's or Least Significant Difference (LSD) tests [70]. We applied the latter to our data, which expose several significant differences, but we limit our comments to the communities on which we focused in this section. It turns out the sentiment of having his best friend at the university is significantly lower in community 20 compared to most others, especially the 16^{th} and 7^{th}, so it can be considered as a characteristic of this community. Students from the community 15 are significantly more inclined to take a loan or to delay a payment than most of the other communities, especially the 16^{th} and 20^{th}, whose students are significantly inclined not to take a loan.

Besides the communities, the nodes of interest detected in the previous section can also be interpreted in terms of nodal attributes. In our data, we highlighted 5 students with very low embeddedness or specific roles (three non hub connectors and one hub connector). We will here only give some points and remarks to illustrate our purpose. First, it is worth noticing two out of three non-hub connectors are girls, and moreover two of them belong to the same community (16) and department (Philosophy). One of them is in 4^{th} year of License. She is an outlier on a question concerning the intention to stay in touch with university friends. Students strongly agreed to this assertion in average, whereas this person clearly thinks the opposite. Moreover, she also states she has a high probability to use old-fashioned products, when she owns cutting edge mobile phone and digital player. This information is of major interest in the context of a marketing strategy, for instance it will allow orienting communication towards social image and acceptance matters. The hub connector is also interesting: he is a boy, in second year of preparatory class in the Law department. Most of his answers to the questions are very close to the average for all the respondents. Nevertheless, contrary to the others, he gives a very high importance to his friends' advice regarding computer and mobile phone purchases. Moreover, contrary to the majority of students, he states he would reduce his other expenditures to be able to afford some products of interest. The marketing strategy has to differ from the case of the previous girl, because he is certainly very well installed socially and possibly aims at keeping a very good social image.

3.2.2 Prediction

The descriptive tools presented in the previous section allow characterizing a community in terms of nodal attributes. This type of analysis is already interesting in terms of interpretation, but predictive methods can bring more precise models regarding the way communities are constituted. First, a model is estimated using the communities as reference groups and taking advantage of the available attributes. Its quality can be assessed in various ways, the simplest being to measure its prediction success rate on instances whose community is known. If the model is considered to fit the data well enough, it can be interpreted by considering which attributes it uses and how it combines them to estimate communities.

We present here two families of statistical tools which can be used to build a predictive model: linear discriminant analysis (LDA) and sigmoid regression. The former was initially designed to predict the value of a nominal variable using numeric attributes, and was later extended to the nominal case under the name of discriminant correspondence analysis. The idea sustaining the method is close to PCA (Principal Component Analysis) and other dimension reduction methods. It consists of projecting the data in a new space maximizing the separation between the communities. The result of the projection is defined by a set of discriminant factors, corresponding to linear combinations of the initial attributes. These factors are then used instead of the attributes to estimate the community of an object. The model is valid under the assumptions of multinormality of the attributes conditionally to the communities and homoscedasticity between communities [70]. Note extensions exist for both non-linear combinations and heteroscedasticity situations.

Two methods exist to derive the discriminant functions: processing all attributes at once (direct approach) or selecting them iteratively (stepwise approach). The second method allows using different criteria [70] to select the attributes and limit their number, it thus results in more parsimonious models. The number of factors is limited by the number of communities and of selected attributes. Each factor can be characterized in terms of its discriminant power, and by interpreting the coefficients associated to the attributes in the corresponding function.

As an example, we tested all the numeric attributes related to our behavioral and sentimental data, which represents a total of 57 attributes. The model obtained with the direct approach has 21discriminant functions and can correctly classify 99.1% of the students. This very high rate has to be nuanced by the fact the model includes many functions, based on all 57 attributes. Obviously, the interpretative value of this model is very weak. We processed separately the behavioral and sentimental attributes, and obtained models based on 21 functions using 31 attributes with a prediction rate of 70.5% for the former, whereas the latter led to 21 functions using 26 attributes with a 69.8% prediction rate. The Anova results of the previous sections were rather promising when considering the discriminant power of the two behavioral and sentimental attributes we tested. However, when considering the discriminant analysis results obtained in this section, it does not seem to be the case for the rest of our data. This suggests both kinds of data do not convey sufficient information to efficiently predict community membership. However, note it is possible to go further, for instance by preprocessing the data to reduce its dimension before performing the discriminant analysis. This could allow improving the readability of the model without losing much predictive power.

The second family of predictive methods is the sigmoid regression, for which one can use two different models: logit or probit. This type of regression is able to predict the value of a dichotomous variable based on numeric and dichotomous variables (its application to nominal variables therefore requires to recode them). It was extended to the prediction of nominal variables, e.g. communities. The two approaches differ mainly in terms of the assumptions and estimation methods they rely on [70]. Probit allows colinearity in the attributes but requires normality, which is not the case of logit. Unlike for discriminant analysis, homoscedasticity is not required.

We applied a multinomial logit regression to the department and class attributes, which are both nominal. The model could be estimated with significantly good fit for both attributes (compared to a null model implementing the hypothesis of no influence of the attributes on the communities). The overall prediction rate is 46.8%, but varies very much depending on the community. For 4 communities (3, 4, 17, 22), it is greater than 80% (with 89.3% as a maximum), and for 9 others (8, 9, 11, 12, 14, 15, 19-21) it is 0%. For the communities we previously focused on (7, 16 and 20) it is of 64.7%, 61.5% and 0%, respectively. This confirms our previous observation: some communities can be efficiently characterized using these factual attributes, but they are not relevant for others. In marketing, this kind of information is at the origin of classic segmentation approaches. In our case, a marketing strategy based only on factual data would have very different effects depending on the targeted communities. It would certainly perform well on communities 3, 4, 17 and 22, but be inefficient on communities such as the 15^{th}. Yet, we previously showed this community was very attractive from a commercial point of view. The fact the network analysis managed to detect this community illustrates how it can be used to complement classic data analysis.

4 Conclusion

In this chapter, we tackled the problem of community detection from the user's point of view. The research is very active in this domain, and so many different tools exist that it is difficult to make an accurate and informed choice. Our aim was to present them, with the will of being as operational as possible. We reviewed the various definitions of the concept of community, and discussed publicly available community detection tools from this perspective. We emphasized other features allowing the user to make an appropriate choice regarding his data and goals, such as the inputs and outputs these tools are able to process. Our goal was to complete the very detailed existing reviews, which already deal with matters concerning the community detection process itself and related computational properties [3,8,9,11]. We also presented practical means of solving secondary problems such as comparing community structures output by different algorithms or corresponding to different levels estimated by a hierarchical algorithm.

We then considered a practical application of community detection to real-world data describing a population of university students. We first concentrated on the topological properties of the network. We chose to ignore general complex network measures, because there again, reviews already describe them in details [63]. Instead, we focused on measures related to the community structures. We illustrated how one can determine the significance of the communities and assess their quality. We also discussed various ways of characterizing individual nodes relatively to the community structure. We then looked at the various methods allowing to take advantage of nodal attributes, which are rather common in some fields such as social sciences. We reviewed descriptive tools and showed how to characterize and interpret the communities. We also illustrated how the application of predictive methods enhances the understanding of the community composition.

However, due to lack of space, we could not perform an exhaustive review and had to discard some methods at each section of our chapter. First, we ignored community detection algorithms able to identify overlapping communities [3,34]. Although there are not many of them yet, compared to those outputting partitions, these approaches are very promising, because many real-world networks include nodes located in-between communities (this was illustrated in the analysis of our data). Second, we only presented general families of definitions of the community concept, when specific variants exist among the hundred community detection algorithms one can find in the literature. The same remark holds for the measures designed to study the significance [72,73] and topological properties [15] of the community structure. Finally, we only mentioned statistical tools in our analysis of the nodal attributes, but some machine learning based approaches are also adapted. For instance, it would be possible to build a very informative predictive model for each community by applying an association rule mining tool [74].

References

1. da Fontura Costa, L., Oliveira Jr., O.N., Travieso, G., Rodrigues, r.A., Villas Boas, P.R., Antiqueira, L., Viana, M.P., da Rocha, L.E.C.: Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. arXiv physics.soc-ph, 0711.3199 (2008).
2. Freeman, L.C.: The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, New York, US-NY (2004)
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75-174 (2010). doi:DOI 10.1016/j.physrep.2009.11.002
4. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 026113 (2004). doi:Artn 026113 Doi 10.1103/Physreve.69.026113
5. Lancichinetti, A., Kivelä, M., Saramäki, J., Fortunato, S.: Characterizing the Community Structure of Complex Networks. PLoS ONE 5(8), e11976 (2010).
6. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Phys Rev E 80(5), 056117 (2009).
7. Labatut, V., Balasque, J.-M.: Business-oriented Analysis of a Social Network of University Students. Paper presented at the ASONAM, Odense, DK,
8. Porter, M.A., Onnela, J.-P., Mucha J., P.: Communities in Networks. arXiv physics.soc-ph, 0902.3788 (2009).
9. Danon, L., Duch, J., Arenas, A., Díaz-Guilera, A.: Community structure identification. In: Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. pp. 93-113. World Scientific, (2007)
10. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E 74(1), 016110 (2006).
11. Newman, M.E.J.: Detecting community structure in networks. Eur Phys J B 38, 321-330 (2004).
12. Mancoridis, S., Mitchell, B.S., Rorres, C., Chen, Y., Gansner, E.R.: Using automatic clustering to produce high-level system organizations of source code. Paper presented at the 6th International Workshop on Program Comprehension, Washington, US-DC,
13. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. PNAS 101(9), 2658-2663 (2004). doi:DOI 10.1073/pnas.0400054101
14. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3), 036106 (2007).
15. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical Properties of Community Structure in Large Social and Information Networks. Paper presented at the WWW, Beijing, CN, Apr
16. Donetti, L., Munoz, M.A.: Detecting network communities: a new systematic and efficient algorithm. J Stat Mech(10), P10012 (2004).
17. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3), 036104 (2006).
18. Lambiotte, R., Delvenne, J.-C., Barahona, M.: Laplacian Dynamics and Multiscale Modular Structure in Networks. arXiv physics.soc-ph, 0812.1770 (2009).
19. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys Rev E 70(6), 066111 (2004).
20. Schuetz, P., Caflisch, A.: Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. Phys Rev E 77(4), 046112 (2008).
21. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J Stat Mech 10, P10008 (2008).
22. Wakita, K., Tsurumi, T.: Finding community structure in mega-scale social networks. arXiv cs.CY, 0702048 (2007).

23. 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).
24. Agarwal, G., Kempe, D.: Modularity-Maximizing Graph Communities via Mathematical Programming. European Physics Journal B 66(3), 409-418 (2008).
25. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. PNAS 104(1), 36-41 (2007).
26. Gleich, D.: Hierarchical Directed Spectral Graph Partitioning. In. Stanford University, (2006)
27. Newman, M.E.J.: Analysis of weighted networks. Phys Rev E 70(5) (2004). doi:Artn 056131 Doi 10.1103/Physreve.70.056131
28. Leicht, E.A., Newman, M.E.J.: Community Structure in Directed Networks. Phys Rev Lett 100(11), 118703 (2008).
29. Hanneman, R.A., Riddle, M.: Introduction to social network methods. University of California, Riverside, US-CA (2005)
30. Luce, R.D.: Connectivity and generalized cliques in sociometric group structure. Psychometrika 15(2), 169–190 (1950).
31. Seidman, S.B.: Network structure and minimum degree. Soc Networks 5(3), 269–287 (1983).
32. Seidman, S.B., Foster, B.L.: A graph theoretic generalization of the clique concept. J Math Sociol 6, 139–154 (1978).
33. Mokken, R.J.: Cliques, clubs and clans. Qual Quant 13, 161–173 (1979).
34. Palla, G., Farkas, I.J., Pollner, P., Derenyi, I., Vicsek, T.: Directed network modules. New Journal of Physics 9, 186 (2007). doi:Artn 186 Doi 10.1088/1367-2630/9/6/186 Pii S1367-2630(07)44249-5
35. Fouss, F., Pirotte, A., Renders, J.-M., Saerens, M.: Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans Knowl Data Eng 19(3), 355-369 (2007).
36. Gan, G.a.M., C. and Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, Philadelphia, US-PA (2007)
37. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, US-NY (1990)
38. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579 (1901).
39. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature 433, 895-900 (2005).
40. Zhou, H.: Network landscape from a Brownian particle's perspective. Phys Rev E 67(4), 041908 (2003).
41. Saerens, M., Fouss, F., Yen, L., Dupont, P.: The principal component analysis of a graph and its relationships to spectral clustering. In: European Conference on Machine Learning, 2004 2004
42. Pons, P., Latapy, M.: Computing communities in large networks using random walks. LNCS 3733, 284-293 (2005).
43. Tong, H., Faloutsos, C., Pan, J.-Y.: Random walk with restart: Fast solutions and applications. Knowl Inf Syst 14(3), 327–346 (2008).
44. Handcock, M.S., Raftery, A.E., Tantrum, J.M.: Model-based clustering for social networks. J Roy Stat Soc A Sta 170, 301-322 (2007).
45. Tyler, R., Wilkinson, D.M., Huberman, B.A.: Email as spectroscopy: Automated discovery of community structure within organizations. In: Deventer, B.V. (ed.) Communities and Technologies. pp. 81-96. Kluwer, (2003)
46. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99(12), 7821-7826 (2002). doi:DOI 10.1073/pnas.1226539799
47. Wu, F., Huberman, B.A.: Finding communities in linear time: a physics approach. Eur Phys J B 38(2), 331-338 (2004). doi:DOI 10.1140/epjb/e2004-00125-x

48. Castellano, C., Cecconi, F., Loreto, V., Parisi, D., Radicchi, F.: Self-contained algorithms to detect communities in networks. Eur Phys J B 38(2), 311–319 (2004).
49. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. PNAS 104(18), 7327-7331 (2007). doi:DOI 10.1073/pnas.0611034104
50. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. PNAS 105(4), 1118 (2008).
51. Ziv, E., Middendorf, M., Wiggins, C.H.: Information-theoretic approach to network modularity. Phys Rev E 71(4) (2005). doi:Artn 046117 Doi 10.1103/Physreve.71.046117
52. van Dongen, S.: Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl 30(1), 121-141 (2008). doi:Doi 10.1137/040608635
53. Donetti, L., Munoz, M.A.: Improved spectral algorithm for the detection of network communities. arXiv physics/0504059v1 (2005).
54. Hofman, J.M., Wiggins, C.H.: Bayesian Approach to Network Modularity. Phys Rev Lett 100(25) (2008).
55. Aldecoa, R., Marin, I.: Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering. PLoS ONE 5(7), e11585 (2010).
56. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal 695(Complex Systems) (2006).
57. O'Madadhain, J., Fisher, D., Smyth, P., White, S., Boey, Y.-B.: Analysis and Visualization of Network Data using. Journal of Statistical Software (2005).
58. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. Paper presented at the International AAAI Conference on Weblogs and Social Media,
59. Rosvall, M., Bergstrom, C.T.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. ArXiv physics.soc-ph, 1010.0431 (2010).
60. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814-818 (2005). doi:Doi 10.1038/Nature03607
61. Rand, W.M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66(336), 846-850 (1971).
62. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2(1), 193-218 (1985).
63. da Fontura Costa, L., Rodrigues, F.A., Travieso, G., Villas Boas, P.R.: Characterization of complex networks: A survey of measurements. Advances in Physics 56(1), 167-242 (2007).
64. Lancichinetti, A., Radicchi, F., Ramasco, J.J.: Statistical significance of communities in networks. Phys Rev E 81(4) (2010).
65. Decaudin, J.M.: La communication Marketing, Concepts, techniques, stratégies. Economica, (2003)
66. Watts, D.C., Dodds, P.S.: Influentials, networks and public opinion formation,. Journal of Consumer Research 34, 441-458 (2007).
67. Kotler, P., Keller, K.L.: Marketing Management- Analysis, planning, implementation and control, 12th ed. Practice Hall International Editions, (2006)
68. Newman, M.E.J.: Mixing patterns in networks. Phys Rev E 67, 026126 (2003).
69. Evrard, Y., Pras, B., Roux, E.: MARKET: Etudes et recherches en Marketing. (2000)
70. Norusis , M.: SPSS 17.0 Guide to Data Analisis. Prentice Hall, Inc., (2008)
71. Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classification. J Am Stat Assoc 49, 732-764 (1954).
72. Rosvall, M., Bergstrom, C.T.: Mapping Change in Large Networks. PLoS ONE 5(1), e8694 (2010). doi:Artn E8694 Doi 10.1371/Journal.Pone.0008694
73. Bianconi, G.a.P., P. and Marsili, M.: Assessing the relevance of node features for network structure. PNAS 106(28), 11433–11438 (2009).
74. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Data Management Systems. Morgan Kaufmann, (2005)

Appendix: Notations

The table bellow summarizes the notations used through this chapter, and indicates the first occurrence of each one of them.

Table 5. Summary of the notations used in this chapter.

Notation	Meaning	1st occurrence
$\langle \blacksquare \rangle$	Average of the quantity $\blacksquare$	Eq. (3)
n	Total number of nodes in the network	Eq. (4)
m	Total number of links in the network	Eq. (7)
n_i	Number of nodes in community i	Eq. (1)
m_{ij}	Number of links between communities i and j	Eq. (1)
q_{ij}	Proportion of links between communities i and j	Eq. (8)
k_i	Number of links between some node and community i	Eq. (4)
e	Embeddedness of some node	Eq. (6)
A_i	Intra-connectivity of community i	Eq. (2)
B_{ij}	Inter-connectivity of communities i and j.	Eq. (2)
${B_{ij}}'$	Alternative inter-connectivity of communities i and j.	Eq. (4)
MQ	Mean quality of a community structure, according to [12]	Eq. (3)
Φ_i	Conductance of community i	Eq. (5)
CV	Coverage of a community structure	Eq. (7)
Q	Modularity of a community structure	Eq. (8)
RI	Rand Index	Eq. (9)
ARI	Adjusted Rand Index	Eq. (10)
B	B-score for community significance	Table 3
d	Density	Table 3
h	Hub dominance of a community	Eq. (11)
ℓ	Average distance	Table 3
k	Degree (number of links) of a node	Table 3
z	Within-community degree of a node relatively to its community	Eq. (13)
P	Participation coefficient of a node to a community structure	Eq. (12)