Advances in Physics,
Vol. 56, No. 1, February 2007, 167–242

Taylor & Francis
Taylor & Francis Group

Characterization of complex networks: A survey of measurements

L. DA F. COSTA*, F. A. RODRIGUES,
G. TRAVIESO and P. R. VILLAS BOAS

Instituto de Física de São Carlos, Universidade de São Paulo,
Caixa Postal 369, 13560-970, São Carlos, SP, Brazil

(*Received 21 August 2006*; *in final form 4 December 2006*)

Each complex network (or class of networks) presents specific topological features which characterize its connectivity and highly influence the dynamics of processes executed on the network. The analysis, discrimination, and synthesis of complex networks therefore rely on the use of measurements capable of expressing the most relevant topological features. This article presents a survey of such measurements. It includes general considerations about complex network characterization, a brief review of the principal models, and the presentation of the main existing measurements. Important related issues covered in this work comprise the representation of the evolution of complex networks in terms of trajectories in several measurement spaces, the analysis of the correlations between some of the most traditional measurements, perturbation analysis, as well as the use of multivariate statistics for feature selection and network classification. Depending on the network and the analysis task one has in mind, a specific set of features may be chosen. It is hoped that the present survey will help the proper application and interpretation of measurements.

Contents

*Corresponding author. Email: luciano@ifsc.usp.br

Advances in Physics
ISSN 0001–8732 print/ISSN 1460–6976 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00018730601170527

1. Introduction

Complex networks research can be conceptualized as lying at the intersection between graph theory and statistical mechanics, which endows it with a truly multidisciplinary nature. While its origin can be traced back to the pioneering works on percolation and random graphs by Flory [1], Rapoport [2–4], and Erdős and Rényi [5–7], research in complex networks became a focus of attention only recently. The main reason for this was the discovery that real networks have characteristics which are not explained by uniformly random connectivity. Instead, networks derived from real data may involve community structure, power law degree distributions and hubs, among other structural features. Three particular developments have contributed to the ongoing related advances: Watts and Strogatz's investigation of small-world networks [8], Barabási and Albert's characterization of scale-free models [9], and Girvan and Newman's identification of the community structures present in many networks (e.g. [10]).

Although graph theory is a well-established and developed area in mathematics and theoretical computer science (e.g., [11, 12]), many of the recent developments in complex networks have taken place in areas such as sociology (e.g., [13, 14]), biology (e.g., [15]) and physics (e.g., [16, 17]). Current interest has focused not only on applying the developed concepts to many real data and situations, but also on studying the dynamic evolution of network topology. Supported by the availability of high performance computers and large data collections, results like the discovery of the scale-free structure of the Internet [18] and of the WWW [19, 20] were of major importance for the increased interest in the new area of complex networks, whose growing relevance has been substantiated by the large number of recent related publications. Reviews of such developments have been presented in four excellent surveys [21–24]; introductory papers [17, 25–27]; several proceedings [28–33]; edited books [16, 34]; and books related to random graphs [11, 35, 36], complex networks theory [37], scientific dissemination [38–44], social networks [45–54], economic systems and political networks [55–58], and WWW and Internet [59–62]. For additional information about the related areas of percolation, disordered systems and fractals see [63–65]; for complex systems see [66, 67].

One of the main reasons behind complex networks popularity is their flexibility and generality for representing virtually any natural structure, including those undergoing dynamic changes of topology. As a matter of fact, every discrete structure such as lists, trees, or even lattices, can be suitably represented as special cases of graphs. It is thus of little surprise that several investigations into complex networks involve the representation of the structure of interest as a network, followed by an analysis of

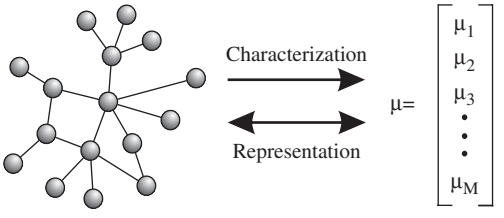


Figure 1. The mapping from a complex network into a feature vector. Generic mappings can be used in order to obtain the characterization of the network in terms of a suitable set of measurements. In case the mapping is invertible, we have a complete representation of the original structure.

the topological features of the obtained representation performed in terms of a set of informative measurements. Another interesting problem consists of measuring the structural properties of evolving networks in order to characterize how the connectivity of the investigated structures changes along the process.

Both such activities can be understood as directed to the *topological characterization* of the studied structures. Another related application is to use the obtained measurements in order to identify different categories of structures, which is directly related to the area of *pattern recognition* [68, 69]. Even when modeling networks, it is often necessary to compare the realizations of the model with real networks, which can be done in terms of the respective measurements. Provided the measurements are comprehensive (ideally the representation by the measurements should be one-to-one or invertible), the fact that the simulated networks yield measurements similar to those of the real counterparts supports the validity of the model.

Particular attention has recently been focused on the relationship between the structure and dynamics of complex networks, an issue which has been covered in two excellent comprehensive reviews [21, 23]. However, relatively little attention has been given to the equally important subject of network measurements (e.g. [70]). Indeed, it is only by obtaining informative quantitative features of the networks topology that they can be characterized and analyzed and, particularly, their structure can be fully related to the respective dynamics. The quantitative description of the networks properties also provides fundamental subsidies for classifying theoretical and real networks into major categories. The present survey's main objective is to provide a comprehensive and accessible review of the main measurements which can be used to quantify important properties of complex networks.

Network measurements are therefore essential as a direct or subsidiary resource in many network investigations, including representation, characterization, classification and modeling. Figure 1 shows the mapping of a generic complex network into the feature vector $\vec{\mu}$, i.e. a vector of related measurements such as average vertex degree, average clustering coefficient, the network diameter, and so on (see sections 18 and 19 for more details about the characterization and classification of real networks). In case the mapping is invertible, in the sense that the network can be recovered from the feature vector, the mapping is said to provide a *representation* of the network (e.g. [68]). An example of invertible mapping for unweighted networks is the adjacency matrix (see section 2). Note, however, that the characterization and

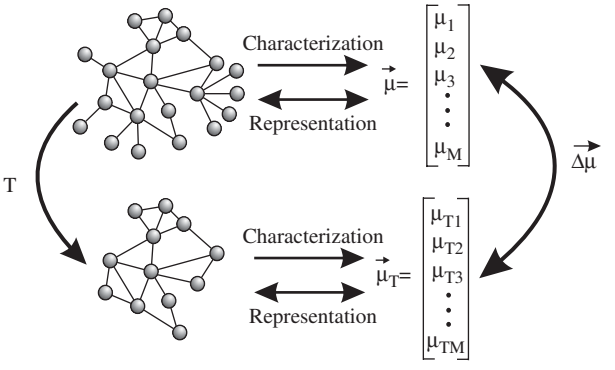


Figure 2. Additional measurements of a complex network can be obtained by applying a transformation T on it and obtaining a new feature vector $\vec{\mu}_T$. The difference $\Delta\vec{\mu}$ between the original and transformed feature vectors can also be considered in order to obtain additional insights about the properties of the original network.

classification of networks does not necessarily require invertible measurements. An interesting strategy which can be used to obtain additional information about the structure of complex networks, involves applying a transformation to the original network and obtaining the measurements from the resulting network, as illustrated in figure 2. In this figure, a transformation T (in this case, deletion of the vertices adjacent to just one other vertex) is applied over the original network to obtain a transformed structure from which new measurements $\vec{\mu}_T$ are extracted. In case the feature vectors $\vec{\mu}$ and $\vec{\mu}_T$ correspond to the same type of measurements, it is also possible to consider the difference between these two vectors in order to obtain additional information about the network under analysis as well as the effects of the transformation.

Perturbations of networks, which can be understood as a special case of the transformation framework outlined above, can also be used to investigate the *sensitivity* of the measurements. Informally speaking, if the measurements considered in the feature vector are such that small changes of the network topology (e.g., add/remove a few edges or vertices) imply large changes in the measurements (large values of $\|\Delta\vec{\mu}\|$), those measurements can be considered as being highly sensitive or unstable. One example of such an unstable measurement for some networks is the average shortest path length between two vertices (see section 16.2).

Another possibility to obtain a richer set of measurements involves the consideration of several instances around the development/growth of the network. A feature vector $\vec{\mu}(t)$ is obtained at each "time" instant t along the growth. Figure 3 shows four instances of an evolving network and the respective trajectory defined in one of the possible feature (or phase) spaces involving two generic measurements μ_1 and μ_2. In such a way, the evolution of a network can now be investigated in terms of a trajectory in a feature's space. Such concepts are presented in more detail in section 16.1, including several models of complex networks.

Both the characterization and classification of natural and human-made structures using complex networks instigate the same important question of *how to choose the most appropriate measurements*. While such a choice should reflect the specific

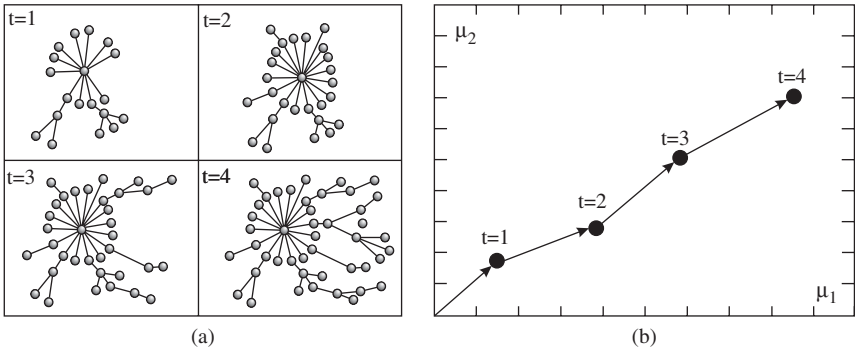


Figure 3. Given a network undergoing some dynamic evolution (a) and a set of measurements (e.g., μ_1 and μ_2), trajectories can be defined in the feature's space (b).

interests and application, it is unfortunate that there is no mathematical procedure for identifying the best measurements. There is an unlimited set of topological measurements, and they are often correlated, implying redundancy in most of the cases. Statistical approaches to decorrelation (e.g., principal component analysis and canonical analysis) can help select and enhance measurements (see section 18), but are not guaranteed to produce optimal results (e.g. [68, 69]). Ultimately, one has to rely on her/his knowledge of the problem and available measurements in order to select a suitable set of features to be considered. For such reasons, it is of paramount importance to have a good knowledge not only of the most representative measurements, but also of their respective properties and interpretation. Although a small number of topological measurements, namely the average vertex degree, clustering coefficient and average shortest path length, were typically considered for complex network characterization during the initial stages of this area, a series of new and more sophisticated features have been proposed and used in the literature along the last years. The fast pace of developments and new results reported in this very dynamic area makes it particularly difficult to follow and to organize the existing measurements.

This review starts by presenting the basic concepts and notation in complex networks and follows by presenting several topological measurements. Illustrations of some of these measurements relating to Erdős-Rényi, Watts-Strogatz, Barabási-Albert, modular and geographical models are also included. The measurements are presented in sections organized according to their main types, including distance-based measurements, clustering coefficients, degree correlations, entropies, centrality, subgraphs, spectral analysis, community-based measurements, hierarchical measurements, and fractal dimensions. A representative set of such measurements is applied to the five considered models and the results are presented and discussed in terms of their cross-correlations and trajectories. The important subjects of measurement selection and assignment of categories to given complex networks are then covered from the light of formal multivariate pattern recognition, including the illustration of such a possibility by using canonical projections and Bayesian decision theory. Table 5 summarizes the main measurements covered in the present survey, as well as the respective symbols and equation numbers.

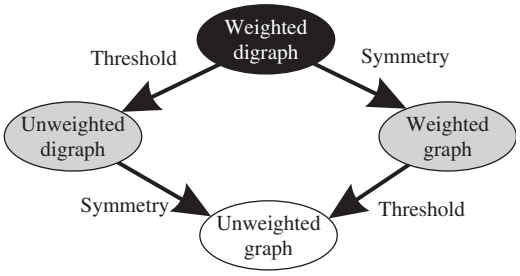


Figure 4. The four main types of complex networks and their transformations. All network types can be derived from the weighted digraph through appropriate transformations.

2. Basic concepts

Figure 4 shows the four main types of complex networks, which include weighted digraphs (directed graphs), unweighted digraphs, weighted graphs and unweighted graphs. The operation of *symmetry* can be used to transform a digraph into a graph, and the operation of *thresholding* can be applied to transform a weighted graph into its unweighted counterpart. These types of graphs and operations are defined more formally in the following, starting from the concept of weighted digraph, from which all the other three types can be derived.

A *weighted directed graph*, G, is defined by a set $\mathcal{N}(G)$ of N *vertices* (or *nodes*), a set $\mathcal{E}(G)$ of M *edges* (or *links*), and a mapping ω: $\mathcal{E}(G) \mapsto \mathbb{R}$. Each vertex can be identified by an integer value $i = 1, 2, \ldots, N$; and each edge can be identified by a pair (i, j) that represents a connection going from vertex i to vertex j to which a weight $\omega(i, j)$ is associated. In the complex network literature, it is often assumed that no self-connections or multiple connections exist; i.e. there are no edges of the form (i, i) and for each pair of edges (i_1, j_1) and (i_2, j_2) it holds that $i_1 \neq i_2$ or $j_1 \neq j_2$ Graphs with self- or duplicate connections are sometimes called *multigraphs*, or *degenerate* graphs. Only non-degenerate graphs are considered henceforth. In an *unweighted digraph*, the edges have no weight, and the mapping ω is not needed. For *undirected graphs* (weighted or unweighted), the edges have no directions; the presence of a edge (i, j) in $\mathcal{E}(G)$ thus means that a connection exist from i to j and from j to i.

A weighted digraph can be completely represented in terms of its *weight matrix* W, so that each element $w_{ij} = \omega(i, j)$ expresses the weight of the connection from vertex i to vertex j. The operation of *thresholding* can be applied to a weighted digraph to produce an unweighted counterpart. This operation, henceforth represented as $\delta_T(W)$, is applied to each element of the matrix W, yielding the matrix $A = \delta_T(W)$. The elements of the matrix A are computed comparing the corresponding elements of W with a specified threshold T; in case $|w_{ij}| > T$ we have $a_{ij} = 1$, otherwise $a_{ij} = 0$. The resulting matrix A can be understood as the *adjacency matrix* of the unweighted digraph obtained as a result of the thresholding operation. Any weighted digraph can be transformed into a graph by using the *symmetry* operation $\sigma(W) = W + W^T$, where W^T is the transpose of W.

For undirected graphs, two vertices i and j are said to be *adjacent* or *neighbors* if $a_{ij} \neq 0$. For directed graphs, the corresponding concepts are those of *predecessor* and

successor: if $a_{ij} \neq 0$ then i is a predecessor of j and j is a successor of i. The concept of adjacency can also be used in digraphs by considering predecessors and successors as adjacent vertices. The *neighborhood* of a vertex i, henceforth represented as $\nu(i)$, corresponds to the set of vertices adjacent to i.

The *degree of a vertex i*, hence k_i, is the number of edges connected to that vertex, i.e. the cardinality of the set $\nu(i)$ (in the physics literature, this quantity is often called "connectivity" [22]). For undirected networks it can be computed as

$$k_i = \sum_j a_{ij} = \sum_j a_{ji}. \tag{1}$$

The *average degree* of a network is the average of k_i for all vertices in the network,

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \frac{1}{N} \sum_{ij} a_{ij}. \tag{2}$$

In the case of directed networks, there are two kinds of degrees: the *out-degree*, k_i^{out}, equal to the number of outgoing edges (i.e. the cardinality of the set of successors), and the *in-degree*, k_i^{in}, corresponding to the number of incoming edges (i.e. the cardinality of the set of predecessors),

$$k_i^{\text{out}} = \sum_j a_{ij}, \tag{3}$$

$$k_i^{\text{in}} = \sum_j a_{ji}. \tag{4}$$

Note that in this case the total degree is defined as $k_i = k_i^{\text{in}} + k_i^{\text{out}}$. The average in- and out-degrees are the same (the network is supposed isolated)

$$\langle k^{\text{out}} \rangle = \langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{ij} a_{ij}. \tag{6}$$

For weighted networks, the definitions of degree given above can be used, but a quantity called *strength* of i, s_i, defined as the sum of the weights of the corresponding edges, is more generally used [71]:

$$s_i^{\text{out}} = \sum_j w_{ij}, \tag{7}$$

$$s_i^{\text{in}} = \sum_j w_{ji}. \tag{8}$$

In the general case, two vertices of a complex network are not adjacent. In fact, most of the networks of interest are sparse, in the sense that only a small fraction of all possible edges are present. Nevertheless, two non-adjacent vertices i and j can be connected through a sequence of m edges $(i, k_1), (k_1, k_2), \ldots, (k_{m-1}, j)$; such set of edges is called a *walk* between i and j, and m is the *length* of the walk. We say that two vertices are *connected* if there is at least one walk connecting them. A *loop* or *cycle* is defined as a walk starting and terminating in the same vertex i and passing only once through each vertex k_n. In case all the vertices and edges along a walk are distinct, the walk is a *path*. Many measurements are based on the length of such connecting paths (see section 4).

Table 1. List of basic symbols used in the text.

Symbol	Concept
$\mathcal{N}(G)$	Set of vertices of graph G
$\mathcal{E}(G)$	Set of edges of graph G
$\lvert\mathcal{X}\rvert$	Cardinality of set $\mathcal{X}$
N	Number of vertices, $\lvert\mathcal{N}(G)\rvert$
M	Number of edges, $\lvert\mathcal{E}(G)\rvert$
W	Weight matrix
w_{ij}	Element of the weight matrix
A	Adjacency matrix
a_{ij}	Element of the adjacency matrix
k_i	Degree of vertex i
k_i^{out}	Out-degree of vertex i
k_i^{in}	In-degree of vertex i
s_i	Strength of vertex i
s_i^{out}	Out-strength of vertex i
s_i^{in}	In-strength of vertex i
$\nu(i)$	Set of neighbors of vertex i
$\lVert X\rVert$	Sum of the elements of matrix X

In undirected graphs, if vertices i and j are connected and vertices j and k are connected, then i and k are also connected. This property can be used to partition the vertices of a graph in non-overlapping subsets of connected vertices. These subsets are called *connected components* or *clusters*.

If a network has too few edges, i.e. the average connectivity of its vertices $\langle k\rangle$ is too small, there will be many isolated vertices and clusters with a small number of vertices. As more edges are added to the network, the small clusters are connected to larger clusters; after some critical value of the connectivity, most of the vertices are connected into a giant cluster, characterizing the percolation [63] of the network. For the Erdős-Rényi graph (see section 3.1) in the limit $N \to \infty$ this happens at $\langle k\rangle = 1$ [35]. Of special interest is the distribution of sizes of the clusters in the percolation point and the fraction of vertices in the giant cluster. The critical density of edges (as well as average and standard deviation) needed to achieve percolation can be used to characterize network models or experimental phenomena. Table 1 lists the basic symbols used in the paper.

3. Complex network models

With the intent of studying the topological properties of real networks, several network models have been proposed. Some of these models have become subject of great interest, including random graphs, the small-world model, the generalized random graph and Barabási-Albert networks. Other models have been applied to the study of the topology of networks with some specific features, as geographical networks and networks with community structure. A comprehensive review of the

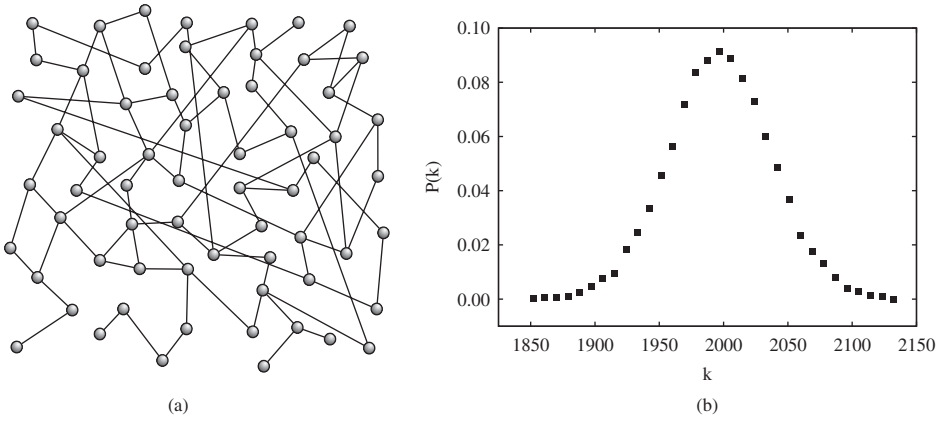


Figure 5. The random graph of Erdős and Rényi: (a) an example and (b) average degree distribution over 10 random networks formed by 10,000 vertices using a probability $p = 0.2$.

various models is not intended here. Instead, the next subsections present some models used in the discussion on network measurements (sections 16, 17, 18 and 19).

3.1. *The random graph of Erdős and Rényi*

The random graph developed by Rapoport [2–4] and independently by Erdős and Rényi [5–7] can be considered the most basic model of complex networks. In their 1959 paper [5], Erdős and Rényi introduced a model to generate random graphs consisting of N vertices and M edges. Starting with N disconnected vertices, the network is constructed by the addition of M edges at random, avoiding multiple and self connections. Another similar model defines N vertices and a probability p of connecting each pair of vertices. The latter model is widely known as Erdős-Rényi (ER) model. Figure 5(a) shows an example of this type of network.

For the ER model, in the large network size limit $N \to \infty$, the average number of connections of each vertex $\langle k \rangle$, given by

$$\langle k \rangle = p(N-1), \tag{9}$$

diverges if p is fixed. Instead, p is chosen as a function of N to keep $\langle k \rangle$ fixed: $p = \langle k \rangle / (N-1)$. For this model, $P(k)$ (the degree distribution, see section 6) is a Poisson distribution (see figure 5(b) and table 2).

3.2. *The small-world model of Watts and Strogatz*

Many real world networks exhibit what is called the *small world* property, i.e. most vertices can be reached from the others through a small number of edges. This characteristic is found, for example, in social networks, where everyone in the world can be reached through a short chain of social acquaintances [39, 44]. This concept originated from the famous experiment made by Milgram in 1967 [72], who

Table 2. Analytical result of some basic measurements for the Erdős-Rényi, Watts-Strogatz and Barabási-Albert network models.

Measurement	Erdős-Rényi	Watts-Strogatz	Barabási-Albert
Degree distribution	$P(k) = \frac{e^{-\langle k\rangle}\langle k\rangle^k}{k!}$	$P(k) = \sum_{i=1}^{\min(k-\kappa,\kappa)} \binom{\kappa}{i} (1-p)^i p^{\kappa-i} \frac{(p\kappa)^{k-\kappa-i}}{(k-\kappa-i)!} e^{-p\kappa}$	$P(k) \sim k^{-3}$
Average vertex degree	$\langle k\rangle = p(N-1)$	$\langle k\rangle = 2\kappa^{\star}$	$\langle k\rangle = 2m$
Clustering coefficient	$C = p$	$C(p) \sim \frac{3(\kappa-1)}{2(2\kappa-1)}(1-p)^3$	$C \sim N^{-0.75}$
Average path length	$\ell \sim \frac{\ln N}{\ln\langle k\rangle}$	$\ell(N,p) \sim p^{\mathcal{T}} f(Np^{\mathcal{T}})^{*}$	$\ell \sim \frac{\log N}{\log(\log N)}$

⋆In WS networks, the value κ represents the number of neighbors of each vertex in the initial regular network (in figure 6, $\kappa = 4$).
*The function $f(u) =$ constant if $u \ll 1$ or $f(u) = \ln(u)/u$ if $u \gg 1$.

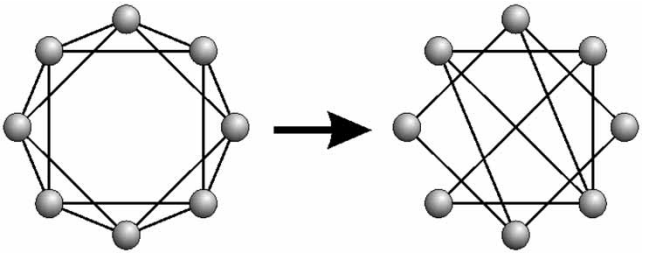


Figure 6. The construction of a small-word network according to Watts and Strogatz: A regular network has its edges rewired with probability p. For $p \approx 0$ the network is regular, with many triangles and large distances, for $p \approx 1$, it becomes a random network, with small distances and few triangles.

found that two US citizens chosen at random were connected by an average of six acquaintances.

Another property of many networks is the presence of a large number of loops of size three, i.e. if vertex i is connected to vertices j and k, there is a high probability of vertices j and k being connected (the clustering coefficient, section 5, is high); for example, in a friendship network, if B and C are friends of A, there is a high probability that B and C are also friends. ER networks have the small world property but a small average clustering coefficient; on the other hand, regular networks with the second property are easy to construct, but they have large average distances. The most popular model of random networks with small world characteristics and an abundance of short loops was developed by Watts and Strogatz [8] and is called the Watts-Strogatz (WS) *small-world model*. They showed that small-world networks are common in a variety of realms ranging from the *C. elegans* neuronal system to power grids. This model is situated between an ordered finite lattice and a random graph presenting the small world property and high clustering coefficient.

To construct a small-word network, one starts with a regular lattice of N vertices (figure 6) in which each vertex is connected to κ nearest neighbors in each direction, totalizing 2κ connections, where $N \gg \kappa \gg \log(N) \gg 1$. Next, each edge is randomly rewired with probability p. When $p = 0$ we have an ordered lattice with high number of loops but large distances and when $p \to 1$, the network becomes a random graph with short distances but few loops. Watts and Strogatz have shown that, in an intermediate regime, both short distances and a large number of loops are present. Figure 7(a) shows an example of a Watts-Strogatz network. Alternative procedures to generate small-world networks based on addition of edges instead of rewiring have been proposed [73, 74], but are not discussed here.

The degree distribution for small-world networks is similar to that of random networks, with a peak at $\langle k \rangle = 2\kappa$ (see also table 2 and figure 7(b)).

3.3. *Generalized random graphs*

A common way to study real networks is to compare their characteristics with the values expected for similar random networks. As the degrees of the vertices are

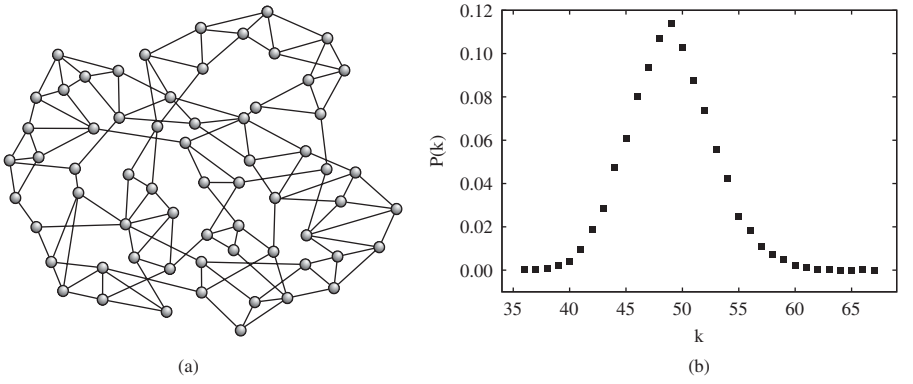


Figure 7. The small-world model of Watts and Strogatz: (a) an example of a network with $N = 64$ vertices, $\kappa = 2$, $p = 0.1$, and (b) average degree distribution over 10 WS networks with 10,000 vertices, $\kappa = 25$ and $p = 0.3$.

important features of the network, it is interesting to make the comparison with networks with the same degree distribution. Models to generate networks with a given degree distribution, while being random in other aspects, have been proposed.

Bender and Canfield [75] first proposed a model to generate random graphs with a pre-defined degree distribution called *configuration model*. Later, Molloy and Reed [76, 77] proposed a different method that produces multigraphs (i.e. loops and multiple edges between the same pair of vertices are allowed).

The common method used to generate this kind of random graph involves selecting a degree sequence specified by a set $\{k_i\}$ of degrees of the vertices drawn from the desired distribution $P(k)$. Afterwards, to each vertex i is associated a number k_i of ''stubs'' or ''spokes'' (ends of edges emerging from a vertex) according to the desired degree sequence. Next, pairs of such stubs are selected uniformly and joined together to form an edge. When all stubs have been used up, a random graph that is a member of the ensemble of graphs with that degree sequence is obtained [78–80].

Another possibility, the *rewiring method*, is to start with a network (possibly a real network under study) that already has the desired degree distribution, and then iteratively choose two edges and interchange the corresponding attached vertices [81]. This rewiring procedure is used in some results presented in section 16.2.

Due to its importance and amenability to analytical treatment, many works deal with this model, including the papers of Newman [23], Aiello *et al.* [82], Chung and Lu [83] and Cohen and Havlin [84].

3.4. *Scale-free networks of Barabási and Albert*

After Watts and Strogatz's model, Barabási and Albert [9] showed that many real systems are characterized by an uneven distribution. Instead of the vertices of these networks having a random pattern of connections with a characteristic degree, as with the ER and WS models (see figures 5 and 7), some vertices are highly connected

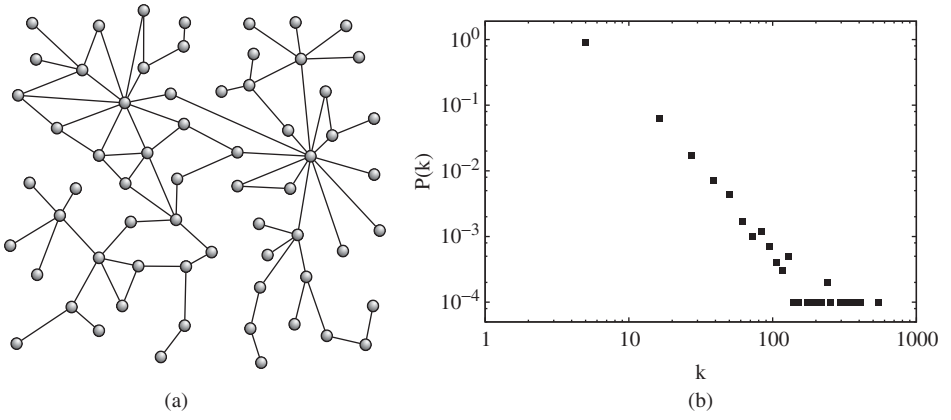


Figure 8. The scale-free network of Barabási and Albert. (a) an example and (b) average degree distribution over 10 Barabási-Albert networks formed by 10,000 vertices using $m = 5$. The degree distribution follows a power law, in contrast to that presented in figure 5.

while others have few connections, with the absence of a characteristic degree. More specifically, the degree distribution has been found to follow a power law for large k,

$$P(k) \sim k^{-\gamma} \tag{10}$$

(see figure 8(b)). These networks are called *scale-free* networks.

A characteristic of this kind of network is the existence of *hubs*, i.e. vertices that are linked to a significant fraction of the total number of edges of the network.

The Barabási-Albert (BA) network model is based on two basic rules: *growth* and *preferential attachment*. The network is generated starting with a set of m_0 vertices; afterwards, at each step of the construction the network *grows* with the addition of new vertices. For each new vertex, m new edges are inserted between the new vertex and some previous vertices. The vertices which receive the new edges are chosen following a *linear preferential attachment* rule, i.e. the probability of the new vertex i to connect with an existing vertex j is proportional to the degree of j,

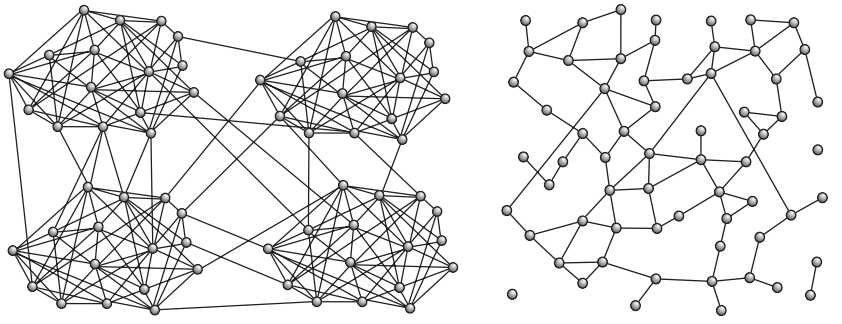
$$\mathcal{P}(i \to j) = \frac{k_j}{\sum_u k_u}. \tag{11}$$

Thus, the most connected vertices have greater a probability of receiving new vertices. This is known as "the rich get richer" paradigm.

Figure 8(a) shows an example of a Barabási-Albert network.

3.5. *Networks with community structure*

Some real networks, such as social and biological networks, present modular structure [10]. These networks are formed by sets or *communities* of vertices such that most connections are found between vertices inside the same community, while connections between vertices of different communities are less common. A model to generate networks with this property was proposed by Girvan and Newman [10].



(a) (b)

Figure 9. (a) An example of a random network with community structure formed by 64 vertices with 4 communities. (b) An example of geographical network formed by 64 vertices.

This model is a kind of random graph constructed with different probabilities. Initially, a set of N vertices is classified into c communities. At each following step, two vertices are selected and linked with probability p_{in}, if they are in the same community, or p_{out}, if they are in different communities. The values of p_{in} and p_{out} should be chosen so as to generate networks with the desired sharpness in the distinction of the communities. When $p_{\text{out}} \ll p_{\text{in}}$, the communities can be easily identified. On the other hand, when $p_{\text{out}} \approx p_{\text{in}}$, the communities become blurred.

Figure 9(a) presents a network generated by using the procedure above.

3.6. *Geographical models*

Complex networks are generally considered as lying in an abstract space, where the position of vertices has no particular meaning. In the case of several kinds of networks, such as protein-protein interaction networks or networks of movie actors, this consideration is reasonable. However, there are many networks where the position of vertices is particularly important as it influences the network's evolution. This is the case for highway networks or the Internet, for example, where the position of cities and routers can be localized in a map and the edges between them correspond to real physical entities, such as roads and optical fibers [85]. This kind of network is called *geographical* or *spatial* network. Other important examples of geographical networks are power grids [86, 87], airport networks [88–90], subway [91] and neural networks [92].

In geographical networks, the existence of a direct connection between vertices can depend on a lot of constraints such as the distance between them, geographical accidents, available resources to construct the network, territorial limitation and so on. The models considered to represent these networks should take these constraints into account.

A simple way to generate geographical networks, used in the results described in sections 16, 17, 18 and 19 is to distribute N vertices at random in a two-dimensional

space Ω and link them with a given probability which decays with the distance, for instance

$$\mathcal{P}(i \to j) \sim e^{-\lambda s_{ij}}, \tag{12}$$

where s_{ij} is the geographical distance of the vertices and λ fixes the length scale of the edges. This model generates a Poisson degree distribution as observed for random graphs and can be used to model road networks (see figure 9(b)). Alternatively, the network development might start with a few nodes while new nodes and connections are added at each subsequent time step (spatial growth). Such a model is able to generate a wide range of network topologies including small-world and linear scale-free networks [93].

4. Measurements related to distance

For undirected, unweighted graphs, the number of edges in a path connecting vertices i and j is called the *length* of the path. A *geodesic path* (or *shortest path*), between vertices i and j, is one of the paths connecting these vertices with minimum length (many geodesic paths may exist between two vertices); the length of the geodesic paths is the *geodesic distance* d_{ij} between vertices i and j. If the graph is weighted, the same definition can be used, but generally one is interested in taking into account the edge weights. Two main possibilities include: first, the edge weights may be proportionally related to some physical distance, for example if the vertices correspond to cities and the weights to distances between these cities through given highways. In this case, one can compute the distance along a path as the sum of the weights of the edges in the path. Second, the edge weights may reflect the strength of connection between the vertices, for example if the vertices are Internet routers and the weights are the bandwidth of the edges, the distance corresponding to each edge can be taken as the reciprocal of the edge weight, and the path length is the sum of the reciprocal of the weight of the edges along the path. If there are no paths from vertex i to vertex j, then $d_{ij} = \infty$. For digraphs, the same definitions can be used, but in general $d_{ij} \neq d_{ji}$, as the paths from vertex i to vertex j are different from the paths from j to i.

Distance is an important characteristic that depends on the overall network structure. The following describes some measurements based on vertex distance.

4.1. *Average distance*

We can define a network measurement by computing the mean value of d_{ij}, known as *average geodesic distance*:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \tag{13}$$

A problem with this definition is that it diverges if there are unconnected vertices in the network. To circumvent this problem, only connected pairs of vertices are included in the sum. This avoids the divergence, but introduces a distortion for networks with many unconnected pairs of vertices, which will show a small value of average distance, expected only for networks with a high number of connections.

Latora and Marchiori [94] proposed a closely related measurement that they called *global efficiency*:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{14}$$

where the sum takes all pairs of vertices into account. This measurement quantifies the efficiency of the network in sending information between vertices, assuming that the efficiency for sending information between two vertices i and j is proportional to the reciprocal of their distance. The reciprocal of the global efficiency is the *harmonic mean* of the geodesic distances:

$$h = \frac{1}{E}. \tag{15}$$

As equation (15) does not present the divergence problem of equation (13), it is therefore a more appropriate measurement for graphs with more than one connected component.

The determination of shortest distances in a network is only possible with global information on the structure of the network. This information is not always available. When global information is unavailable, navigation in a network must happen using limited, local information and a specific algorithm. The effective distance between two vertices is thus generally larger than the shortest distance, and dependent on the algorithm used for navigation as well as network structure [95].

4.2. *Vulnerability*

In infrastructure networks (like WWW, the Internet, energy supply, etc), it is important to know which components (vertices or edges) are crucial to optimum functioning. Intuitively, the critical vertices of a network are their hubs (vertices with higher degree), however there are situations in which they are not necessarily the most vital for the performance of the system which the network underlies. For instance, all vertices of a network in the form of a binary tree have equal degree, therefore there is no hub, but disconnection of vertices closer to the root and the root itself have a greater impact than of those near the leaves. This suggests that networks have a hierarchical property, which means that the most crucial components are those in higher positions in the hierarchy.

A way to find critical components of a network is by looking for the most vulnerable vertices. If we associate the performance of a network with its global efficiency, equation (14), the *vulnerability* of a vertex can be defined as the drop in performance when the vertex and all its edges are removed from the network [96]

$$V_i = \frac{E - E_i}{E}, \tag{16}$$

where E is the global efficiency of the original network and E_i is the global efficiency after the removal of the vertex i and all its edges. As suggested by Gol'dshtein *et al.* [96], the ordered distribution of vertices with respect to their vulnerability V_i is related to the network hierarchy, thus the most vulnerable (critical) vertex occupies the highest position in the network hierarchy.

A measurement of network vulnerability [97] is the maximum vulnerability for all of its vertices:

$$V = \max_i \; V_i. \tag{17}$$

5. Clustering and cycles

A characteristic of the Erdős-Rényi model is that the local structure of the network near a vertex tends to be a tree. More precisely, the probability of loops involving a small number of vertices goes to 0 in the large network size limit. This is in marked contrast with the profusion of short loops which appear in many real-world networks. Some measurements proposed to study the cyclic structure of networks and the tendency to form sets of tightly connected vertices are described in the following.

5.1. *Clustering coefficients*

One way to characterize the presence of loops of order three is through the *clustering coefficient*.

Two different clustering coefficients are frequently used. The first, also known as *transitivity* [98], is based on the following definition for undirected unweighted networks:

$$C = \frac{3N_\Delta}{N_3}, \tag{18}$$

where N_Δ is the number of triangles in the network and N_3 is the number of connected triples. The factor three accounts for the fact that each triangle can be seen as consisting of three different connected triples, one with each of the vertices as central vertex, and assures that $0 \leq C \leq 1$. A triangle is a set of three vertices with edges between each pair of vertices; a connected triple is a set of three vertices where each vertex can be reached from each other (directly or indirectly), i.e. two vertices must be adjacent to another vertex (the central vertex). Therefore we have

$$N_\Delta = \sum_{k>j>i} a_{ij}a_{ik}a_{jk}, \tag{19}$$

$$N_3 = \sum_{k>j>i} (a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}), \tag{20}$$

where the a_{ij} are the elements of the adjacency matrix A and the sum is taken over all triples of distinct vertices i, j, and k only one time.

The second definition of the clustering coefficient of a given vertex i [8] is given by:

$$C_i = \frac{N_\Delta(i)}{N_3(i)}, \tag{21}$$

where $N_\Delta(i)$ is the number of triangles involving vertex i and $N_3(i)$ is the number of connected triples having i as the central vertex:

$$N_\Delta(i) = \sum_{k>j} a_{ij}a_{ik}a_{jk}, \tag{22}$$

$$N_3(i) = \sum_{k>j} a_{ij}a_{ik}, \tag{23}$$

If k_i is the number of neighbors of vertex i, then $N_3(i) = k_i(k_i-1)/2$. $N_\Delta(i)$ counts the number of edges between neighbors of i. Representing the number of edges between neighbors of i as l_i, equation (21) can be rewritten as:

$$C_i = \frac{2l_i}{k_i(k_i-1)}. \tag{24}$$

Using C_i, an alternative definition of the network clustering coefficient (different from that in equation (18)) is

$$\tilde{C} = \frac{1}{N}\sum_i C_i. \tag{25}$$

The difference between the two definitions is that the average in equation (18) gives the same weight to each triangle in the network, while equation (25) gives the same weight to each vertex, resulting in different values because vertices of higher degree are possibly involved in a larger number of triangles than vertices of smaller degree.

For weighted graphs, Barthélemy *et al.* [71] introduced the concept of *weighted clustering coefficient* of a vertex,

$$C_i^w = \frac{1}{s_i(k_i-1)}\sum_{k>j}\frac{w_{ij}+w_{ik}}{2}a_{ij}a_{ik}a_{jk}, \tag{26}$$

where the normalizing factor $s_i(k_i-1)$ (s_i is the strength of the vertex, see section 2) assures that $0 \le C_i^w \le 1$. From this equation, a possible definition of clustering coefficient for weighted networks is

$$C^w = \frac{1}{N}\sum_i C_i^w. \tag{27}$$

Another definition for clustering in weighted networks [99] is based on the intensity of the triangle subgraphs (see section 12.2),

$$\tilde{C}_i^w = \frac{2}{k_i(k_i-1)}\sum_{k>j}(\hat{w}_{ij}\hat{w}_{jk}\hat{w}_{ki})^{1/3}, \tag{28}$$

where $\hat{w}_{ij} = \hat{w}_{ij}/\max_{ij} w_{ij}$.

Given the clustering coefficients of the vertices, the clustering coefficient can be expressed as a function of the degree of the vertices:

$$C(k) = \frac{\sum_i C_i\delta_{k_ik}}{\sum_i \delta_{k_ik}}, \tag{29}$$

where δ_{ij} is the Kronecker delta. For some networks, this function has the form $C(k) \sim k^{-\alpha}$. This behavior has been associated with a hierarchical structure of the

network, with the exponent α being called its *hierarchical exponent* [100]. Soffer and Vázquez [101] found that this dependence of the clustering coefficient with k is to some extent due to the degree correlations (section 6) of the networks, with vertices of high degree connecting with vertices of low degree. They suggested a new definition of clustering coefficient without degree correlation bias:

$$\tilde{C}_i = \frac{l_i}{\omega_i}, \tag{30}$$

where l_i is the number of edges between neighbors of i and ω_i is the maximum number of edges possible between the neighbors of vertex i, considering their vertex degrees and the fact that they are necessarily connected with vertex i.

5.2. *Cyclic coefficient*

Kim and Kim [102] defined a coefficient to measure how cyclic a network is. The *local cyclic coefficient* of a vertex i is defined as the average of the inverse of the sizes of the smallest cycles formed by vertex i and its neighbors,

$$\theta_i = \frac{2}{k_i(k_i-1)} \sum_{k>j} \frac{1}{S_{ijk}} a_{ij}a_{ik}, \tag{31}$$

where S_{ijk} is the size of the smallest cycle which passes through vertices i, j and k. Note that if vertices j and k are connected, the smallest cycle is a triangle and $S_{ijk}=3$. If there is no loop passing through i, j and k, then these vertices are treelike connected and $S_{ijk}=\infty$. The cyclic coefficient of a network is the average of the cyclic coefficient of all its vertices:

$$\theta = \frac{1}{N}\sum_i \theta_i. \tag{32}$$

5.3. *Structure of loops*

High clustering coefficient and power-law degree distribution are ubiquitous in most real networks [24]. However the clustering coefficient is insufficient for comprehensive analysis of complex networks [103]. Depending on the network topology, cycles composed by more than three vertices can be fundamental for transport and network maintenance [104]. An interesting issue regards the investigation of cycles with length larger than three, such as by using the *grid coefficient*, $c_{4,i}$, which is defined as the fraction of all quadrilaterals (cycles of length four) passing by the vertex i divided by the maximum possible number of quadrilaterals sharing the vertex i. Such a coefficient can be generalized for cycles of length n as the fraction of all cycles of length n that pass through the vertex i divided by the maximum number of those type of cycles that could pass by i. Also, Lind *et al.* [105] defined a clustering coefficient for squares, $C_4(i)$, i.e. the probability that two neighbors of node i share a common

neighbor (different from i),

$$C_4(i) = \frac{\sum_{j=1}^{k_i} \sum_{l=j+1}^{k_i} q_i(j,l)}{\sum_{j=1}^{k_i} \sum_{l=j+1}^{k_i} [a_i(j,l) + q_i(j,l)]}, \tag{33}$$

where j and l label neighbors of node i, $q_i(j, l)$ are the number of common neighbors between j and l. Also, $a_i(j, l) = (k_j - \eta_i(j, l))(k_l - \eta_i(j, l))$ with $\eta_i(j, l) = 1 + q_i(j, l) + \theta_{jl}$ and $\theta_{jl} = 1$ if neighbors j and l are connected with each other and 0 otherwise.

The estimated distribution of cycles can be used for network characterization [106]. Rozenfeld *et al*. [104] proposed a *loopiness exponent* determined in terms of the characteristic cycle length. Moreover, a way to compute cycles in networks is by using the adjacency matrix A [107]. The number of loops of order 3 is given as,

$$N_3 = \frac{1}{6} \sum_i (A^3)_{ii}, \tag{34}$$

and for orders 4 and 5,

$$N_4 = \frac{1}{8}\left[\sum_i (A^4)_{ii} - 2\sum_i (A^2)_{ii}(A^2)_{ii} + \sum_i (A^2)_{ii}\right], \tag{35}$$

$$N_5 = \frac{1}{10}\left[\sum_i (A^5)_{ii} - 5\sum_i (A^2)_{ii}(A^3)_{ii} + 5\sum_i (A^3)_{ii}\right]. \tag{36}$$

Such loops have been studied in the context of autonomous systems by Bianconi *et al*. [108], who showed that the statistical distribution of loops of order 3, 4 and 5 remains stable during the network evolution. Other works have addressed the statistical estimation of loops in complex networks, including the work by Bianconi and Marsili, who studied the size of loops and Hamiltonian cycles in random scale-free networks [109], and the impact of degree correlations in loops present in scale-free networks [110].

5.4. *Rich-Club coefficient*

In science, influential researchers of some areas tend to form collaborative groups and publish papers together [111]. This tendency is observed in other real networks and reflect the tendency of hubs to be well connected with each other. This phenomenon, known as *rich-club*, can be measured by the *rich-club coefficient*, introduced by Zhou and Mondragon [112]. The rich-club of degree k of a network G is the set of vertices with degree greater than k, $\mathcal{R}(k) = \{v \in \mathcal{N}(G) | k_v > k\}$. The rich-club coefficient of degree k is given by

$$\phi(k) = \frac{1}{|\mathcal{R}(k)|(|\mathcal{R}(k)| - 1)} \sum_{i,j \in \mathcal{R}(k)} a_{ij} \tag{37}$$

(the sum corresponds to two times the number of edges between vertices in the club). This measurement is similar to that defined before for the clustering coefficient (see equation (24)), giving the fraction of existing connections among vertices with degree higher than k.

Colizza *et al.* [111] derived an analytical expression of the rich-club coefficient, valid for uncorrelated networks,

$$\phi_{\text{unc}}(k) \sim \frac{k^2}{\langle k \rangle N}. \tag{38}$$

The definition of the *weighted rich-club coefficient* for weighted networks is straightforward. If $\mathcal{R}^w(s)$ is the set of vertices with strength greater than s, $\mathcal{R}^w(s) = \{v \in \mathcal{N}(G) | s_v > s\}$,

$$\phi^w(s) = \frac{\sum_{i,j \in \mathcal{R}^w(s)} w_{ij}}{\sum_{i \in \mathcal{R}^w(s)} s_i} \tag{39}$$

(the sum in the numerator gives two times the weight of the edges between elements of the rich-club, the sum in the denominator gives the total strength of the vertices in the club).

6. Degree distribution and correlations

The degree is an important characteristic of a vertex [113]. Based on the degree of the vertices, it is possible to derive many measurements for the network. One of the simplest is the *maximum degree*:

$$k_{\max} = \max_i k_i. \tag{40}$$

Additional information is provided by the *degree distribution*, $P(k)$, which expresses the fraction of vertices in a network with degree k. An important property of many real world networks is their power law degree distribution [9]. For directed networks there are an out-degree distribution $P^{\text{out}}(k^{\text{out}})$, an in-degree distribution $P^{\text{in}}(k^{\text{in}})$, and the joint in-degree and out-degree distribution $P^{\text{io}}(k^{\text{in}}, k^{\text{out}})$. The latter distribution gives the probability of finding a vertex with in-degree k^{in} and out-degree k^{out}. Similar definitions considering the strength of the vertices can be used for weighted networks. An objective quantification of the level to which a log-log distribution of points approach a power law can be provided by the respective Pearson coefficient, which is henceforth called *straightness* and abbreviated as *st*.

It is often interesting to check for correlations between the degrees of different vertices, which have been found to play an important role in many structural and dynamical network properties [114]. The most natural approach is to consider the correlations between two vertices connected by an edge. This correlation can be expressed by the joint degree distribution $P(k, k')$, i.e. as the probability that an arbitrary edge connects a vertex of degree k to a vertex of degree k'. Another way to express the dependence between vertex degrees is in terms of the *conditional probability* that an arbitrary neighbor of a vertex of degree k has degree k' [31, 115],

$$P(k'|k) = \frac{\langle k \rangle P(k, k')}{kP(k)}. \tag{41}$$

Notice that $\sum_{k'} P(k'|k) = 1$. For undirected networks, $P(k, k') = P(k', k)$ and $k'P(k|k')\ P(k') = kP(k'|k)\ P(k)$. For directed networks, k is the degree at the tail of

the edge, k' is the degree at the head, both k and k' may be in-, out-, or total degrees, and in general $P(k, k') \neq P(k', k)$. For weighted networks the strength s can be used instead of k.

$P(k, k')$ and $P(k|k')$ characterize formally the vertex degree correlations, but they are difficult to evaluate experimentally, especially for fat-tailed distributions, as a consequence of the finite network size and the resulting small sample of vertices with high degree. This problem can be addressed by computing the *average degree of the nearest neighbors* of vertices with a given degree k [116], which is given by

$$k_{\mathrm{nn}}(k) = \sum_{k'} k' P(k'|k). \tag{42}$$

If there are no correlations, $k_{\mathrm{nn}}(k)$ is independent of k, $k_{\mathrm{nn}}(k) = k^2/k$. When $k_{\mathrm{nn}}(k)$ is an increasing function of k, vertices of high degree tend to connect with vertices of high degree, and the network is classified as *assortative*, whereas whenever $k_{\mathrm{nn}}(k)$ is a decreasing function of k, vertices of high degree tend to connect with vertices of low degree, and the network is called *disassortative* [117].

Another way to determine the degree correlation is by considering the Pearson correlation coefficient of the degrees at both ends of the edges [117]:

$$r = \frac{(1/M)\sum_{j>i} k_i k_j a_{ij} - \left[(1/M)\sum_{j>i}(1/2)(k_i + k_j)a_{ij}\right]^2}{(1/M)\sum_{j>i}(1/2)\left(k_i^2 + k_j^2\right)a_{ij} - \left[(1/M)\sum_{j>i}(1/2)(k_i + k_j)a_{ij}\right]^2}, \tag{43}$$

where M is the total number of edges. If $r > 0$ the network is assortative; if $r < 0$, the network is disassortative; for $r = 0$ there is no correlation between vertex degrees.

Degree correlations can be used to characterize networks and to validate the ability of network models to represent real network topologies. Newman [117] computed the Pearson correlation coefficient for some real and model networks and discovered that, although the models reproduce specific topological features such as the power law degree distribution or the small-world property, most of them (e.g., the Erdő-Rényi and Barabási-Albert models) fail to reproduce the assortative mixing ($r = 0$ for the Erdő-Rényi and Barabási-Albert models). Further, it was found that the assortativity depends on the type of network. While social networks tend to be assortative, biological and technological networks are often disassortative [24]. The latter property is undesirable for practical purposes, because disassortative networks are known to be resilient to simple target attack, at the least. So, for instance, in disease propagation, social networks would ideally be vulnerable (i.e. the network is dismantled into connected components, isolating the focus of disease) and technological and biological networks should be resilient against attacks. The degree correlations are related to the network evolution process and, therefore, should be taken into account in the development of new models as done, for instance, in the papers by Catanzaro *et al.* [118] on social networks, Park and Newman [119] on the Internet, and Berg *et al.* [120] on protein interaction networks. Degree correlations also have strong influence on dynamic processes like instability [121], synchronization [122, 123] and spreading [115, 124, 125]. For additional discussions about dynamic process as in networks see Ref. [24].

7. Networks with different vertex types

Some networks include vertices of different types. For example, in a sociological network where the vertices are people and the edges are a social relation between two persons (e.g., friendship), one may be interested in answering questions like: how probable is a friendship relationship between two persons of different economic classes? In this case, it is interesting to consider that the vertices are not homogeneous, having different types. In the following, measurements associated with such kind of networks are discussed.

7.1. *Assortativity*

For networks with different types of vertices, a type mixing matrix E can be defined, with elements e_{st} such that e_{st} is the number of edges connecting vertices of type s to vertices of type t (or the total strength of the edges connecting the two vertices of the given types, for weighted networks). It can be normalized as

$$\hat{E} = \frac{E}{\|E\|}, \tag{44}$$

where $\|X\|$ (cardinality) represents the sum of all elements of matrix X.

The probability of a vertex of type s having a neighbor of type t therefore is

$$P^{(\text{type})}(t|s) = \frac{\hat{e}_{st}}{\sum_u \hat{e}_{su}}. \tag{45}$$

Note that $\sum_t P^{(\text{type})}(t|s) = 1$.

$P^{(\text{type})}(t|s)$ and $\hat{E}$ can be used to quantify the tendency in the network of vertices of some type to connect to vertices of the same type, called *assortativity*. We can define an assortativity coefficient [23, 126] as:

$$\tilde{\mathbb{Q}} = \frac{\sum_s P^{(\text{type})}(s|s) - 1}{N_T - 1}, \tag{46}$$

where N_T is the number of different vertex types in the network. It can be seen that $0 \leq \tilde{\mathbb{Q}} \leq 1$, where $\tilde{\mathbb{Q}} = 1$ for a perfectly assortative network (only edges between vertices of the same type) and $\tilde{\mathbb{Q}} = 0$ for random mixing. But there is a problem with this definition because each vertex type has the same weight in $\tilde{\mathbb{Q}}$, regardless of the number of vertices of that type. An alternative definition that avoids this problem [127] is:

$$\mathbb{Q} = \frac{\operatorname{Tr} \hat{E} - \|\hat{E}^2\|}{1 - \|\hat{E}^2\|}. \tag{47}$$

It is interesting to associate the vertex type to its degree. The Pearson correlation coefficient of vertex degrees, equation (43), can be considered as an assortativity coefficient for this case.

7.2. *Bipartivity degree*

A special case of disassortativity is that of bipartite networks. A network is called *bipartite* if its vertices can be separated into two sets such that edges exist only

between vertices of different sets. It is a known fact that a network is bipartite if and only if it has no loops of odd length (e.g. [128]). Although some networks are bipartite by construction, others, like a network of sexual contacts, are only approximately bipartite. A way to quantify how much a network is bipartite is therefore needed. A possible measurement is based on the number of edges between vertices of the same subset in the best possible division [128],

$$b = 1 - \frac{\sum_{ij} a_{ij}\, \delta_{\vartheta(i),\, \vartheta(j)}}{\sum_{ij} a_{ij}}, \tag{48}$$

where $\vartheta(i)$ maps a vertex i to its type and δ is the Kronecker delta. The smallest value of b for all possible divisions is the bipartivity of the network. The problem with this measurement is that its computation is NP-complete, due to the necessity of evaluating b for the best possible division. A measurement that approximates b but is computationally easier was proposed in [128], based on a process of marking the minimum possible number of edges as responsible for the creation of loops of odd length.

Another approach is based on the subgraph centrality [129] (section 12.3). The subgraph centrality of the network, equation (91), is divided in part due to even closed walks and part due to odd closed walks (a closed walk is a walk, possibly with repetition of vertices, ending on the starting vertex). As odd closed walks are not possible in bipartite networks, the fraction of the subgraph centrality of the network due to even closed walks can be used as the bipartivity degree [129]:

$$\beta = \frac{SC_{\text{even}}}{SC} = \frac{\sum_{j=1}^{N} \cosh \lambda_j}{\sum_{j=1}^{N} e^{\lambda_j}}, \tag{49}$$

where SC is the subgraph centrality of the network (section 12.3), SC_{even} is the subgraph centrality due to the even closed walks and the λ_j are the eigenvalues of the adjacency matrix of the network.

8. Entropy and energy

Entropy and energy are key concepts in thermodynamics, statistical mechanics [130] and information theory [131]. Entropy has important physical implications related to the amount of "disorder" and information in a system [132]. In information theory, entropy describes how much randomness is present in a signal or random event [133]. These concepts can be applied to complex networks.

8.1. *Entropy of the degree distribution*

The *entropy of the degree distribution* provides an average measurement of the heterogeneity of the network, which can be defined as

$$H = -\sum_{k} P(k) \log P(k). \tag{50}$$

The maximum value of entropy is obtained for a uniform degree distribution and the minimum value $H_{\min} = 0$ is achieved whenever all vertices have the same degree [134].

Network entropy has been related to the robustness of networks, i.e. their resilience to attacks [134], and the contribution of vertices to the network entropy is correlated with lethality in protein interactions networks [135].

Solé and Valverde [136] suggested the use of the remaining degree distribution to compute the entropy. The *remaining degree* of a vertex at one end of an edge is the number of edges connected to that vertex not counting the original edge. The remaining degree distribution can be computed as

$$q(k) = \frac{(k+1)P(k+1)}{\langle k \rangle}. \tag{51}$$

The entropy of the remaining degree is given by

$$H^* = -\sum_k q(k) \log q(k). \tag{52}$$

8.2. *Search information, target entropy and road entropy*

The structure of a complex network is related to its reliability and information propagation speed. The difficulty while searching information in the network can be quantified through the information entropy of the network [137, 138]. Rosvall *et al*. [139] introduced measurements to quantify the information associated to locating a specific target in a network. Let $p(i, b)$ be a shortest path starting at vertex i and ending at vertex b. The probability of following this path in a random walk is

$$\mathcal{P}[p(i,b)] = \frac{1}{k_i} \prod_{j \in p(i,b)} \frac{1}{k_j - 1}, \tag{53}$$

where k_j is the degree of vertex j and the product includes all vertices j in the path $p(i, b)$ with the exclusion of i and b. The *search information*, corresponding to the total information needed to identify one of the shortest paths between i and b, is given by

$$\mathcal{S}(i,b) = -\log_2 \sum_{\{p(i,b)\}} \mathcal{P}[p(i,b)], \tag{54}$$

where the sum is taken over all shortest paths $p(i, b)$ from i to b.

The average search information characterizes the ease or difficulty of navigation in a network and is given by [139]

$$\mathcal{S} = \frac{1}{N^2} \sum_{ib} \mathcal{S}(i,b). \tag{55}$$

This value depends on the structure of the network. As discussed by Rosvall *et al*. [139], city networks are more difficult to navigate than their random counterparts.

In order to measure how difficult it is to locate vertices in the network starting from a given vertex i, the *access information* is used,

$$\mathcal{A}_i = \frac{1}{N} \sum_b \mathcal{S}(i,b), \tag{56}$$

which measures the average number of "questions" needed to locate another vertex starting from i. To quantify how difficult is to find the vertex b starting from the other vertices in the network, the *hide information* is used,

$$\mathcal{H}_b = \frac{1}{N}\sum_i \mathcal{S}(i,b). \tag{57}$$

Note that the average value of $\mathcal{A}_i$ and $\mathcal{H}_b$ for a network is $\mathcal{S} : \sum_i \mathcal{A}_i = \sum_b \mathcal{H}_b = \mathcal{S}N$.

Considering the exchange of messages in the network, it is possible to define entropies in order to quantify the predictability of the message flow. Assuming that messages always flow through shortest paths and all pairs of vertices exchange the same number of messages at the same rate, the following entropies can be defined [137]:

$$\mathcal{T}_i = -\sum_{ij} a_{ji}c_{ij}\log_2 c_{ij}, \tag{58}$$

$$\mathcal{R}_i = -\sum_{ij} a_{ji}b_{ij}\log_2 b_{ij}, \tag{59}$$

where a_{ji} is an element of the adjacency matrix, c_{ij} is the fraction of messages targeted at vertex i that comes through vertex j, and b_{ij} is the fraction of messages that goes through vertex i coming from vertex j. In addition, $\mathcal{T}_i$ is the *target entropy* of vertex i and $\mathcal{R}_i$ is the *road entropy* of vertex i. Low values of these entropies mean that the vertex from where the next message originates (*to* vertex i or passing *through* vertex i) can be easily predicted.

As a general measurement of the flows of messages, we can define target and road entropies for the network as averages among all vertices

$$\mathcal{T} = \frac{1}{N}\sum_i \mathcal{T}_i, \tag{60}$$

$$\mathcal{R} = \frac{1}{N}\sum_i \mathcal{R}_i. \tag{61}$$

As shown in [137], these quantities are related to the organization of the network: a network with a low value of $\mathcal{T}$ has a star structure and a low value of $\mathcal{R}$ means that the network is composed by hubs connected in a string.

Further works related to searchability in networks have been reported by Trusina *et al*. [140], who defined search information weighted by the traffic on the network, and Rosvall *et al*. [141], who studied networks with higher order organization like modular or hierarchical structure.

8.3. *Energy of complex networks*

By using concepts of statistical mechanics [130], it is possible to define the energy associated with networks and the respective partition function. Based on such concepts, Bianconi [142] proposed a theoretical approach to describe the emergence of scale-free degree distribution or finite-scale degree distribution in complex networks. In particular, the energy associated to a degree distribution N_k is given as

$$E(\{N_k\}) = \log(\mathcal{N}_G), \tag{62}$$

where $\mathcal{N}_G$ is the number of indistinguishable networks that can be constructed from the distribution N_k, and

$$\mathcal{N}_G = e^{E(\{N_k\})} = \prod_k k!^{N_k}. \tag{63}$$

The entropy of each distribution, $S(\{N_k\})$, is defined as

$$e^{\mathcal{S}}(\{N_k\}) = \mathcal{N}_{N_k} = \frac{(2L)!}{\prod_k (kN_k)!}, \tag{64}$$

where $\{N_k\}$ is the number of ways in which it is possible to distribute $2L$ edges into a degree sequence $\{k_1, \ldots, k_N\}$. By using such concepts, Bianconi showed that the optimal degree distribution with respect to the free energy minimization is obtained for scale-free degree distribution [142].

9. Centrality measurements

In networks, the greater the number of paths in which a vertex or edge participates, the higher the importance of this vertex or edge for the network. Thus, assuming that the interactions follow the shortest paths between two vertices, it is possible to quantify the importance of a vertex or a edge in terms of its *betweenness centrality* [143] defined as:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)}, \tag{65}$$

where $\sigma(i, u, j)$ is the number of shortest paths between vertices i and j that pass through *vertex* or *edge* u, $\sigma(i, j)$ is the total number of shortest paths between i and j, and the sum is over all pairs i, j of distinct vertices.

When one takes into account the fact that the shortest paths might not be known and instead a search algorithm is used for navigation (see section 4.1), the betweenness of a vertex or edge must be defined in terms of the probability of it being visited by the search algorithm. This generalization, which was introduced by Arenas *et al.* [144], subsumes the betweenness centrality based on random walks as proposed by Newman [145].

The *central point dominance* is defined as [143]

$$CPD = \frac{1}{N-1} \sum_i (B_{\max} - B_i), \tag{66}$$

where $B_{\max}$ is the largest value of betweenness centrality in the network. The central point dominance will be 0 for a complete graph and 1 for a star graph in which there is a central vertex included in all paths. Other centrality measurements can be found in the interesting survey by Koschützki *et al.* [146].

10. Spectral measurements

The spectrum of a network corresponds to the set of eigenvalues λ_i ($i = 1, 2, \ldots, N$) of its adjacency matrix A. The spectral density of the network is defined as [147, 148]:

$$\rho(\lambda) = \frac{1}{N}\sum_i \delta(\lambda - \lambda_i), \tag{67}$$

where $\delta(x)$ is the Dirac delta function and ρ approaches a continuous function as $N \to \infty$; e.g., for Erdős-Rényi networks, if p is constant as $N \to \infty$, $\rho(\lambda)$ converges to a semicircle [147]. Also, the eigenvalues can be used to compute the lth-moments,

$$M_l = \frac{1}{N}\sum_{i_1, i_2, \ldots, i_l} a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_l i_1} = \frac{1}{N}\sum_i (\lambda_i)^l. \tag{68}$$

The eigenvalues and associated eigenvectors of a network are related to the diameter, the number of cycles and connectivity properties of the network [147, 148]. The quantity $D_l = NM_l$ is the number of paths returning to the same vertex in the graph passing through l edges. Note that these paths can contain already visited vertices. In a tree-like graph, a return walk is only possible going back through the already visited edges, the presence of odd moments is a sure sign of cycles in the graph; in particular, as a walk can go through three edges and return to its starting vertex only by following three different edges (if self-connections are not allowed), D_3 is related with the number of triangles in the network [148].

In addition, spectral analysis allows the determination that a network is bipartite (if it does not contain any odd cycle [129], see section 7.2), characterizing models of real networks [149, 150], and visualizing networks [151]. In addition, spectral analysis of networks is important to determine communities and subgraphs, as discussed in the next section.

11. Community identification and measurements

Many real networks present an inhomogeneous connecting structure characterized by the presence of groups whose vertices are more densely interconnected to one another than with the rest of the network. This modular structure has been found in many kinds of networks such as social networks [152, 153], metabolic networks [154] and in the worldwide flight transportation network [89]. Figure 10 presents a network with a well-defined community structure.

Community identification in large networks is particularly useful because vertices belonging to the same community are more likely to share properties and dynamics. In addition, the number and characteristics of the existing communities provide subsidies for identifying the category of a network as well as understanding its dynamic evolution and organization. In the case of the World Wide Web, for instance, pages related to the same subject are typically organized into communities, so that the identification of these communities can help the task of seeking for information. Similarly, in the case of the Internet, information about communities formed by routers geographically close one another can be considered in order to improve the flow of data.

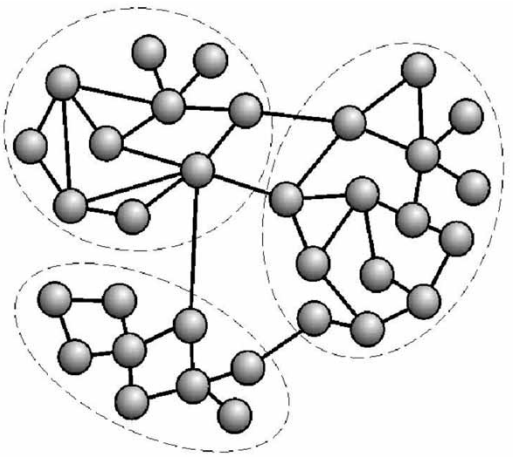


Figure 10. A network with community structure represented by the dashed lines. The communities are the groups of more intensely interconnected vertices.

Despite the importance of the concept of community, there is no consensus about its definition. An intuitive definition was proposed by Radichi *et al.* [155] based on the comparison of the edge density among vertices. Communities are defined in a strong and a weak sense. In a *strong sense*, a subgraph is a community if all of its vertices have more connections between them than with the rest of the network. In a *weak sense*, on the other hand, a subgraph is a community if the sum of all vertex degrees inside the subgraph is larger than outside it. Though these definitions are intuitive, one of their consequences is that every union of communities is also a community. To overcome this limitation a hierarchy among the communities can be assumed *a priori*, as discussed by Reichardt and Bornholdt [156], who defined community in networks as the spin configuration that minimizes the energy of the spin glass by mapping the community identification problem onto finding the ground state of a infinite range Potts spin glass [157, 158].

Another fundamentally related problem concerns how to best divide a network into its constituent communities. In real networks, no information is generally available about the number of existing communities. In order to address this problem, a measurement of the quality of a particular division of networks was proposed by Newman and Girvan [159], called *modularity* and typically represented by Q. If a particular network is split into c communities, Q can be calculated from the symmetric $c \times c$ *mixing matrix* E whose elements along the main diagonal, e_{ii}, give the fraction of connections between vertices in the same community i while the other elements, e_{ij} $(i \neq j)$ identify the fraction of connections between vertices in the different communities i and j. This is similar to the definition used to compute assortativity, section 7. The calculation of Q can then be performed as follows:

$$Q = \sum_i \left[e_{ii} - \left(\sum_j e_{ij} \right)^2 \right] = \mathrm{Tr}E - \| E^2 \|. \tag{69}$$

The situation $Q = 1$ identifies networks formed by disconnected modules. This quantity has been used in many community-finding algorithms, as briefly reviewed in the following.

Though there are many ways to define modularity, a generally accepted definition of a module does not exist [160]. The definitions described above estimate the modularity in terms of a given partitioning. Ziv *et al*. [161] proposed the modularity to be defined in terms of information entropy (see section 8). This algorithm, which has been called the *Network Information Bottleneck*, tends to allow performance better than the algorithm based on betweenness centrality of Girvan and Newman [159].

It should be noted that this review of community finding methods focused on the subject of how specific network measurements have been adopted to identify the communities. Since we do not attempt to provide a comprehensive study of this important subject, the interested reader should refer to recent papers by Newman [162] and Danon *et al*. [163] for further information and a more complete review on community finding methods. The following discussion has been organized into subsections according to the nature of the adopted methodology.

11.1. *Spectral methods*

Spectral methods are based on the analysis of the eigenvectors of matrices derived from the networks [164]. These methods have been discussed in a recent survey by Newman [165]. The quantity measured corresponds to the eigenvalues of matrices associated with the adjacency matrix. These matrices can be the *Laplacian matrix* (also known as Kirchhoff matrix),

$$L = D - A, \tag{70}$$

or the *Normal matrix*,

$$\tilde{A} = D^{-1}A, \tag{71}$$

where D is the diagonal matrix of vertex degrees with elements $d_{ii} = \sum_j a_{ij}, d_{ij} = 0$ for $i \neq j$.

A particular method, called *spectral bisection* [165–167], is based on the diagonalization of the Laplacian matrix. If the network is separated into c disconnected components, L will be block diagonal and have c degenerated eigenvectors, all corresponding to eigenvalue 0. However, if the separation is not clear, the diagonalization of L will produce one eigenvector with eigenvalues 0 and $c-1$ eigenvalues slightly different from 0. The spectral bisection considers the case when $c = 2$ and the division of the network is obtained assigning positive components of the eigenvector associated with the second eigenvalue (the positive eigenvalue most close to 0) to one community and the negative ones to another community. Particularly, the second eigenvalue, called *algebraic connectivity*, is a measurement of how good the division is, with small values corresponding to better divisions. Although spectral bisection is easy to implement, it tends to be a poor approach for detecting communities in real networks [165]. There are many alternative methods based on spectral analysis [168], to be found in refs [162, 163].

Recently, Newman [169] proposed a method which reformulates the modularity concept in terms of the eigenvectors of a new characteristic matrix for the network,

called *modularity matrix*. For each subgraph g, its modularity matrix $B^{(g)}$ has elements

$$b_{ij}^{(g)} = a_{ij} - \frac{k_i k_j}{2M} - \delta_{ij} \sum_{u \in \mathcal{N}(g)} \left[a_{iu} - \frac{k_i k_u}{2M} \right], \tag{72}$$

for vertices i and j in g. Thus, in order to split the network in communities, first the modularity matrix is constructed and its most positive eigenvalue and corresponding eigenvector are determined. According to the signs of the elements of this vector, the network is divided into two parts (vertices with positive elements are assigned to a community and vertices with negative elements to another). Next, the process is repeated recursively to each community until a split which makes zero or a negative contribution to the total modularity is reached. Following this idea, Newman proposed a new definition of communities as indivisible subgraphs, i.e. subgraphs whose division would not increase the modularity. Currently, this method is believed to be the most precise, as it is able to find a division with the highest value of modularity for many networks [169].

11.2. *Divisive methods*

In a divisive method, the underlying idea is to find the edges which connect different communities and remove them in a iterative form, breaking the network into disconnected groups of vertices. The computation of modularity can be used afterwards to determine the best division of the network. Next we give a brief description of the most known divisive methods according to the adopted measurement used to choose the vertex to remove.

11.2.1. Betweenness centrality. The most popular divisive method is the Girvan-Newman algorithm [10]. Because different communities are connected by a small number of edges, this method considers that bottlenecks are formed at the edges which connect communities, through which all shortest paths should pass. In order to measure this traffic-related property in networks, the algorithm uses the concept of *edge betweenness* [10], see section 9. Edges with high betweenness are progressively removed. After removing each edge, the betweenness of each remaining edge must be calculated again.

Although this algorithm represents a powerful alternative to determine communities (as shown in figure 11), it has some disadvantages. The main one is its high computational cost. As discussed by Girvan and Newman [10], the entire algorithm runs in worst-case time $O(M^2N)$ on networks with M edges and N vertices. In order to overcome this limitation, some improvements in the algorithm were proposed including the Tyler's algorithm [170], which introduced a stochastic element to the method, restricting the calculation of the betweenness only to a partial set of edges and using statistics to estimate the real betweenness.

11.2.2. Edge clustering coefficient. A different approach was proposed by Radicchi *et al*. [155] (see also [171]), which is based on counting short loops of order l (triangles for $l = 3$) in networks. The algorithm is similar to Girvan and Newman's method, but instead of the betweenness centrality, it computes the *edge clustering coefficient*.

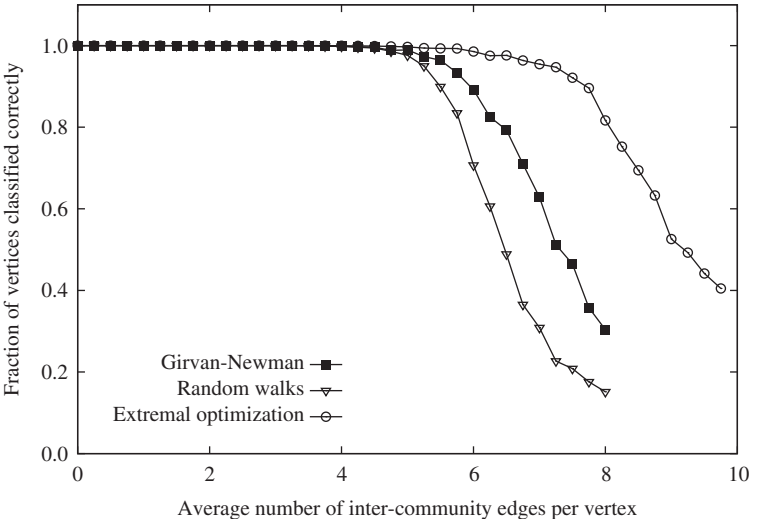


Figure 11. Comparison of precision between the methods developed by Girvan and Newman (see section 11.2.1), the same method based on random walks [179], and the method developed by Duch and Arenas, based on extremal optimization (see section 11.4.1). Each point in this graph is an average of 100 realizations of networks created by using the model described in section 3.5, with 128 vertices organized into 4 communities and varying the density of connections inside and outside communities.

This measurement is based on the fact that edges which connect communities tend to exhibit a small value for this coefficient. The clustering coefficient of edge (i, j) is calculated as

$$C_{ij} = \frac{Z_{ij} + 1}{\min(k_i - 1, k_j - 1)}, \tag{73}$$

where Z_{ij} is the number of triangles to which (i, j) belongs. This method can be generalized to more complex loops, e.g., squares. Though this method is simple and fast ($O(M^4/N^2)$), it fails whenever the network has a small average clustering coefficient, because the value of C_{ij} will be small for all edges. This suggests that the method will work well only when applied to networks with a high average clustering coefficient, such as social networks [163].

11.3. *Agglomerative methods*

Some networks are characterized by the fact that the vertices belonging to each community present similar features. So, it is in principle, possible to obtain the communities by considering such similarities between vertices. In contrast to divisive methods, agglomerative approaches start with all vertices disconnected and then apply some similarity criterion to progressively join them and obtain the communities. It is interesting to note that this type of method presents a direct relationship with pattern recognition and clustering theory and algorithms (e.g. [68, 172–174]),

which have been traditionally used in order to group individuals represented by a vector of features into meaningful clusters.

11.3.1. Similarity measurements. One important family of agglomerative methods is known as *hierarchical clustering* [13, 68, 172, 173], which starts with N vertices and no edges. Edges are added progressively to the network in decreasing order of similarity, starting with the pair with strongest similarity [162, 175]. To evaluate the similarity associated with edge (i,j), a possibility is to use the so called *Euclidian distance*, given by

$$\sqrt{\sum_{k\neq i,j}(a_{ik}-a_{jk})^2}, \tag{74}$$

or the *Pearson correlation* between vertices as represented in the adjacency matrix, defined as

$$\frac{(1/N)\sum_k (a_{ik}-\mu_i)(a_{jk}-\mu_j)}{\sigma_i\sigma_j}, \tag{75}$$

where $\mu_i = (1/N)\sum_j a_{ij}$ and $\sigma_i^2 = (1/(N-1))\sum_j (a_{ij}-\mu_i)^2$.

Although this method is fast, the obtained division of the network is not generally satisfactory for real networks, as discussed in [162].

11.4. *Maximization of the modularity*

Newman [159, 176] proposed a method based on joining communities in such a way as to maximize the modularity. In this method, two communities i and j are joined according to a measurement of affinity, given by the change of the modularity Q of the network (equation (69)) when the communities are joined

$$\Delta Q_{ij} = 2\left(e_{ij} - \frac{\sum_j e_{ij}\sum_i e_{ij}}{2M}\right). \tag{76}$$

Thus, starting with each vertex disconnected and considering each of them as a community, we repeatedly join communities together in pairs, choosing at each step the joining that results in the greatest increase (or smallest decrease) in the modularity Q. This process can be repeated until the whole network is contained in one community. Currently, as discussed by Danon *et al.* [163], the Newman's method is believed to be the fastest one, running in $O(N\log^2 N)$. Also, this method is more precise than the traditional method based on betweenness centrality [159]. However, as discussed by Danon *et al.* [177], the fast Newman's method has a limitation when the size of the communities is not homogeneous, as a newly joined community i has the new values of e_{ij} in equation (76) increased, and tends to be chosen for new joining. In real networks, the distribution of sizes of communities tends to follow a power law. So, this approach fails in many real networks. In order to overcome this limitation, it was proposed [177] to normalize the value of ΔQ by the number of edges in community i,

$$\Delta\hat{Q}_{ij} = \frac{\Delta Q_{ij}}{\sum_j e_{ij}} = \frac{2}{\sum_j e_{ij}}\left(e_{ij} - \frac{\sum_j e_{ij}\sum_i e_{ij}}{2M}\right). \tag{77}$$

This alteration on the local modularity makes the method more precise while not affecting its execution time.

11.4.1. Extremal optimization. The extremal optimization method proposed by Duch and Arenas [178] is a heuristic search for optimizing the value of the modularity Q. The *local modularity* represents the contribution of individual vertex i to the modularity Q. If c_i is the community of vertex i, the local modularity is given by

$$q_i = \sum_j a_{ij}\,\delta_{c_i,\,c_j} - k_i \sum_{c_k} e_{c_i c_k}, \tag{78}$$

where $e_{c_i c_j}$ are the elements of the community mixing matrix (page 196) and δ is the Kronecker delta. In order to keep the value of this contribution in the interval $[-1, 1]$ and independent of vertex degree, it should be normalized by the degree of the vertex, i.e. $\hat{q}_i = q_i/k_i$. The value of $\hat{q}_i$ is used as the *fitness* for the extremal optimization algorithm. A heuristic search is performed to obtain the maximum value of the modularity. Initially, the network is split into two random partitions with the same number of vertices. After each step, the system self-organizes by moving the vertex with lowest fitness from one partition to another. The process stops when the maximum value of Q is reached. After that, all links between both partitions are deleted and the optimization of Q proceeds recursively considering every resulting connected component. The process of community identification finishes when the value of Q cannot be improved further.

Although this method is not particularly fast, scaling as $O(N^2 \log N)$, it can achieve high modularity values [178]. By comparing the precision of some methods as presented in figure 11, we can see that the extremal optimization method is more precise than the methods based on removing edges with highest betweenness centrality value. Moreover, it is clear that the computation of betweenness centrality by counting the number of shortest paths passing through each edge is more precise than calculating this coefficient by random walks [159].

11.5. *Local methods*

More recently, some methods have been developed to detect the local community of a vertex based only on local information about the network topology. One such method was proposed by Bagrow and Bolt [180], which is based on the change of the hierarchical degree between two consecutive distances (see section 13). Starting from a vertex v_0, the vertices of successive hierarchical rings are added to the community, as long as the relation between the successive hierarchical degrees is greater than a specified threshold α

$$\frac{k_d(v_0)}{k_{d-1}(v_0)} > \alpha. \tag{79}$$

When the expansion reaches a distance d for which the above condition fails, the community stops growing.

Despite its favorable speed, this approach has an important limitation: the division is precise only when v_0 is equidistant from all parts of its enclosing community's boundary [181]. In order to overcome this drawback, it has been suggested [180] that the algorithm be executed N times starting from each vertex and then achieve a

consensus about the detected communities. However, this approach increases the execution time of the algorithm.

Another local method was proposed by Clauset [181] which is based on computing the *local modularity*. The idea is that of a step-by-step growth of the community together with the exploration of the network. The community $\mathcal{C}$ starts with only the original vertex v_0. When a vertex is explored, a list of its neighbors is known. The set $\mathcal{U}$ is a list of all vertices that are not in $\mathcal{C}$ but are adjacent to some of its vertices; the set $\mathcal{B}$ (the *boundary* of $\mathcal{C}$) is the subset of vertices in $\mathcal{C}$ that are adjacent of at least one vertex in $\mathcal{U}$. The local modularity is defined as the ratio of the number of edges with one end point in $\mathcal{B}$ and neither end point in $\mathcal{U}$ to the number of edges with end points in $\mathcal{B}$. Considering undirected networks, this can be written as

$$R = \frac{\sum_{i\in\mathcal{B}, j\in\mathcal{C}} a_{ij}}{\sum_{i\in\mathcal{B}, j} a_{ij}}. \tag{80}$$

The algorithm consists in choosing iteratively from the set $\mathcal{U}$ the vertex that would result in the largest increase (or smallest decrease) in the value of R when added to $\mathcal{C}$. The iteration stops when a pre-defined number of vertices was included in the community.

11.6. *Method selection*

Despite the many interesting alternative methods, including those briefly reviewed above, it should be noted that the problem of community finding remains a challenge because no single method is fast and sensitive enough to ensure ideal results for general, large networks, a problem which is compounded by the lack of a clear definition of communities. If communities are to be identified with high precision, the spectral method proposed by Newman [169] is a good choice. However, if priority is assigned to speed, methods such as those using greedy algorithms (runs in $O(N \log^2 N)$) should be considered [176]. In brief, the choice of the best method to be used depends on the configuration of the problem and the kind of desired results [163].

One fact that should have become clear from our brief review of community finding approaches is the essential importance of the choice of the *measurements* adopted to express the separation of the communities. As a matter of fact, such measurements ultimately represent an objective definition of communities. Therefore, an interesting perspective for further research would be to consider the possible adaptation and combination of some of the measurements reported in this survey with the specific objective of community characterization.

11.7. *Roles of vertices*

After community identification, it is possible to determine the role of vertices [154] by using the *z-score of the within-module degree*, z_i, and the *participation coefficient*, P_i. The z-score measures how "well-connected" vertex i is to the other vertices in the community, being defined by

$$z_i = \frac{q_i - \bar{q}_{s_i}}{\sigma_{q_{s_i}}}, \tag{81}$$

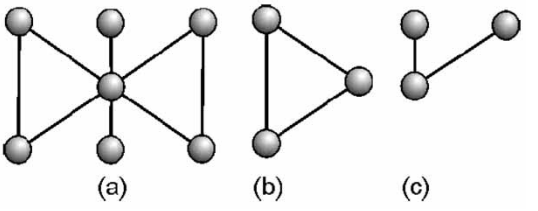


Figure 12. A network such as that in (a) includes several subgraphs, such as cycles (b) and trees (c).

where q_i is the number of connections i makes with other vertices in its own community s_i, $\bar{q}_{s_i}$ is the average of q over all vertex in s_i, and $\sigma_{q_{s_i}}$ is the standard deviation of q in s_i.

The participation coefficient measures how "well-distributed" the edges of vertex i are among different communities,

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{q_{is}}{k_i}\right)^2, \tag{82}$$

where q_{is} is the number of edges from vertex i to community s and k_i is the degree of vertex i. This value is zero if all edges are within its own community and it is close to one if its edges are uniformly distributed among all communities. Based on these two index, a zP parameter-space can be constructed, allowing the classification of vertices into different roles (see e.g. [154]).

12. Subgraphs

A graph g is a *subgraph* of the graph G if $\mathcal{N}(g) \subseteq \mathcal{N}(G)$ and $\mathcal{E}(g) \subseteq \mathcal{E}(G)$, with the edges in $\mathcal{E}(g)$ extending over vertices in $\mathcal{N}(g)$. If g contains all edges of G that connect vertices in $\mathcal{N}(g)$, the subgraph g is said to be *implied* by $\mathcal{N}(g)$. Important subgraphs include loops, trees (connected graphs without loops) and complete subnetworks (cliques). Figure 12 shows a network and some subnetworks. There are many ways to define subgraphs in networks. An interesting way to describe the topology of real networks in terms of subgraphs is by using the *k-core decomposition*. The k-core is obtained by removing from the network all vertices with degree smaller than k. After such a removal, some vertices in the resulting network may have degree less than k; such vertices are removed and the network is analyzed again. When no further removal is possible, the non-empty resulting subgraph is called k-core of the original network [182]. An important application of such a concept is in network visualization [183]. In this approach, the network is peeled layer by layer and the structure is displayed from the outmost shells. An algorithm for this type of visualization is publicly available, namely the Large Network Visualization tool [183].

Recent studies about the properties and applications of k-core decomposition in real networks have been performed [184, 185]. Important statistical properties of k-core are discussed by Dorogvtesev *et al.* [182] and investigations about topology of the Internet using k-core decomposition are presented by Carmi *et al.* [186].

Moreover, protein interaction networks are analyzed in terms of k-cores by Wuchty and Almaas [187]. The k-core approach has also been applied in order to predict the function of proteins [188].

There are other ways to define subgraphs than by analyzing the occurrence of subgraphs previously defined. An interesting kind of such subgraphs is called motifs, which are discussed below.

12.1. *Network motifs*

Network motifs are subgraphs that appear more frequently in a real network than could be statistically expected [189–191] (see figure 13). Figure 14 shows some

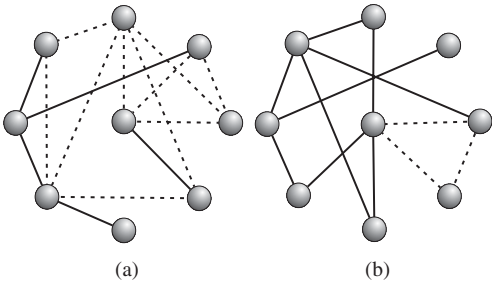


Figure 13. In a real network (a), the number of motifs (represented here by three vertices linked by dashed lines) is greater than in an equivalent random network (b).

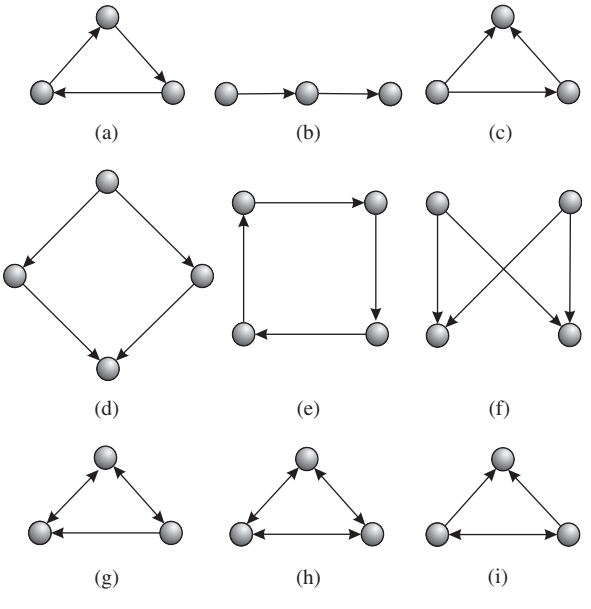


Figure 14. Some types of motifs: (a) three-vertex feedback loop, (b) three chain, (c) feed-forward loop, (d) bi-parallel, (e) four-vertex feedback loop, (f) bi-fan, (g) feedback with two mutual dyads, (h) fully connected triad and (i) uplinked mutual dyad.

possible motifs of directed networks and their conventional names. To find the motifs in a real network, the number of occurrences of subgraphs in the network is compared with the expected number in the ensemble of networks with the same degree for each vertex. A large number of randomized networks from this ensemble is generated in order to compute the statistics of occurrence of each subgraph of interest. If the probability of a given subgraph to appear at least the same number of times as in the real network is smaller than a given threshold (usually 0.01), the subgraph is considered a motif of the network.

In order to quantify the significance of a given motif, its *Z-score* can be computed. If $N_i^{(\text{real})}$ is the number of times that a motif i appears in the real network, $\langle N_i^{(\text{rand})} \rangle$ the ensemble average of its number of occurrences, and $\sigma_i^{(\text{rand})}$ the standard deviation of the number of occurrences, then:

$$Z_i = \frac{N_i^{(\text{real})} - \langle N_i^{(\text{rand})} \rangle}{\sigma_i^{(\text{rand})}}. \tag{83}$$

It is also possible to categorize different networks by the Z-scores of their motifs: networks that show emphasis on the same motifs can be considered as part of the same family [192]. For this purpose, the *significance profile* of the network can be computed. The significance profile is a vector that, for each motif of interest i, is used to compute the importance of this motif with respect to other motifs in the network:

$$SP_i = \frac{Z_i}{\sum_j Z_j^2}. \tag{84}$$

It is interesting to note that motifs are related to network evolution. As described by Milo *et al*. [190], different kinds of networks present different types of motifs e.g., for transcription networks of *Saccharomyces cerevisiae* and *Escherichia coli* two main motifs are identified: feed-forward loop and bi-fan; for neurons: feed-forward loop, bi-fan and bi-parallel; for food-webs: three chain and bi-parallel; for electronic circuits: feed-forward loop, bi-fan, bi-parallel, four-node feedback and three node feedback loop; and for the WWW: feedback with two mutual dyads, fully connected triad and uplinked mutual dyad. Thus, motifs can be considered as building blocks of complex networks and many papers have been published investigating the functions and evolution of motifs in networks [15].

12.2. *Subgraphs and motifs in weighted networks*

In weighted networks, a subgraph may be present with different values for the weights of the edges. Onnela *et al*. [99] suggested a definition for the *intensity* of a subgraph based on the geometric mean of its weights on the network. Given a subgraph g, its intensity is defined by

$$I(g) = \left(\prod_{(i,j) \in \mathcal{E}(g)} w_{ij} \right)^{1/n_g}, \tag{85}$$

where $n_g = |\mathcal{E}(g)|$ is the number of edges of subgraph g.

In order to verify whether the intensity of a subgraph is small because all its edges have small weight values or just one of the weights is too small, the *coherence* of the

subgraph Ψ, defined as the ratio between geometric and arithmetic mean of its weights, can be used:

$$\Psi(g) = \frac{I(g)n_g}{\sum_{(i,j)\in\mathcal{E}(g)} w_{ij}}. \tag{86}$$

All possible subgraphs of the weighted graph can be categorized into sets of *topologically equivalent* subgraphs.† Let M be one such set of topologically equivalent subgraphs. The intensity of M is given by $I_M = \sum_{g\in M} I(g)$ and its coherence by $\Psi_M = \sum_{g\in M} \Psi(g)$. An intensity score ZI_M can be accordingly defined by

$$ZI_M = \frac{I_M - \langle I_M^{(\text{rand})}\rangle}{\sigma_{I_M}^{(\text{rand})}} \tag{87}$$

and the coherence score,

$$Z\Psi_M = \frac{\Psi_M - \langle \Psi_M^{(\text{rand})}\rangle}{\sigma_{\Psi_M}^{(\text{rand})}}, \tag{88}$$

where $\langle I_M^{(\text{rand})}\rangle$ and $\sigma_{I_M}^{(\text{rand})}$ are the mean and the standard deviation of the intensities in a randomized graph ensemble; $\langle \Psi_M^{(\text{rand})}\rangle$ and $\sigma_{\Psi_M}^{(\text{rand})}$ are the average and the standard deviation of the coherence in the randomized ensemble. When the network is transformed to its unweighted version, ZI_M and $Z\Psi_M$ tend to Z (see equation (83)).

12.3. *Subgraph centrality*

A way to quantify the centrality of a vertex based on the number of subgraphs in which the vertex takes part has been proposed [193]. The respective measurement, called *subgraph centrality*, considers the number of subgraphs that constitute a closed walk starting and ending at a given vertex i, with higher weights given to smaller subgraphs. This measurement is related to the moments of the adjacency matrix, equation (68):

$$SC_i = \sum_{k=0}^{\infty} \frac{(A^k)_{ii}}{k!}, \tag{89}$$

where $(A^k)_{ii}$ is the ith diagonal element of the kth power of the adjacency matrix A, and the factor $k!$ assures that the sum converges and that smaller subgraphs have more weight in the sum. Subgraph centrality can be easily computed [193] from the spectral decomposition of the adjacency matrix,

$$SC_i = \sum_{j=1}^{N} v_j(i)^2 e^{\lambda_j}, \tag{90}$$

†Two subgraphs are topologically equivalent if the only difference is the weight of the existing edges.

where λ_j is the jth eigenvalue and $v_j(i)$ is the ith element of the associated eigenvector. This set of eigenvectors should be orthogonalized. The subgraph centrality of a graph is given by [129]:

$$SC = \frac{1}{N}\sum_{i=1}^{N} SC_i = \frac{1}{N}\sum_{i=1}^{N} e^{\lambda_j}. \tag{91}$$

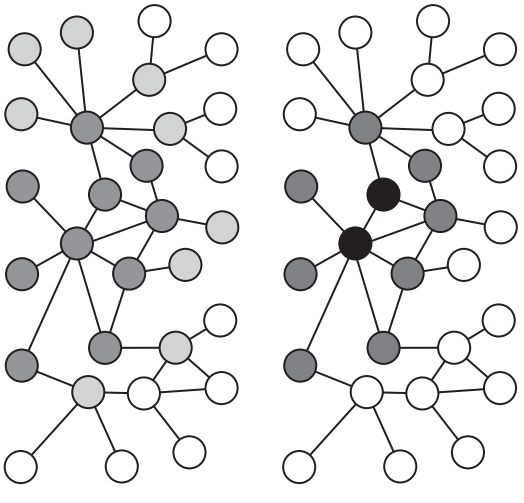
13. Hierarchical measurements

Using concepts of mathematical morphology [194–197], it is possible to extend some of the traditional network measurements and develop new ones (e.g. [198–200]). Two fundamental operations of mathematical morphology are *dilation* and *erosion* (see figure 15). Given a subgraph g of a graph G, the complement of g, denoted $\bar{g}$, is the subgraph implied by the set of vertices in G that are not in g,

$$\mathcal{N}(\bar{g}) = \mathcal{N}(G) \backslash \mathcal{N}(g)$$

($\backslash$ is the operator of set difference). The dilation of g is the subgraph $\delta(g)$ implied by the vertices in g plus the vertices directly connected to a vertex in g. The erosion of g, denoted $\varepsilon(g)$, is defined as the complement of the dilation of the complement of g:

$$\varepsilon(g) = \overline{\delta(\bar{g})}.$$



(a) (b)

Figure 15. Example of morphological operations: (a) Dilation: the dilation of the initial subnetwork (dark gray vertices) corresponds to the dark and light gray vertices; (b) Erosion: the erosion of the original subnetwork, given by the dark gray vertices in (a), results in the subnetwork represented by the black vertices in (b).

These operations can be applied repeatedly to generate the d-dilations and d-erosions:

$$\delta_d(g) = \underbrace{\delta(\delta(\ldots(g)\ldots))}_{d}, \tag{92}$$

$$\varepsilon_d(g) = \underbrace{\varepsilon(\varepsilon(\ldots(g)\ldots))}_{d}. \tag{93}$$

The first operation converges to the entire network G and the second converges to an empty network.

The d-ring of subgraph g, denoted $R_d(g)$, is the subgraph implied by the set of vertices

$$\mathcal{N}(\delta_d(g))\backslash\mathcal{N}(\delta_{d-1}(g));$$

the rs-ring of g, denoted $R_{rs}(g)$, is the subgraph implied by

$$\mathcal{N}(\delta_s(g))\backslash\mathcal{N}(\delta_{r-1}(g)).$$

Note that $R_d(g) = R_{dd}(g)$. The same definitions can be extended to a single vertex considering the subgraph implied by that vertex, and to an edge considering the subgraph formed by the edge and the two vertices that it connects. In the case of a single vertex i the abbreviations $R_d(i)$ and $R_{rs}(i)$ are used. For example, in figure 16, $R_1(15)$ includes the vertices {8, 14, 16, 17}; $R_2(15)$ includes {1, 13, 18, 19}; for the graph g implied by the vertices {1, 15, 22} (in black), $R_1(g)$ includes the vertices in white: {2, 3, 4, 5, 6, 7, 8, 9, 14, 16, 17}.

The *hierarchical degree* of a subgraph g at distance d, henceforth represented as $k_d(g)$, can be defined as the number of edges connecting rings $R_d(g)$ to $R_{d+1}(g)$. Note that $k_0(i)$ is equal to k_i.

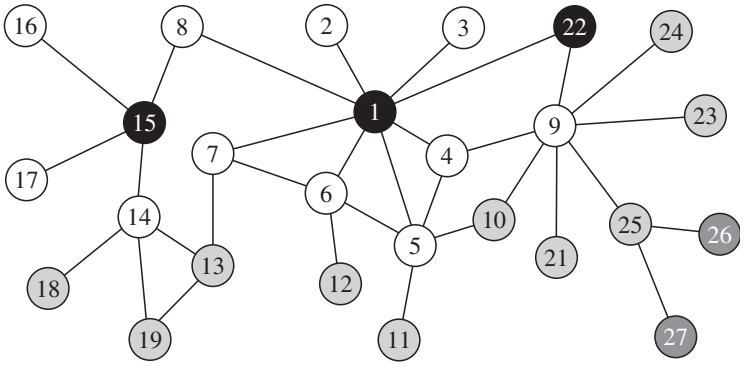


Figure 16. The subgraph of interest is defined by black vertices, $g = \{1, 15, 22\}$. The first hierarchical level of g is given by the first dilation around g, represented by the white vertices; the second hierarchical level is obtained dilating the subnetwork again, represented by the gray vertices. The hierarchical degree of the first level is given by the number of edges from white to light gray vertices, $k_1(g) = 12$, and the hierarchical degree of the second level is the number of edges from light gray to dark gray vertices, $k_2(g) = 2$.

Another measurement which can be hierarchically extended is the clustering coefficient. The *rs-clustering coefficient* of g, $C_{rs}(g)$, can be defined as the number of edges in the respective *rs*-ring n_{rs}, divided by the total of possible edges between the vertices in that ring, i.e. for undirected networks

$$C_{rs}(g) = \frac{2n_{rs}(g)}{|\mathcal{N}(R_{rs}(g))|(|\mathcal{N}(R_{rs}(g))| - 1)}. \tag{94}$$

Other possible hierarchical measurements are briefly described in the following. The *convergence ratio* at distance d of subgraph g, $cv_d(g)$, corresponds to the ratio between the hierarchical subgraph degree at distance $d-1$ and the number of vertices in the ring at distance d; it can be understood as the average number of edges received by each vertex in the hierarchical level d from the previous level,

$$cv_d(g) = \frac{k_{d-1}(g)}{|\mathcal{N}(R_d(g))|}. \tag{95}$$

It is also possible to define the *divergence ratio*, which corresponds to the reciprocal of the convergence ratio

$$dv_d(g) = \frac{|\mathcal{N}(R_d(g))|}{k_{d-1}(g)}. \tag{96}$$

14. Fractal dimension

Fractals are objects or quantities that display self-similarity (or self-affinity) in all scales. For complex networks, the concept of self-similarity under a length-scale transformation was not expected because of the small world property, implying the average shortest path length of a network increases logarithmically with the number of vertices. However, Song *et al.* [201] analyzed complex networks by using fractal methodologies and verified that real complex networks may consist of self-repeating patterns on all length scales.

In order to measure the fractal dimension of complex networks, a *box counting method* and a *cluster growing method* has been proposed [201]. In the former, the network is covered with N_B boxes, where all vertices in each of them are connected by a minimum distance smaller than l_B. N_B and l_B are found to be related by

$$N_B \sim l_B^{-d_B}, \tag{97}$$

where d_B is the *fractal box dimension* of the network.

For the cluster growing method, a seed vertex is chosen at random and a cluster is formed by vertices distant at most l from the seed. This process is repeated many times and the average mass of resulting clusters is calculated as a function of l, resulting in the relation

$$\langle M_c \rangle \sim l^{d_f}, \tag{98}$$

where the average mass $\langle M_c \rangle$ is defined as the number of vertices in the cluster and d_f is the *fractal cluster dimension*.

For a network whose vertices have a typical number of connections, both exponents are the same, but this is not the case for scale-free networks.

Another scaling relation is found with a renormalization procedure based on the box counting method [201]. A renormalized network is created with each box of the original network transformed into a vertex and two new vertices are connected if at least one edge exists between vertices of the corresponding boxes in the original network. By considering the degree k' of each vertex of the renormalized network versus the maximum degree k in each box of the original network we have that:

$$k' \approx l_B^{-d_k} k, \tag{99}$$

The exponents γ (of the power law of the degree distribution), d_B and d_k are related by [201]:

$$\gamma = 1 + \frac{d_B}{d_k}. \tag{100}$$

Thus, scale-free networks, characterized by the exponent γ, can also be described by the two length invariant exponents d_B and d_k.

15. Other measurements

This section describes additional, complementary measurements related to network complexity, edge reciprocity and matching index.

15.1. *Network complexity*

It might be of interest to quantify the 'complexity' of a network. Lattices and other regular structures, as well as purely random graphs, should have small values of complexity. Some recent proposals are briefly presented below.

Machta and Machta [202] proposed the use of the computational complexity of a parallel algorithm [203] for the generation of a network as a complexity measurement of the network model. If there is a known parallel algorithm for the generation of the network of order $\mathcal{O}(f(N))$, with $f(x)$ a given function, then the complexity of the network model is defined as $\mathcal{O}(f(N))$. For example, Barabási-Albert networks can be generated in $\mathcal{O}(\log \log N)$ parallel steps [202].

Meyer-Ortmanns [204] associated the complexity of the network with the number of topologically non-equivalent graphs generated by splitting vertices and partitioning the edges of the original vertex among the new vertices, the transformations being restricted by some constraints to guarantee the generation of valid graphs.

The *off-diagonal complexity*, proposed by Claussen [205] is defined as an entropy of a specially defined vertex-vertex edge correlation matrix. An element with indexes (k, l) of this matrix has contributions from all edges that connect a vertex of degree k to a vertex of degree l (only values $k > 1$ are used).

15.2. *Edge reciprocity*

i is linked to vertex j, is vertex j also linked to vertex i? Such information helps to obtain a better characterization of the network, can be used to test network models against real networks and gives indication of how much information is lost when the direction of the edges is discarded (e.g. for the computation of some measurements that only apply to undirected networks).

A standard way to obtain information about reciprocity is to compute the fraction of bilateral edges:

$$\varrho = \frac{\sum_{ij} a_{ij} a_{ji}}{M}, \tag{101}$$

where M is the total number of edges.

A problem with this measurement is that its value is only relevant with respect to a random version of the network, as networks with higher connectivity tend to have a higher number of reciprocal edges due exclusively to random factors. Garlaschelli and Loffredo [206] proposed the use of the correlation coefficient of the adjacency matrix:

$$\rho = \frac{\sum_{ij} (a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{ij} (a_{ij} - \bar{a})^2}, \tag{102}$$

where $\bar{a}$ is the mean value of the elements of the adjacency matrix. This expression, known as *edge reciprocity*, simplifies to

$$\rho = \frac{\varrho - \bar{a}}{1 - \bar{a}}. \tag{103}$$

This value is an absolute quantity, in the sense that values of ρ greater than zero imply larger reciprocity than the random version (*reciprocal* networks), while values below zero imply smaller reciprocity than a random network (*antireciprocal* networks). This concept can be easily extended to weighted networks by substituting a_{ij} for w_{ij} in the above expressions.

15.3. *Matching index*

A *matching index* can be assigned to each edge in a network in order to quantify the similarity between the connectivity of the two vertices adjacent to that edge [207]. A low value of the matching index identifies an edge that connects two dissimilar regions of the network, thus possibly playing an important role as a shortcut between distant network regions [207]. The matching index of edge (i, j) is computed as the number of matching connections of vertices i and j (i.e. connections to the same other

vertex k), divided by the total number of connections of both vertices (excluding connections between i and j),

$$\mu_{ij} = \frac{\sum_{k \neq i,j} a_{ik} a_{jk}}{\sum_{k \neq j} a_{ik} + \sum_{k \neq i} a_{jk}}. \quad (104)$$

For directed networks, matching connections are only those in the same direction, and incoming and outgoing connections of vertices i and j should be considered separately. The matching index has also been adapted to consider all the immediate neighbors of a node, instead of a single edge [208].

16. Measurements of network dynamics and perturbation

This section covers two important related issues, namely the use of trajectories to characterize the dynamical evolution of complex networks connectivity and a brief discussion about the sensitivity of measurements to perturbations [209].

16.1. *Trajectories*

As indicated in the Introduction of this work, in the following we analyze the behavior of trajectories (see figure 17) defined by tuples of measurements as the analyzed network undergoes progressive modification, such as during their growth. The network models considered for the illustration of trajectories include: Erdős-Rényi random graphs (ER), random networks with community structure (CN), Watts-Strogatz small-worlds (WS), Geographical Networks (GN), and Barabási and Albert scale-free networks (BA) (see section 3 for a description of these models). The number of vertices considered was 500, 1000 and 2000, and the number of edges varies so that the average vertex degree ranges from 4 to 204, increasing by steps of 20. In the case of the GN model, the vertices were randomly distributed through a square box of unit size. The λ parameter in equation (12) was adjusted in order to guarantee the desired average degree. The CN networks included four communities interconnected by using $p_{\text{out}}/p_{\text{in}}$ (see section 3) equal to 5%, 10% and 15%. In the WS model, the probability p of rewiring the edges was 0.0002, 0.02 and 0.1. For the sake of better visualization, the trajectories of the WS and CN models were drawn separately from the other cases. The direction of evolution of the trajectories as more edges are included is indicated by arrows in figure 17. These results are discussed subsequently with respect to several pairs of measurements.

16.1.1. Average clustering coefficient and average shortest path length. By inspecting the trajectories associated to this pair of measurements (see figure 17(a)), two distinct behaviors can be identified. First, the average clustering coefficient $\tilde{C}$ exhibits a high variation while the average shortest path length ℓ remains almost constant with addition of edges for the ER, CN and BA models. Second, an opposite effect is observed for GN and WS models. In the latter case, the ℓ value undergoes a steep decrease, while staying almost constant for the other network models. This effect is related to the fact that GN and WS models are formed by vertices that tend to link to closer neighbors. Hence, with the addition of edges the number of long-range connections may decrease ℓ while $\tilde{C}$ remains almost unchanged. Furthermore, ℓ

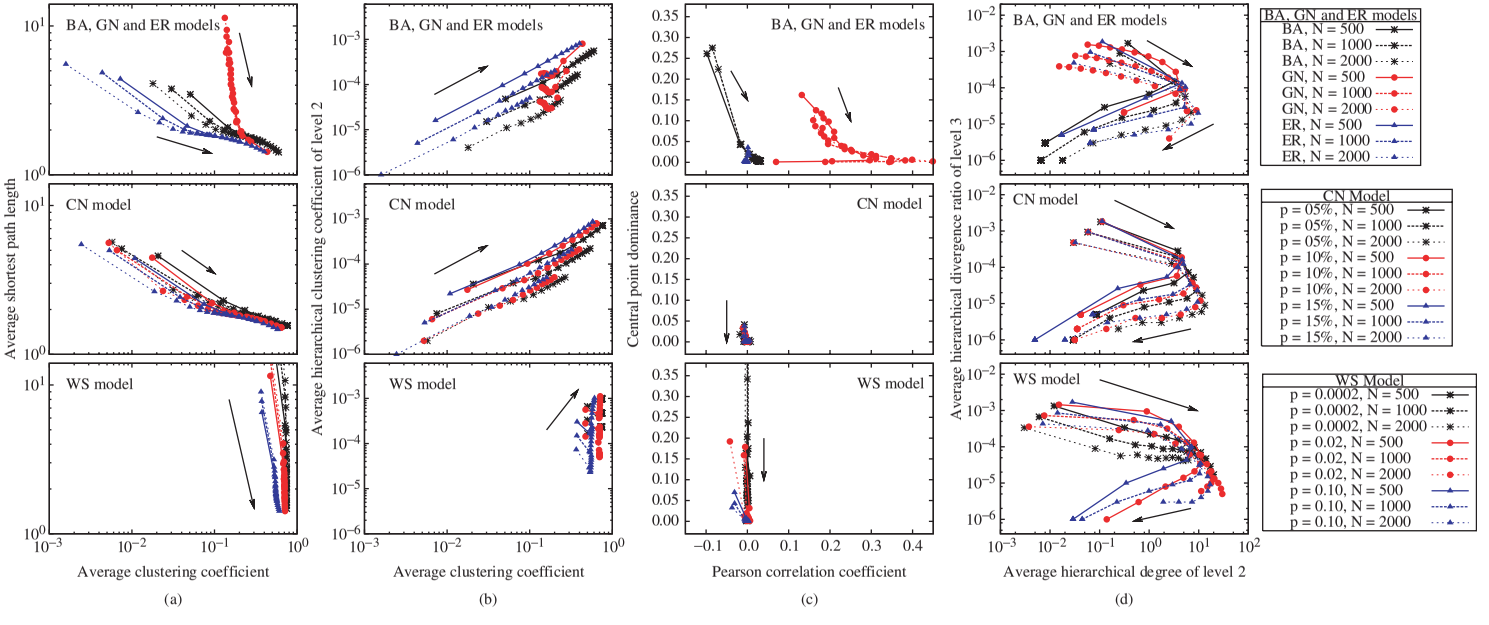


Figure 17. Trajectories defined by pairs of measurements. Each point corresponds to 10 network model realizations. Network sizes used are 500, 1000, and 2000; average degrees vary from 4 to 204 in steps of 20; for the community model, p_{out}/p_{in} is 5%, 10%, and 15%; for the WS model the rewiring probability values are 0.0002, 0.02, and 0.1.

decreases faster for WS model than for GN, while $\tilde{C}$ remains larger for the former model than in the other cases. This effect can be explained by the fact that the WS model is more regular than the GN and has larger $\tilde{C}$.

In the case of ER, CN and BA models, the values of ℓ and $\tilde{C}$ are smaller than for the other models. For ℓ the connections are not limited by proximity, adjacency or geography. At the same time, loops of order three appear when new edges are added to them, increasing $\tilde{C}$.

Another interesting fact observed from figure 17(a) is that all curves converge on the same point, corresponding to fully connected graphs, as the networks become denser. Therefore, ℓ and $\tilde{C}$ tend to unit value at that stage.

16.1.2. Average clustering coefficient and average hierarchical clustering coefficient of second level. The combination of $\tilde{C}$ and the average hierarchical clustering coefficient of second level† $\langle C_2(i)\rangle$, where the average is taken over all vertices in the network, see figure 17(b), tends to follow a power law for all trajectories except in the case of the GN and WS models, whose curves have a minimum value for $\langle C_2(i)\rangle$. Nonetheless, the highest growth rate is observed for the trajectories of WS model after the minimum value of $\langle C_2(i)\rangle$ is reached.

Another interesting characteristic of this combination of measurements is that $\tilde{C}$ is greater than $\langle C_2(i)\rangle$. This can be explained by the fact of $\langle C_2(i)\rangle$ is related to the presence of loops of order five without additional connections between their vertices [200]. Since loops of higher orders are less likely to appear in the considered networks, $\tilde{C}$ tends to become larger than $\langle C_2(i)\rangle$.

16.1.3. Pearson correlation coefficient and central point dominance. For all considered network models, except for the GN case, the Pearson correlation coefficient of vertex degrees represented by r (section 6), is close to zero even with the addition of new edges, as can be seen in figure 17(c), which shows the trajectories defined by the pair of measurements r and central point dominance CPD (section 9) as the average degree increases. This property can be explained by the fact that in ER, CN and WS models the edges are placed irrespectively to vertex degree, while the BA model is based on preferential growth [117], which leads to non-assortative mixing (i.e. no correlation between vertex degrees). The r value for the GN model is greater than zero in almost all cases because its growing dynamics is based on the geographic proximity of vertices. As the position of vertices is randomly chosen, some regions may by result be highly populated, implying the respective vertices have a high probability of becoming highly interconnected. On the other hand, vertices belonging to the regions barely populated have small chances to become "hubs" while still having a good chance of being connected. These two opposite behaviors tend to imply a r value greater than zero.

The central point dominance is a measurement of the maximum betweenness of any point in the network [143] (see section 9 for further details). By observing figure 17(c), one can see that most network models exhibit average values of this measurement close to zero, except for the BA, GN and WS cases. In BA networks,

†Note that $\tilde{C}$ is identical to the average hierarchical clustering coefficient at the first level.

values significantly larger than zero only occur in the beginning of the growth process (i.e. in the presence of few edges).

For WS models, the way in which they are normally constructed (see section 3) directly contributes to producing a network with modular structure, hence a high *CPD* value. Nevertheless, when new edges are added, the network gets denser and the value of this measurement goes to zero. In CN models, the *CPD* coefficient depends on the relation between the average vertex degrees inside and outside communities, i.e. when the network is highly modular, the *CPD* value tends to become larger.

16.1.4. Average hierarchical degree of second level and average hierarchical divergence ratio of third level. As shown in figure 17(d), all curves obtained for the average hierarchical degree of second level† $\langle k_2(i)\rangle$ and the average hierarchical divergence ratio of level three $\langle dv_3(i)\rangle$ have similar behavior. When the networks are sparse and new edges are added, increasing the average vertex degree, the average hierarchical counterpart increases until a maximum value. Afterwards, since the networks have a finite size, further increase of the connectivity tends to reduce the number of hierarchical levels in the networks and, as consequence, the average hierarchical vertex degrees of levels higher than one tend to decrease. The hierarchical divergence ratio of level three decreases with larger average vertex degree.

16.1.5. Discussion. As presented in figure 17, each measurement is specifically sensitive to the effects of addition of new edges to a network. Interestingly, the sensitivity also depends strongly on the network model. Some trajectories were closer to one another for specific network models as a consequence of inherent structural similarities. This effect is particularly pronounced in trajectories defined by the average clustering coefficient and average shortest path length, where two classes of trajectories appear, one for ER, BA and CN, and another for GN and WS.

The analysis of network dynamics provides insights about model similarities. If network trajectories evolve in a similar fashion, it is possible to infer that these networks have similar structure concerning the respective pair of measurements. However, for other measurements, this similarity may be weaker or non-existent. For instance, in the space defined by the average clustering coefficient and average shortest path length, the curves obtained for ER and BA evolve in similar fashion. This behavior is not observed in the space defined by the central point dominance and Pearson correlation coefficient. Also, by inspecting the trajectories, it is possible to determine the correlation between measurements during the network's evolution. For instance, the dynamics of the average clustering coefficient and the average hierarchical clustering coefficient of second level present correlation for ER, BA and CN.

The trajectory-based study described here can be immediately extended to real network analysis and modeling. In the case of the WWW, for instance, by inspecting its evolution in the measurements space it is possible to develop more precise models

†Notice that $\langle k_2(i)\rangle$ (average taken over all vertices i in the network) depends on the network connectivity.

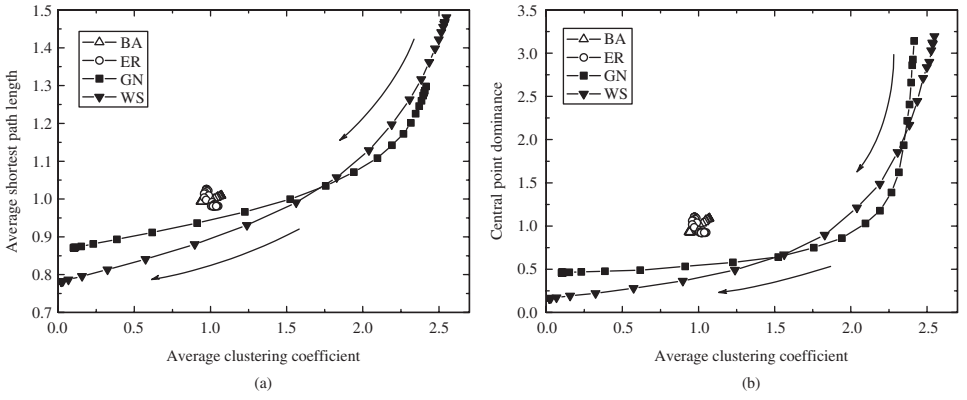


Figure 18. Example of perturbations. Each point corresponds to 100 realizations of networks with $N = 1000$ and $\langle k \rangle = 6$.

to represent and characterize its structure. For citation networks, it is possible to characterize the networks generated for different knowledge areas and obtain insights about their structure and evolution. All in all, trajectories provide a visually clear and accessible interpretation about the evolution of complex networks connectivity and dynamics.

16.2. *Perturbation analysis*

Another important property of a given measurement relates to how much it changes when the networks undergo small *perturbations* (e.g., rewiring, edge or vertex attack, weight changes, etc.). For instance, the shortest path length provides a good example of a particularly sensitive measurement, in the sense that the modification of a single connection may have great impact in its value. The quantification of the sensitivity of measurements to different types of perturbations and networks therefore provides valuable information to be considered for characterization, analysis and classification of complex networks.

Interesting insights can be obtained as far as this subject is concerned by performing progressive perturbations in a specific network and observing the respective relative variations in the measurement of interest. Figure 18 shows the trajectories obtained in the $\{\tilde{C}, \ell\}$ and $\{\tilde{C}, CPD\}$ feature spaces shown in figure 18(a) and (b), respectively, while considering the rewiring method [81] (see also section 3.3) progressively performed for the BA, ER, GN and WS models ($N = 1000$, $\langle k \rangle = 6$). The values shown in this figure were normalized through division by the respective averages of the measurements in order to provide suitable visual comparison.

Each successive point along the trajectories, which are indicated by arrows in figure 18, was obtained after a number of rewirings which are successive integer powers of two (for the sake of obtaining more uniform visualization). It is clear from the obtained results, that the sensitivities of the two pairs of measurements vary substantially with respect to the type of network under consideration. More specifically, the widest variations were observed for the GN and WS models. The stable trajectories obtained for the shortest path length in the case of the ER and BA

models are direct consequences of the fact that these networks are inherently characterized by low overall average shortest path length. A similar situation is verified for the clustering coefficient, which tends to be small in those two types of networks. The evolution of the trajectory regarding the clustering coefficient and average shortest path length for the GN and WS networks is a direct consequence of the fact that the progressive edge rewiring tends to strongly reduce those two measurements. The marked difference of sensitivity of measurements to perturbations depending on the type of network model suggests that quantifications of the sensitivity (e.g. the standard deviation or entropy) can be potentially useful as additional measurements for network identification.

17. Correlation analysis

Although a virtually infinite number of measurements can be obtained for quantifying the connectivity of complex networks, a varying degree of redundancy will be observed between their pairwise combinations. For instance, the node degree and the clustering coefficient are uncorrelated for most networks [101]. Moreover, the intensity of such correlations may depend on the specific type of complex network. In the present work we quantify the degree of correlation between two measurements x and y in an ensemble of n networks, understood as two random variables, in terms of the Pearson correlation coefficient [68, 210], defined as

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2} \sqrt{\sum_{i=1}^{n} (y_i - \langle y \rangle)^2}}, \tag{105}$$

where x_i and y_i are the measurements of the ith network of the ensemble, and $\langle x \rangle$ and $\langle y \rangle$ are their respective averages. The value of the Pearson correlation varies from -1 (negative correlation) to 1 (positive correlation), with null value indicating that the two random variables are uncorrelated.

The correlation between two measurements can be related to redundancy, in the sense that two completely correlated features (e.g. $r_{xy} = 1$) provide no additional information than any of them taken separately. On the other hand, even a highly correlated pair of measurements can provide additional information for the characterization and separation between analyzed networks (e.g. [68]). Figure 19 shows an example of two-dimensional feature space involving two classes which, though highly correlated, provide valuable information for the separation between the two categories (e.g. by using the dashed line).

In order to quantify the redundancies between the considered measurements, we calculated the respective pairwise Pearson correlation coefficients for the BA, ER and GN models ($N = 1000$ and $\langle k \rangle = 4$). Table 3 shows the values obtained with respect to each model and also considering all models together ('All').

Several interesting facts can be inferred from this table. First, particularly high absolute values of correlations have been obtained for the BA model, with low absolute values observed for the ER and GN cases. This seems to represent a particularly interesting property of BA networks. Another interesting finding regards the fact that the correlations obtained for specific network models not necessarily agree

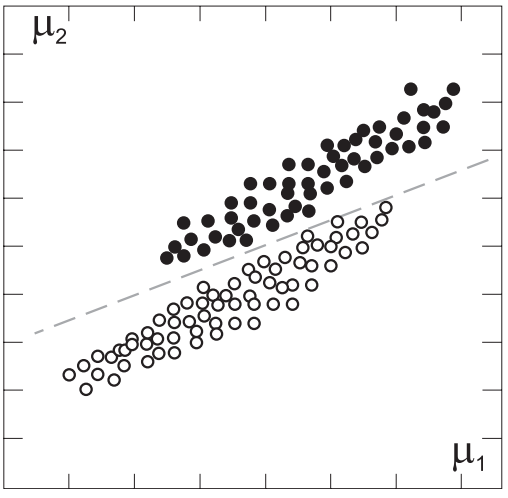


Figure 19. Example of two measurements which, though correlated, can still contribute to category identification.

with those obtained when the three models are considered together. This is the case, for instance, of the low correlation observed between the measurement average shortest path length ℓ and log-log degree distribution straightness st for each of the three individual models and high correlation otherwise obtained when these three models are considered jointly. This interesting behavior can be immediately explained by considering figure 20, which illustrates that three low correlation groups can result in global alignment, therefore implying the relatively strong overall negative correlation. Such situations indicate that the individual and global correlations can provide complementary information about different types of relationships.

It is also clear from the results in table 3 that particularly high correlations were obtained between the average shortest path length ℓ and the vertex degree at the second hierarchical level $\langle k_2(i) \rangle$. This fact suggests that this specific hierarchical vertex degree may be considered, at least for the three considered types of networks, as an estimation of the average shortest path length, allowing substantial computational saving. Another interesting result is that the highest correlations were obtained for the BA model, as a possible consequence of the presence of the respective hubs. For instance, the correlation between the average shortest path length and the average clustering coefficient was found to be equal to -0.63 for the BA models. This is a consequence of the fact that additional links tend to be established with the hubs and therefore contribute to higher clustering and shortest paths.

18. Multivariate statistical methods for dimensionality reduction and measurement selection

The intrinsic statistical variability of the connectivity of real and simulated complex networks, even when produced by the same process or belonging to the same class,

Table 3. Correlations between measurements for the BA, ER and GN models and "All" jointly. The values were estimated from 1000 realizations for each model of networks with $N = 1000$ and $\langle k \rangle = 4$.

		st	r	$\tilde{C}$	ℓ	CPD	$\langle k_2(i) \rangle$	$\langle C_2(i) \rangle$	$\langle dv_3(i) \rangle$
st	BA	1.00							
	ER	1.00							
	GN	1.00							
	All	1.00							
r	BA	−0.22	1.00						
	ER	−0.01	1.00						
	GN	−0.13	1.00						
	All	0.71	1.00						
$\tilde{C}$	BA	0.06	−0.29	1.00					
	ER	−0.01	0.07	1.00					
	GN	0.04	−0.00	1.00					
	All	0.31	0.82	1.00					
ℓ	BA	−0.01	0.38	−0.63	1.00				
	ER	−0.06	0.04	−0.08	1.00				
	GN	−0.10	0.02	0.03	1.00				
	All	0.69	0.96	0.88	1.00				
CPD	BA	−0.09	0.23	0.39	−0.58	1.00			
	ER	−0.61	0.10	0.03	0.07	1.00			
	GN	−0.05	−0.02	0.03	0.23	1.00			
	All	−0.87	−0.44	0.02	−0.41	1.00			
$\langle k_2(i) \rangle$	BA	0.01	−0.30	0.63	−0.99	0.60	1.00		
	ER	0.04	0.03	0.08	−0.90	−0.06	1.00		
	GN	0.08	0.28	−0.02	−0.65	−0.13	1.00		
	All	−0.96	−0.80	−0.43	−0.79	0.85	1.00		
$\langle C_2(i) \rangle$	BA	0.02	0.02	0.58	−0.74	0.59	0.76	1.00	
	ER	−0.03	0.04	0.45	−0.16	0.02	0.19	1.00	
	GN	−0.00	0.09	0.59	0.18	0.07	−0.11	1.00	
	All	0.37	0.86	0.99	0.91	−0.05	−0.49	1.00	
$\langle dv_3(i) \rangle$	BA	0.01	0.26	−0.57	0.91	−0.52	−0.94	−0.69	1.00
	ER	0.03	−0.10	−0.01	−0.25	−0.01	−0.16	−0.04	1.00
	GN	−0.02	−0.28	−0.09	−0.03	−0.00	−0.50	−0.21	1.00
	All	−0.14	−0.74	−0.97	−0.79	−0.18	0.27	−0.96	1.00

implies that sound characterization, comparison and classification of networks should take into account not only the average measurements, but also additional information about their variability including higher statistical moments (e.g., variance, kurtosis, etc.) as well as multivariate statistical distribution of the measurements. For example, realizations obtained by using the Barabási-Albert (BA) model with fixed parameters will produce networks which, though not identical, will have equivalent statistical distribution of their properties. Figure 21 shows a scatterplot obtained by considering 1000 realizations of the BA model with $N = 1000$ and $m = 3$ with respect to the measurements $(r, \tilde{C}, \ell)$, where r is the Pearson correlation coefficient of vertex degrees; $\tilde{C}$, the average clustering coefficient; and ℓ, the average shortest path length. Although the obtained points form a well-defined cluster around the average point (−0.0653, 0.0365, 3.255), there is a significant dispersion of cases

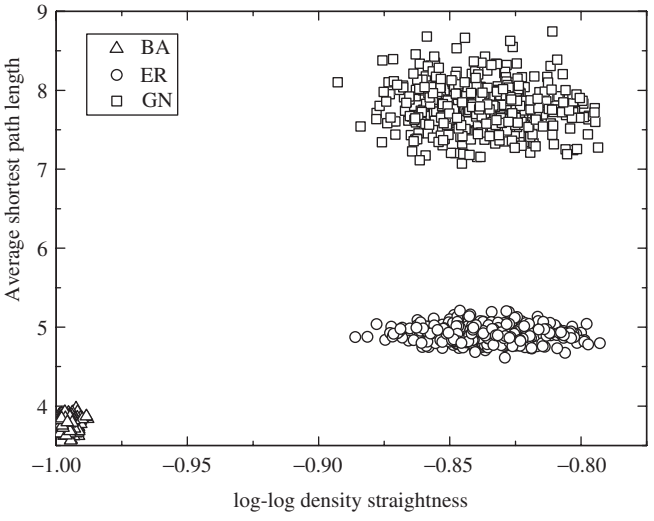


Figure 20. Example of scatterplot showing the low correlation between log-log degree density straightness and the average shortest path length for all the individual models and the high correlation for all models together. The networks have $N = 1000$ and $\langle k \rangle = 4$; 500 realizations of each model were used.

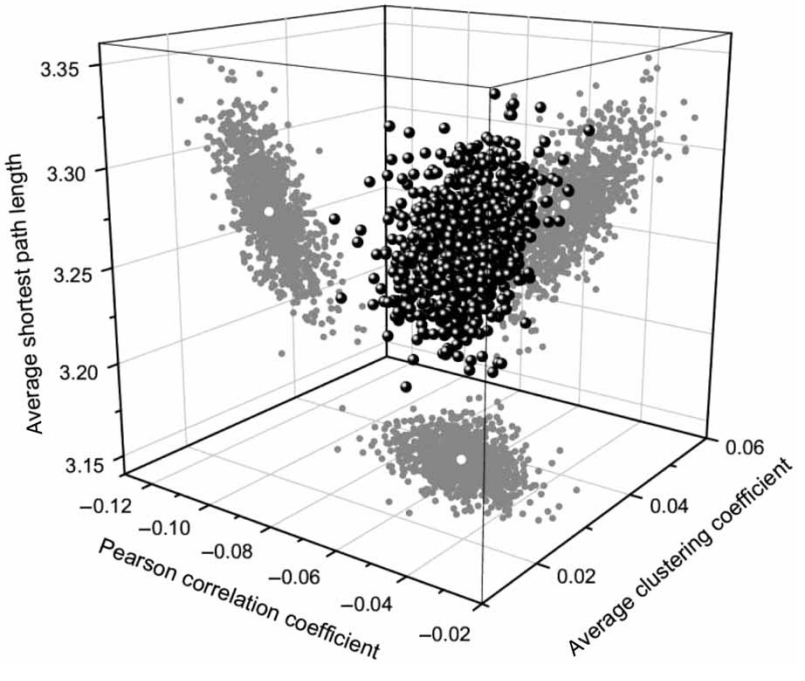


Figure 21. The spatial distribution in the $(r, \tilde{C}, \ell)$ phase space of 1000 different realizations of the BA model with $N = 1000$ and $m = 3$. The distribution has been projected onto the three main planes (gray shadows) for the purpose of better visualization. The *white circles* in the middle of the gray shadows represent the mean projected into those planes.

around this center, implying that additional statistical measurements other than the mean need to be used for proper characterization of the network under analysis. Therefore, any objective attempt at characterizing, comparing or classifying complex networks needs to take into account statistical distributions in phase spaces such as that in figure 21. Such an important task can be effectively accomplished by using traditional and well-established concepts and methods from *Multivariate Statistics* (e.g., [68, 69, 211]) and *Pattern Recognition* (e.g. [68, 69, 212]).

It should be observed that while too small sets of measurements can prove to be insufficient to characterize a network, many highly correlated features (as illustrated in the previous section) may not contribute substantially to the overall understanding of the connectivity. It is therefore interesting to consider statistical methods capable of reducing the dimensionality of the feature space while retaining the contribution of the more meaningful measurements. As far as the choice and interpretation of network measurements are concerned, two multivariate methods stand out as being particularly useful, namely *Principal Component Analysis* — PCA (e.g. [68, 69]) and *Canonical Variable Analysis* (e.g., [69, 211]). While the former procedure allows the reduction of the dimensionality of the measurement space, obtained in terms of projections so as to concentrate the variation of the data along the first new axes (i.e. those associated to the highest covariance matrix eigenvalues), the latter method implements such projections so as to achieve best separation, in terms of inter and intra-class distances (see below), between the involved classes of networks under analysis. In both these methods, the variables associated to each of the axes in the new, dimensionally reduced feature space, correspond to linear combinations of the original measurements. Consequently, some indication about the contribution of each measurement for the description of the statistical distribution of the studied networks can be obtained by considering the absolute values of the respective weights in the linear combination. Such a procedure can be applied in order to help identify the most meaningful measurements.

The current section presents and illustrates in a self-contained and accessible fashion these two dimensionality reduction methods from multivariate statistics (PCA and canonical analysis). The potential for applications of these methods is illustrated with respect to three reference complex network models — namely Erdős and Rényi random graph (ER), Barabási-Albert (BA) and Geographical Network model (GN), against which some real-world networks are classified.

18.1. *Principal component analysis*

Let the connectivity properties of a set of R complex networks, irrespective of their type or origin, be described in terms of P scalar measurements x_i, $i = 1, 2, \ldots, P$, organized as the feature vector $\vec{x} = (x_1, x_2, \ldots, x_P)^T$. The covariance matrix K can be estimated as

$$K = \frac{\left(\vec{x} - \langle\vec{x}\rangle\right)\left(\vec{x} - \langle\vec{x}\rangle\right)^T}{R}, \tag{106}$$

where $\langle\vec{x}\rangle$ is the average feature vector, each element of which corresponds to the average of the respective measurement. As K is a real and symmetric $P \times P$ matrix, a set of P decreasing eigenvalues λ_i and respectively associated eigenvectors $\vec{v}_i$ can

be obtained. Moreover, if all eigenvalues are distinct, the eigenvectors will be orthogonal.† These eigenvectors can be stacked to obtain the transformation matrix T, i.e.

$$T = \begin{bmatrix} \leftarrow & \overrightarrow{v_1} & \rightarrow \\ \leftarrow & \overrightarrow{v_2} & \rightarrow \\ & \cdots & \\ \leftarrow & \overrightarrow{v_P} & \rightarrow \end{bmatrix}. \tag{107}$$

The original feature vectors $\vec{x}$ can now be transformed into a new coordinates reference through the following linear transformation corresponding to axes rotation:

$$\vec{X} = T\vec{x} \tag{108}$$

which defines the *principal component projections*.

It can be shown [213] that the distribution of points in the new phase space obtained by the above transformation is such that the largest variance is observed along the first axis, followed by decreasing variances along the subsequent axes, with the initial axes being called *principal*. Such an important property allows, by considering only the principal eigenvalues, the original cloud of points to be projected along phase spaces of a smaller dimensionality p. In order to do so, the transformation matrix is constructed while taking into account only the first T_p eigenvectors associated to the largest eigenvalues, i.e.

$$T_p = \begin{bmatrix} \leftarrow & \overrightarrow{v_1} & \rightarrow \\ \leftarrow & \overrightarrow{v_2} & \rightarrow \\ & \cdots & \\ \leftarrow & \overrightarrow{v_p} & \rightarrow \end{bmatrix}. \tag{109}$$

Figure 22 shows the effect of projecting the cloud of points in figure 21 onto the two main axes so that the variance of the samples is maximized. Although useful for implementing dimensionality reduction — which favors visualization, redundancy reduction, and computational savings — the principal component analysis method is limited as it does not explicitly consider the category of each individual. This limitation is overcome in the canonical variable analysis described below.

18.2. *Canonical variable analysis*

The method known as *canonical variable analysis* provides a powerful extension of principal component analysis by performing the projections so as to optimize the separation between the known categories of objects. Before presenting the method, we introduce a series of scatter measurements from which the overall criterion for class separation is defined.

†Otherwise, orthogonal eigenvectors can still be assigned to repeated eigenvalues.

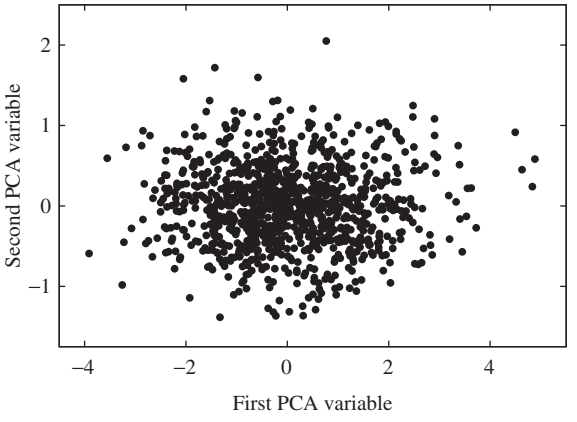


Figure 22. The principal component projection of the distribution of measurements in figure 21. Measurement values were first normalized by subtracting the corresponding mean value and dividing by the standard deviation to avoid biases due to the different absolute values. The first and second PCA variable have projecting vectors (−0.005, 0.707, −0.707) and (0.006, 0.707, 0.707) in the space defined by (r, C, ℓ) respectively.

Let us consider that the R complex networks of interest can be divided into N_c classes, each one with N_i objects and identified as C_i, $i = 1, 2, \ldots, N_c$, and that each object ξ is represented by its respective feature vector $\overrightarrow{x_\xi} = (x_1, x_2, \ldots, x_P)^T$ (see the previous section). The *total scatter matrix*, S, expressing the overall dispersion of the measurements [68] is defined as follows

$$S = \sum_{\xi=1}^{R} \left(\overrightarrow{x_\xi} - \langle \vec{x} \rangle\right)\left(\overrightarrow{x_\xi} - \langle \vec{x} \rangle\right)^T. \tag{110}$$

The *scatter matrix for each class* C_i is given as

$$S_i = \sum_{\xi \in C_i} \left(\overrightarrow{x_\xi} - \langle \vec{x} \rangle_i\right)\left(\overrightarrow{x_\xi} - \langle \vec{x} \rangle_i\right)^T, \tag{111}$$

where $\langle \vec{x} \rangle_i$ is the average feature vector of the class C_i.

The *intraclass scatter matrix*, accounting for the dispersion inside each of the classes, is defined as

$$S_{\text{intra}} = \sum_{i=1}^{N_c} S_i. \tag{112}$$

Finally, the *interclass scatter matrix*, characterizing the dispersion between each pair of classes, is given as

$$S_{\text{inter}} = \sum_{i=1}^{N_c} N_i \left(\langle \vec{x} \rangle_i - \langle \vec{x} \rangle\right)\left(\langle \vec{x} \rangle_i - \langle \vec{x} \rangle\right)^T. \tag{113}$$

It can be verified that

$$S = S_{\text{intra}} + S_{\text{inter}}. \tag{114}$$

The objective of the canonical analysis method is to maximize the interclass dispersion while minimizing the intraclass scattering (e.g. [211]). This can be achieved through the following linear transformation

$$\Gamma \overrightarrow{x}_{\xi} = \Gamma \overrightarrow{x_{\xi}}, \tag{115}$$

where $\Gamma = [\overrightarrow{\gamma_1}, \overrightarrow{\gamma_2}, \ldots, \overrightarrow{\gamma_P}]^T$ is chosen so that $\overrightarrow{\gamma_1}$ maximizes the ratio

$$\frac{\overrightarrow{\gamma_1}^T S_{\text{inter}} \overrightarrow{\gamma_1}}{\overrightarrow{\gamma_1}^T S_{\text{intra}} \overrightarrow{\gamma_1}}, \tag{116}$$

and $\overrightarrow{\gamma_j}, j = 2, 3, \ldots, P$, maximizes a similar ratio and

$$\overrightarrow{\gamma_j}^T S_{\text{intra}} \overrightarrow{\gamma_j} = 0. \tag{117}$$

It can be shown that the vectors $\overrightarrow{\gamma_1}, \overrightarrow{\gamma_2}, \ldots, \overrightarrow{\gamma_P}$ correspond to the eigenvectors of the matrix $S_{\text{intra}}^{-1} S_{\text{inter}}$.

Figure 23 illustrates a phase space of reduced dimensionality (a) containing two distributions of observations, as well as the respective PCA (b) and canonical analysis (c) projections considering two dimensions. The potential of the canonical approach for implementing dimensionality reduction while favoring well-separated clusters is evident from this example.

19. Bayesian decision theory for network classification

Another situation in multivariate statistics which is particularly important for complex network research concerns network identification. Indeed, it is often a critical issue to decide which of several reference models a given theoretical or experimentally obtained network belongs. This important problem can be approached in a sound way by using Bayesian decision theory [69], a well-established methodology which, provided good probabilistic models of the properties of the networks are available, allows near-optimal classification performance.†

The elegant and sound methodology known as *Bayesian decision theory* provides an intuitive and effective means for classifying objects into a given set of categories. In principle, it is assumed that the mass probabilities P_i, as well as the conditional probability densities, $p(\overrightarrow{x_{\xi}}|C_i)$, are all given or can be properly estimated (e.g., by using parametric or non-parametric methods, see [68, 69, 212]). The mass probability P_i corresponds to the probability that an object, irrespective of its properties, belongs to class C_i, and therefore can be estimated from the respective relative frequency. The conditional probabilities $p(\overrightarrow{x_{\xi}}|C_i)$ provide a statistical model of how the measurements in the feature vectors are distributed inside each category. Given an object

†Optimal performance is guaranteed in case the involved mass and conditional properties are completely known (see section 19 and [68, 69]).

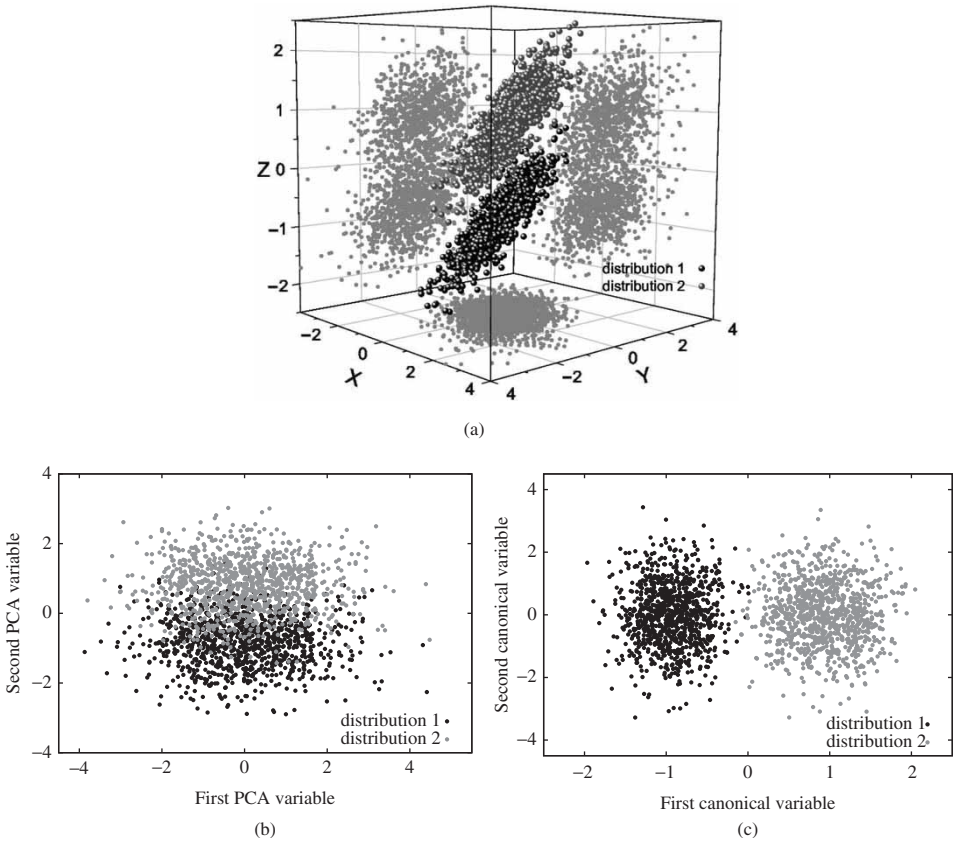


Figure 23. A phase space (scatterplot) containing two distributions of points (a) and respective PCA (b) and canonical (c) projections. Note that neither the projections into the three main planes (a) nor the PCA projection (b) can separate the distributions, which is suitably accomplished by the canonical projection (c).

with unknown classification, the most likely category c to be assigned to it is the one for which the respectively observed feature vector $\vec{x}$ produces the highest value of $P_\xi p(\vec{x}|C_\xi)$. In case the probability functions are not available, it is still possible to use approximate classification methods such as *k-nearest neighbors* (e.g. [69]), which consists of identifying the set of the k individuals which are closer (i.e. smaller distance between feature vectors) to the sample to be classified, and take as the resulting category that corresponding to the most frequent class among the nearest neighbors.

Let us illustrate the above concepts and methodology in terms of a situation involving three categories C_1, C_2 and C_3 of complex networks, namely Geographical Network (GN), Watts-Strogatz small-world network (WS) and Erdős and Rényi random graph (ER), characterized in terms of their normalized average shortest path length l and Pearson correlation coefficient of vertex degrees r. The corresponding scatterplot is shown in figure 24(a).

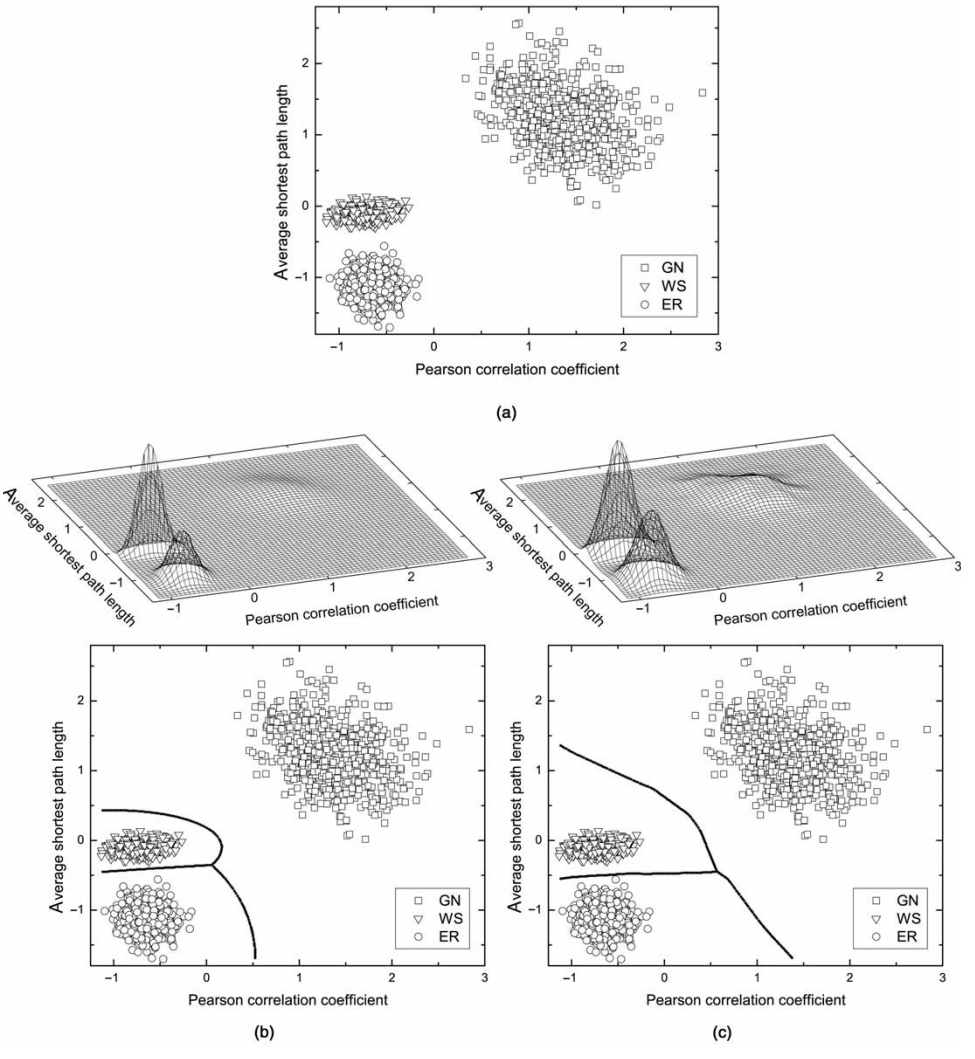


Figure 24. A scatterplot of normalized measurements containing several complex networks derived from three main categories, i.e., Geographical Network (GN), Watts-Strogatz (WS) and Erdős and Rényi random graph (ER) models (a), and respectively fitted decision regions obtained by using the Bayes method considering parametric (b) and non-parametric (c) estimation. The network parameters are $N = 250$, $\langle k \rangle = 20$, with 1000 realizations for each model; the rewiring probability in the WS model is 0.4.

Usually, we do not know the mass and conditional probabilities of each type of networks, so they have to be estimated from the available data. This stage can be understood as the *training phase* of the Bayesian decision theory method. There are two main ways to estimate the probabilities required: parametric and non-parametric. In the former, the mathematical form of the probability functions is known (e.g., normal distribution) and the respective parameters (mean and

covariance matrix, in the case of normal distributions) need to be estimated; in the latter, the mathematical type of the densities is unknown, being estimated, e.g., through some interpolation procedure such as the Parzen windows methodology [69]. In this method, the original discrete distribution of observations in the features space is represented in terms of a sum of Dirac's deltas and convolved with a smooth function such as a Gaussian function. The final result of this operation is a sum of the Gaussian functions at each position of the Dirac's deltas, weighted by the respective amplitude of the latter functions. As such, the Parzen windows methodology provide a means for filling in the empty spaces between the original observation, allowing proper interpolation of the overall density function.

Once the training phase is concluded, new objects whose classes are to be determined have their measurements estimated and used to identify, among the probability distributions of the trained models, which are the most likely respective classes. This categorization procedure corresponds to the *decision phase* of the Bayesian methodology. The Bayes rule can then be expressed as:

$$\text{if } f(\vec{x}_i|c_a)P(c_m) = \max_{b=1,\,m}\{f(\vec{x}_i|c_b)P(c_b)\} \text{ then select } c_a, \tag{118}$$

where $\vec{x}_i$ is the vector that stores the network set of measurements and c_a is the class of networks associated to the model a.

Figure 24(b) illustrates the parametric approach, considering three normal density distributions, applied to the data in figure 24(a). These distributions were defined by having their parameters (namely average vector and covariance matrix) estimated from the respective experimental measurements. The separating frontiers are shown in the projection at the bottom of the figure. The decision regions obtained by using non-parametric estimation through Parzen windows are shown in figure 24(c).

Note that a very high dimensional feature space implies that a substantially high number of individuals must be considered in order to obtain properly estimated (i.e. not too sparse) densities. Therefore, it is essential to limit the number of measurements to a small set of more discriminative features. An interesting alternative involves the use of canonical projections in order to reduce the dimensionality of the problem. A key open question which is briefly addressed in this section regards which of the several topological measurements available for complex networks characterization can yield the best characterization and discrimination among the principal network models.

19.1. *Combining canonical variable analysis and bayesian decision theory*

An interesting possibility for classifying networks involves the application of canonical variable analysis followed by Bayesian decision theory (e.g., [68, 211, 214]). More specifically, the observations considered for the training stage are projected into a reduced dimensional feature space by using canonical analysis, so that the Bayesian decision method is applied not over the larger original features space, but onto a more manageable and representative features space. This possibility is explored in this section in order to address the important issue of classifying experimental complex networks into three reference categories defined by the Barabási-Albert (BA), Erdős and Rényi random graph (ER) and Geographical

Network (GN) models. The following experimental networks are considered in our experiments:

US Airlines Transportation Network (USATN): The USATN is composed by 332 US airports in 1997, connected by flights. The data was collected from the Pajek datasets [215]. This kind of network exhibits a power law behavior as described in [89, 216].

Protein-Protein Interaction Network of Saccharomyces Cerevisiae (PPIN): PPIN is formed by 1922 proteins linked according to identified direct physical interactions [217]. A dataset is available at the Center for Complex Network Research (The University of Notre Dame). The vertex degree distributions of protein-interaction networks tend to follow a power law [217].

Autonomous System (AS): In the Internet, an AS is a collection of IP networks and routers under the control of one entity that presents a common routing policy to the Internet. Each AS is a large domain of IP addresses that usually belongs to one organization such as a university, a business enterprise, or an Internet Service Provider. In this type of networks, two vertices (AS) are connected if there is at least one physical link between them. This kind of network has been described as corresponding to the Barabási-Albert model [218]. The data considered in our work is available at the web site of the National Laboratory of Applied Network Research (http://www.nlanr.net). We used the data collected in Feb. 1998, with the network containing 3522 vertices and 6324 edges.

Transcriptional Regulation Network of the E. coli (TRNE): In this network, the vertices represent operons (an operon is a group of contiguous genes that are transcribed into a single mRNA molecule) and each edge is directed from an operon that encodes a transcription factor to another operon that is regulated by that transcription factor. Hence, this kind of network, which is believed to be scale free [189], plays an important role in controlling gene expression. We used the undirected version of the network analyzed by Shen-Orr *et al.* [189], which is formed by 577 interactions and 424 operons. The original network was transformed into the undirected form by the operation of *symmetry* described in section 2.

Delaunay Network (DLN): This network was obtained by distributing a set of points (the vertices) uniformly (but with an exclusion radius in order to avoid too close points) along a unit square and obtaining the edges from the respective Delaunay triangulation (e.g. [219]). Therefore, each point defines a tile in the respective Voronoi diagram,† and every pair of adjacent vertices are connected (see figure 25). The connectivity of this type of geometrical structure, henceforth called *Delaunay network*, is therefore completely determined by the adjacency between the vertices, which is in turn defined by the geographical distribution of the vertices. As such, Voronoi networks provide one interesting extreme case of geographical networks where only the immediate spatial neighborhood is considered for connection. The network considered here contains 251 vertices and 700 edges.

†Each Delaunay triangulation has as dual a Voronoi tessellation. Each vertex in the former structure is associated to one of the sides of the Voronoi cells, and vice-versa (e.g. [219]).

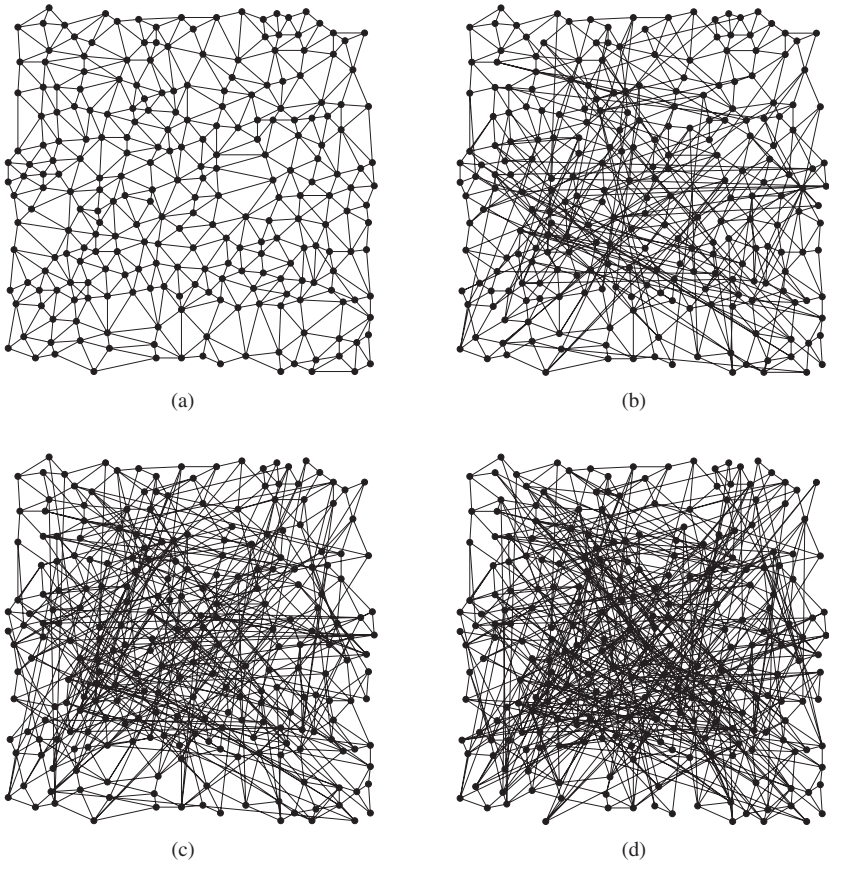


Figure 25. The Delaunay geographical network (DLN) for several numbers of random rewirings: original (a) and after 60 (b), 120 (c) and 200 (d) rewirings.

Progressively rewired (degree preserving) versions of this network were also considered in order to illustrate the evolution of trajectories in decision spaces. Figure 25 illustrates four of these successive configurations.

A total of three sets of 300 realization of each reference model (BA, ER and GN) were generated. The networks for each set were designed to have average vertex degrees near the experimental value. The model and experimental networks were characterized in terms of the following measurements: straightness st, average vertex degree $\langle k \rangle$, Pearson correlation coefficient of vertex degrees r, average clustering coefficient $\tilde{C}$, average shortest path length ℓ, central point dominance CPD, average hierarchical degree of second level $\langle k_2(i) \rangle$, average hierarchical clustering coefficient of second level $\langle C_2(i) \rangle$ and average hierarchical divergence ratio of the third level $\langle dv_3(i) \rangle$.

In order to provide a general and representative view of the effect of these measurements in the classification of real networks, we considered the following combinations of measurements:

(i) $\{\ell, st\}$,
(ii) $\{\langle k \rangle, \tilde{C}, \ell\}$,

(iii) $\{\langle k_2(i)\rangle, \langle C_2(i)\rangle, \langle dv_3(i)\rangle\}$,
(iv) $\{st, r, CPD\}$,
(v) $\{\langle k\rangle, \tilde{C}, \ell, st, r, CPD\}$,
(vi) $\{\langle k\rangle, \tilde{C}, \ell, \langle k_2(i)\rangle, \langle C_2(i)\rangle, \langle dv_3(i)\rangle\}$,
(vii) $\{st, r, CPD, \langle k_2(i)\rangle, \langle C_2(i)\rangle, \langle dv_3(i)\rangle\}$,
(viii) all measurements.

Table 4 shows the results, i.e. the theoretical model and respective average vertex degree which have been associated to each experimental network by the classification procedure, obtained for each of these configurations. More specifically, each experimental network was classified as having the same category as the theoretical model defining the decision region in the canonical projection space where the feature vector of the experimental data was mapped.

A number of interesting facts can be inferred from table 4. To begin with, the compatibility between the type of network model expected and obtained for each of the experimental networks varies considerably for each case. The best compatibility was obtained for the DLN, i.e. the identified model was compatible with the expected type (geographical) for all considered combinations of measurements. Compatible average vertex degrees have also been obtained for cases (iii), (vi)–(viii). Figure 26 illustrates the location of this network in the scatterplot defined by the canonical projection of the combination of all measurements. In this figure, which also shows the separating frontiers of the decision regions, the experimental network DLN (represented as ♦) resulted closer to GN with average vertex degree of 6. PPIN implied the highest number of incompatible classifications which, instead of being identified as a BA network (as could be expected [217]), was understood as GN except for the cases $\{\ell, st\}$ and $\{st, r, CPD\}$. A similar situation was verified regarding the average vertex degrees. Figures 27(c) and (d) show the resulting position of this network within the scatterplots obtained by canonical projection of the combination of all measurements (c) and all except those hierarchical (d). Note the good agreement between the resulting categories obtained for these two cases. In both cases, the PPIN resulted very close to the GN with average vertex degree of 3.03.

A particularly interesting result has been obtained for the USATN, which tended to appear well away from all theoretical groups in most cases, as illustrated in the scatterplot shown in figure 27(a) with respect to the case $\{\langle k\rangle, \tilde{C}, \ell\}$. Intermediate results were obtained for the other networks. For instance, TRNE has been classified as expected (i.e. as a BA network) in 2 cases, identified as an ER in only one case and as a GN in 5 cases. Figure 27(b) shows the position of this network in the scatterplot defined for all measurements. Note that TRNE appears almost in the middle of the ER and GN types for average vertex degree of 2.45.

It is also possible to use hierarchical clustering algorithms (e.g. [68, 69, 200]) in order to obtain additional information about the relationship between the analyzed networks. Figure 28 shows the dendrogram obtained for the situation depicted in figure 27(c) by using Ward's agglomerative method. In this method the networks, initially treated as individual clusters, are progressively merged in order to guarantee minimal dispersion inside each cluster. The linkage distance is shown along the y-axis, indicating the point where the clusters are merged (the sooner two clusters are merged, the most similar they are). The similarity between the cases belonging to

Table 4. The classes assigned to the real networks by considering each combination of measurements. The classes in bold mean wrong identified model and, in italic style, wrong average vertex degree.

		Identified networks for the following combinations:							
Experimental network	Expected network	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
US Airlines transportation network (USATN) $\langle k\rangle = 12.8$	BA/GN $\langle k\rangle = 12.8$	*BA* $\langle k\rangle = 10.0$	*GN** $\langle k\rangle = 10.0$	GN* $\langle k\rangle = 12.8$	*BA* $\langle k\rangle = 10.0$	*BA** $\langle k\rangle = 10.0$	*BA** $\langle k\rangle = 10.0$	*GN** $\langle k\rangle = 14.0$	BA* $\langle k\rangle = 12.0$
Autonomous System (AS) $\langle k\rangle = 3.59$	BA $\langle k\rangle = 3.59$	*BA* $\langle k\rangle = 6.0$	**GN** $\langle k\rangle = \mathbf{3.59}$	*BA* $\langle k\rangle = 6.0$	BA $\langle k\rangle = 4.0$	**GN** $\langle k\rangle = \mathbf{3.59}$	**GN** $\langle k\rangle = \mathbf{3.59}$	*BA* $\langle k\rangle = 6.0$	**GN** $\langle k\rangle = \mathbf{3.59}$
Transcriptional regulation Network of the *E. coli* (TRNE) $\langle k\rangle = 2.45$	BA $\langle k\rangle = 2.45$	BA $\langle k\rangle = 2.0$	**GN** $\langle k\rangle = \mathbf{2.45}$	***GN*** $\langle k\rangle = \mathbf{4.0}$	*BA* $\langle k\rangle = 4.0$	**ER** $\langle k\rangle = \mathbf{2.45}$	**ER** $\langle k\rangle = \mathbf{2.45}$	**GN** $\langle k\rangle = \mathbf{2.45}$	ER $\langle k\rangle = 2.45$
Protein-Protein interaction Network of the Saccharomyces Cerevisiae (PPIN) $\langle k\rangle = 3.03$	BA $\langle k\rangle = 3.03$	***ER*** $\langle k\rangle = 2.0$	**GN** $\langle k\rangle = 3.03$	***GN*** $\langle k\rangle = \mathbf{2.0}$	***ER*** $\langle k\rangle = \mathbf{2.0}$	**GN** $\langle k\rangle = \mathbf{3.03}$	**GN** $\langle k\rangle = \mathbf{3.03}$	***ER*** $\langle k\rangle = \mathbf{2.0}$	**GN** $\langle k\rangle = \mathbf{3.03}$
Delaunay Network (DLN) $\langle k\rangle = 6.0$	GN $\langle k\rangle = 6.0$	*GN* $\langle k\rangle = 4.0$	*GN* $\langle k\rangle = 4.0$	GN $\langle k\rangle = 6.0$	*GN* $\langle k\rangle = 4.0$	*GN* $\langle k\rangle = 4.0$	GN $\langle k\rangle = 6.0$	GN $\langle k\rangle = 6.0$	GN $\langle k\rangle = 6.0$

* Class identified well away from all considered theoretical models (see, for instance, figure 27(c)).

Table 5. Summary of discussed measurements.

Measurement	Symbol	Equation
Mean geodesic distance	ℓ	(13)
Global efficiency	E	(14)
Harmonic mean distance	h	(15)
Vulnerability	V	(17)
Network clustering coefficient	C and $\tilde{C}$	(18) and (25)
Weighted clustering coefficient	C^w	(27)
Cyclic coefficient	θ	(32)
Maximum degree	$k_{\max}$	(40)
Mean degree of the neighbors	$k_{\text{nn}}(k)$	(42)
Degree-degree correlation coefficient	r	(43)
Assortativity coefficient	$\tilde{\mathbb{Q}}, \mathbb{Q}$	(46) and (47)
Bipartivity degree	b and β	(48) and (49)
Degree Distribution entropy	$H(i)$	(50)
Average search information	$\mathcal{S}$	(55)
Access information	$\mathcal{A}_i$	(56)
Hide information	$\mathcal{H}_i$	(57)
Target entropy	$\mathcal{T}$	(60)
Road entropy	$\mathcal{R}$	(61)
Betweenness centrality	B_i	(65)
Central point dominance	CPD	(66)
lth moment	M_l	(68)
Modularity	Q	(69)
Participation coefficient	P_i	(82)
z-score	z_i	(83)
Significance profile	SP_i	(84)
Subgraph centrality	SC	(91)
Hierarchical clustering coefficient	C_{rs}	(94)
Convergence ratio	$cv_d(i)$	(95)
Divergence ratio	$dv_d(i)$	(96)
Edge reciprocity	ϱ and ρ	(101) and (102)
Matching index of edge (i, j)	μ_{ij}	(104)

each of the three types of networks is reflected by the fact that three respective main branches are obtained in the dendrogram in figure 28. The GN cluster incorporates the experimental protein-protein network, to which it is most closely related by the measurements. Note that the GN group, including the protein-protein network, is significantly different from the ER and BA models at the right-hand side of the figure, as indicated by the high linkage distance at which these two groups (i.e. the GN and ER/BA) are merged.

The results discussed above illustrate the classification procedure and its potential for identifying the category of networks of unknown nature. The fact that the assigned category sometimes varies according to the choice of measurements suggests the presence of specific topological features in some experimental networks which are not fully compatible with any of the assumed theoretical reference models. Indeed, the consideration of a more comprehensive set of measurements can, in principle, provide a more meaningful subclassification of the networks. Such a possibility is particularly important in the case of scale-free networks, which are known to involve

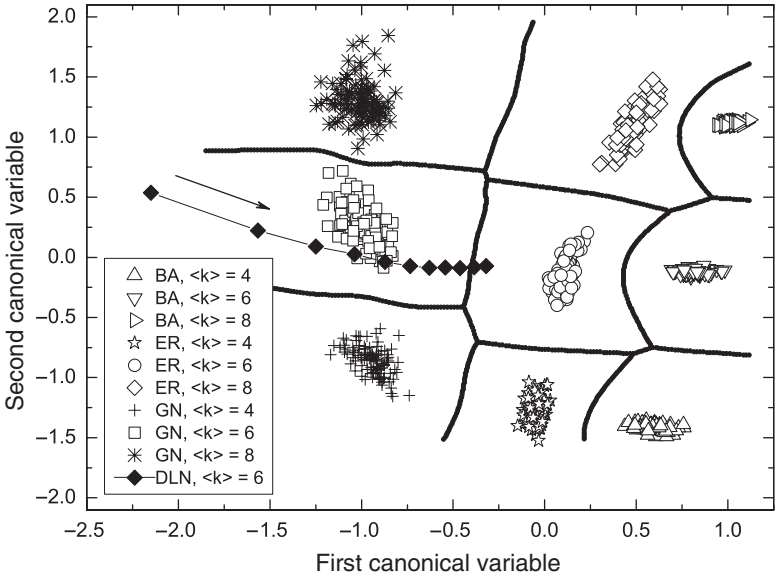


Figure 26. Separating frontiers between the decision regions in the scatterplots obtained by canonical analysis for the DLN. The separating frontiers were obtained by Bayesian decision theory. Note the trajectory defined by the mapping of the progressively rewired versions of the original DLN network, extending from the GN towards the ER region with $\langle k \rangle = 6$.

subtypes [24]. For instance, TRNE has been identified in our experiments as having BA type while considering two measurements (i.e. $\{st, \ell\}$), but was understood as a GN model by considering three measurements (i.e. $\{\langle k \rangle, \tilde{C}, \ell\}$) and as ER when we considered six measurements (i.e. $\{\langle k \rangle, \tilde{C}, \ell, st, r, CPD\}$).

It should be always kept in mind that the consideration of an excessive number of measurements may ultimately compromise the quality of the classification.

Methodologies such as the canonical analysis followed by Bayesian classification can be used to identify the features which contribute particularly to the correct classifications. This can be done by considering the measurements which contribute more intensely to the canonical projections providing the largest number of correct classifications. A simpler methodology involves the application of the principal component analysis to remove the redundancies between the measurements. In the case of a reduced number of measurements, it is also possible to consider all the respective combinations and identify which of them yields the best classifications. Another interesting possibility for investigating complex network connectivity is to consider outliers analysis (e.g. [208]). The reader interested in additional information on multivariate statistics and feature selection is referred to the specialized literature (e.g., [68, 69, 211, 212]) for more in-depth discussion and coverage. Many other methods from multivariate statistical analysis, including hierarchical clustering and structural equation modeling, can also be valuable for investigations in complex network research. Though the potential of hierarchical clustering for suggesting relationships between the classes is briefly illustrated in the following, further information about such methods can be found in textbooks such as [68, 69, 211–213, 220].

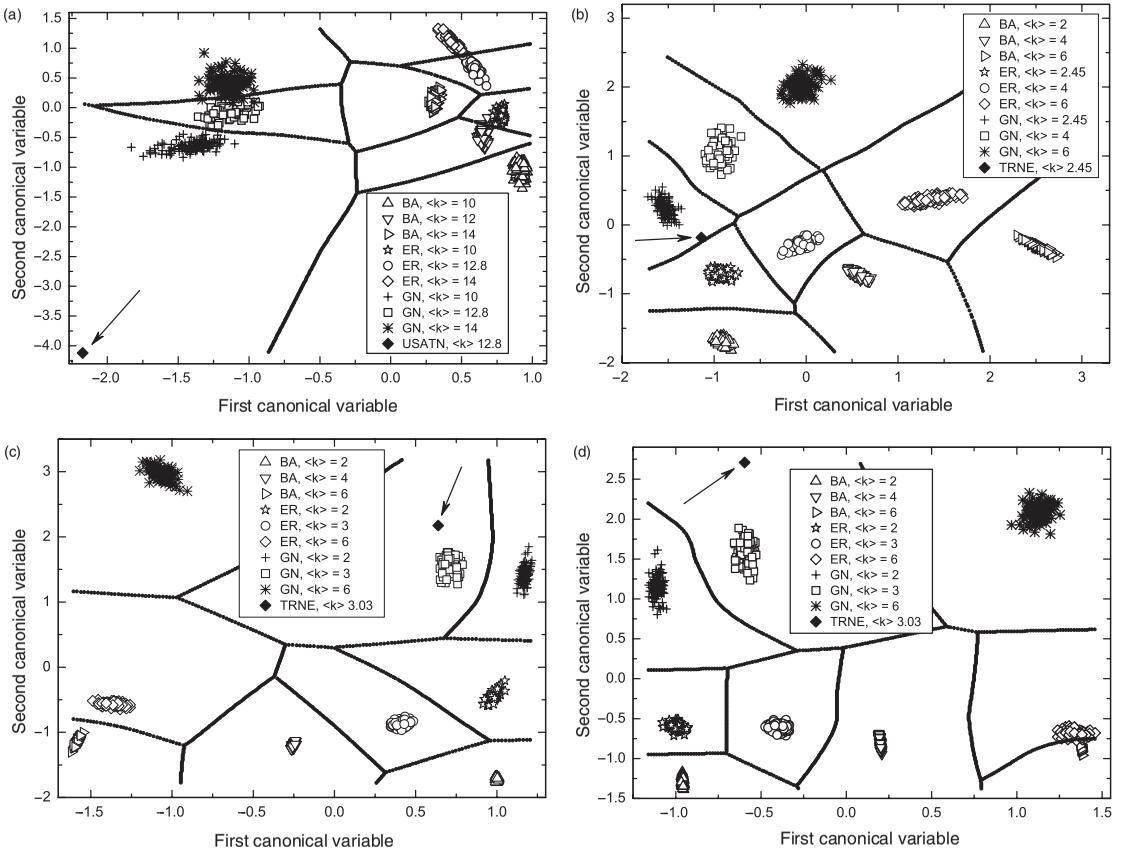


Figure 27. Examples of classification by canonical variable analysis and Bayesian decision theory: (a) US Airlines Transportation Network (USATN); (b) the Transcriptional Regulation Network of the *E. coli* (TRNE); and (c) the Protein-Protein Interaction Network of the *Saccharomyces Cerevisiae* (PPIN), considering all measurements; (d) the same protein network as in (c) but excluding the hierarchical measurements. Note the presence of the separating frontiers between the decision regions in the scatterplots. The arrows indicate the mapped experimental networks.

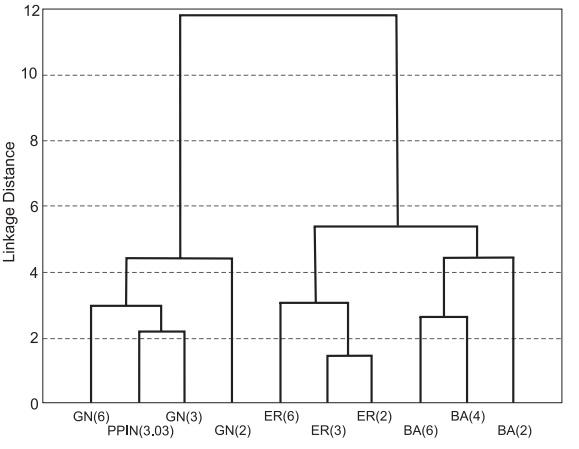


Figure 28. Dendrogram obtained for the protein-protein interaction network considering all measurements except those hierarchical. Note that the BA, ER and GN networks resulted in well-separated branches, while the protein-protein network was included into the latter group.

20. Concluding remarks

Measurements of the connectivity and topology of complex networks are essential for the characterization, analysis, classification, modeling and validation of complex networks. Although initially limited to simple features such as vertex degree, clustering coefficient and shortest path length, several novel, powerful measurements have been proposed. We hope it has been made clear that the several available measurements often provide complementary characterization of distinct connectivity properties of the structures under analysis. It is only by becoming familiar with such measurements that one can expect to identify proper sets of features to be used for the characterization of complex networks. The current survey has been organized to provide a comprehensive coverage of not only the most traditional measurements but also complementary alternatives which, though not so frequently used, can provide valuable resources for characterizing specific topological properties of complex networks. Special attention was also given to the application of measurements in community finding algorithms, an important issue in complex network research.

In addition to presenting such measurements according to coherent categories, we also addressed issues such as visualization, in terms of trajectories defined by measurements, of complex network growth. As illustrated by the results presented, which considered several important theoretical network models, such trajectories clearly reflect, in graphical terms, important tendencies exhibited by different network categories as their average degree is increased. Another important point to be kept in mind in network measurements is correlations. While high correlation between a pair of measurements indicates that they are largely redundant, our results show that the own correlation values vary from one network model to another, providing further useful information for network characterization. Another important property of a specific measurement is its sensitivity to small perturbations in the network, such as the inclusion or removal of edges or vertices. We illustrated that different measure-

ments can behave very differently with respect to such induced changes. Because one of the most challenging issues related to network categorization regards the choice of the features to be taken into account, we provided a self-contained discussion about how multivariate statistics concepts and methods can be applied for that aim. More specifically, we showed how high dimensional measurement spaces can be effectively projected, by using principal component analysis, into lower-dimensional spaces favoring visualization and application of computationally intensive measurements. We also described how two useful methods, namely canonical analysis and Bayesian decision theory, can be combined to provide the means for semi-automated identification of the effective linear combinations of measurements, in the sense of allowing good discrimination between network categories. The potential of such multivariate methodologies was illustrated for theoretical models and experimental networks. The results clearly suggested that considering a comprehensive set of measurements can provide more complete characterization of the topological properties of the networks to the point of requiring a revision of the traditional classification of experimental networks into subclasses or new models.

All in all, this survey provides for the first time, an integrated presentation and discussion of a comprehensive set of measurements previously covered in separate works. In addition, it addresses important issues related to the application of these measurements for characterization and classification of networks, including dynamic representations in terms of trajectories, redundancy between measurements as quantified by correlations, perturbation effects and a powerful multivariate framework for classification of networks of unknown category. The systematic application of such concepts and tools is poised to yield a wealth of new results in the study of complex networks.

Acknowledgements

We are grateful to Lucas Antiqueira, Carlos A.-A. Castillo-Ocaranza, Ernesto Estrada, A. Díaz-Guilera, Shalev Itzkovitz, Marcus Kaiser, Xiang Lee, Jon Machta, Adilson E. Motter, Osvaldo N. Oliveira-Jr, Andrea Scharnhorst, Matheus Viana, and Duncan Watts for comments and suggestions. Luciano da F. Costa is grateful to FAPESP (procs. 99/12765-2 and 05/00587-5), CNPq (proc. 308231/03-1) and the Human Frontier Science Program (RGP39/2002) for financial support. Francisco A. Rodrigues is grateful to FAPESP (proc. 04/00492-1) and Paulino R. Villas Boas is grateful to CNPq (proc. 141390/2004-2).

References

[1] P.J. Flory, J. Amer. Chem. Soc. **63** 3083 (1941).
[2] A. Rapoport, Bull. Math. Biophy. **13** 85 (1951).
[3] A. Rapoport, Bull. Math. Biophy. **15** 523 (1953).
[4] A. Rapoport, Bull. Math. Biophy. **19** 257 (1957).
[5] P. Erdős and A. Rényi, Publ. Math., **6** 290 (1959).
[6] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5** 17 (1960).
[7] P. Erdős and A. Rényi, Acta Math. Sci. Hung. **12** 261 (1961).

[8] D.J. Watts and S.H. Strogatz, Nature **393** 440 (1998).
[9] A.-L. Barabási and R. Albert, Science **286** 509 (1999).
[10] M. Girvan and M.E.J. Newman, Proc. Nat. Acad. Sci. USA **99** 7821 (2002).
[11] B. Bollobás, *Modern Graph Theory. Graduate Texts in Mathematics* (Springer-Verlag, New York, 1998).
[12] D.B. West, *Introduction to Graph Theory* (Prentice Hall, London, 2001).
[13] J.P. Scott, *Social Network Analysis: A Handbook* (Sage Publications, 2000).
[14] M.E.J. Newman and J. Park, Phys. Rev. E **68** 036122 (2003).
[15] A.-L. Barabási and Z.N. Oltvai, Nature **5** 101 (2004).
[16] S. Bornholdt and H.G. Schuste (Eds) *Handbook of Graphs and Networks: From the Genome to the Internet*, (Wiley-VCH, London, 2003).
[17] L.A.N. Amaral and J.M. Ottino, Eur. Phy. J. B, **38** 147 (2004).
[18] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comp. Commun. Rev. **29** 251 (1999).
[19] R. Albert, H. Jeong, and A.-L. Barabási, Nature **401** 130 (1999).
[20] A.-L. Barabási, R. Albert, and H. Jeong, Phys. A **281** 69 (2000).
[21] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74** 48 (2002).
[22] S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51** 1079 (2002).
[23] M.E.J. Newman, SIAM Rev. **45** 167 (2003).
[24] S. Boccaletti, V. Latora, Y. Moreno, M. Chaves, and D.-U. Hwang, Phys. Rep. **424** 175 (2006).
[25] B. Hayes, Amer. Scien. **88** 9 (2000).
[26] B. Hayes, Amer. Scien. **88** 104 (2000).
[27] A.-L. Barabási and E. Bonabeau, Scien. Amer. **288** 60 (2003).
[28] U. Brandes and D. Wagner (Eds) *Graph-Theoretic Concepts in Computer Science*, Lecture Notes in Computer Science, Konstanz, Germany, June 15–17 2000. 26th International Workshop, Springer.
[29] P.L. Garrido and J. Marro (Eds) *Modeling Complex Systems*, volume 661 of *American Institute of Physics Conference Proceedings*, Spain, 2003. Seventh Granada Lectures, Melville: New York.
[30] R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera (Eds) *Statistical Mechanics of Complex Networks*, volume 625 of *Lecture Notes in Physics* (Springer, Berlin, 2003).
[31] M. Boguñá, R. Pastor-Satorras and A. Vespignani (Eds) *Statistical Mechanics of Complex Networks*, volume 625 of *Lectures and Notes in Physics* (Springer, Berlin, 2003).
[32] E. Ben-Naim, H. Frauenfelder and Z. Toroczkai (Eds), *Complex Networks*, Lecture Notes in Physics (Springer Verlag, Berlin, 2004).
[33] D. Bonchev and D.H. Rouvray (Eds), *Complexity in Chemistry, Biology, and Ecology*, Mathematical and Computational Chemistry (Springer, Berlin, 2005).
[34] M.E.J. Newman, A.-L. Barábasi and D.J. Watts (Eds), *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).
[35] B. Bollobás, *Random graphs* (Academic Press, Inc., New York, 1985).
[36] R. Diestel, *Graph Theory* (Springer, Berlin, 2000).
[37] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of Networks – From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[38] S. Strogatz, *Sync: The Emerging Science of Spontaneous Order* (Hyperion, London, 2003).
[39] D.J. Watts, *Small worlds: the dynamics of networks between order and randomness* (Princeton University Press, Princeton, 1999).
[40] B.A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information* (The MIT Press, Massachusetts, 2001).
[41] M. Castells, *The Internet Galaxy* (Oxford University Press, New York, 2001).
[42] A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means* (Plume, 2002).
[43] M. Buchanan, *Nexus: Small Worlds and the Groundbreaking Science of Networks* (Norton, New York, 2002).
[44] D.J. Watts, *Six Degrees. The Science of a Connected Age* (W.W. Norton & Company, 2003).
[45] M. Kochen, *The Small World* (Ablex Publishing Corporation, New Jersey, 1989).

[46] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[47] P. Hage and F. Harary, *Island Networks: Communication, Kinship and Classification Structures in Oceania* (Cambridge University Press, New York, 1996).
[48] W.E. Baker, *Networking Smart: How To Build Relationships for Personal and Organizational Success* (Backinprint.com, 2000).
[49] W.E. Baker, *Achieving Success Through Social Capital: Tapping Hidden Resources in Your Personal and Business Networks* (Jossey-Bass, London, 2000).
[50] R.R. McNeill and W.H. McNeill, *The Human Web: A Bird's-Eye View of World History* (W.W. Norton & Company, New York, 2003).
[51] P.R. Monge and N.S. Contractor, *Theories of Communication Networks* (Oxford Univeristy Press, New York, 2003).
[52] P.J. Carrington, J. Scott and S. Wasserman (Eds), *Models and Methods in Social Network Analysis* (Cambridge University Press, Cambridge, 2005).
[53] L.C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science* (Empirical Press, New York, 2004).
[54] P. Csermely, *Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks* (Springer, Berlin, 2006).
[55] D. Messner, *The Network Society: Economic Development and International Competitiveness as Problems of Social Governance* (Frank Cass Publishers, Portland, 1997).
[56] Ross Dawson, *Living Networks: Leasing your Company, Customers, and Partners in the Hyper-Connected Economy* (Prentice Hall, New Jersey, 2003).
[57] C. Westland, *Financial Dynamics: A System for Valuing Technology Companies* (John Wiley & Sons, London, 2003).
[58] Y. Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale University Press, Yale, 2006).
[59] M. Dodge and R. Kitchin, *Mapping Cyberspace* (Routledge, New York, 2001).
[60] M. Dodge and R. Kitchin, *Atlas of Cyberspace* (Addison-Wesley, Great Britian, 2001).
[61] P. Baldi, P. Frasconi and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms* (John Wiley & Sons, England, 2003).
[62] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004).
[63] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor and Francis, London, 1994).
[64] A. Bunde and S. Havlin, *Fractals in Science* (Springer, Berlin, 1995).
[65] A. Bunde and S. Havlin, *Fractals and Disordered Systems* (Springer, Berlin, 1996).
[66] Y. Bar-Yam, *Dynamics of Complex Systems* (Perseus Books, New York, 1992).
[67] N. Boccara, *Modeling Complex Systems* (Springer-Verlag New York, 2004).
[68] L. da F. Costa and R.M. Cesar Jr, *Shape Analysis and Classification: Theory and Practice* (CRC Press, New York, 2001).
[69] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification* (John Wiley & Sons, Inc., New York, 2001).
[70] E. Ziv, R. Koytcheff, M. Middendorf and C. Wiggins, Phys. Rev. E **71** 016110 (2005).
[71] M. Barthélemy, A. Barrat, R. Pastor-Satorras and A. Vespignani, Physica A **346** 34 (2005).
[72] S. Milgram, Psy. Today **1** 60 (1967).
[73] R. Monasson, Eur. Phys. J. B 12 (1999).
[74] M.E.J. Newman and D.J. Watts, Phys. Rev. Lett. A **263** 341 (1999).
[75] E.A. Bender and E.R. Can, J. Combinat. Theory, Ser. A **24** 296 (1978).
[76] M. Molloy and B. Reed, Rand. Struct. Algor. **6** 161 (1995).
[77] M. Molloy and B. Reed, Prob. Comp. **7** 295 (1998).
[78] M.E.J. Newman (edited by S. Bornholdt and H.G. Schuster), *Handbook of Graphs and Networks: From the Genome to the Internet*, (Wiley-VCH, New York, 2003).
[79] M.E.J. Newman, D.J. Watts and S.H. Strogatz, Proc. Nat. Acad. Sci. USA **99** 2566 (2002).
[80] M.E.J. Newman, S.H. Strogatz and D.J. Watts, Phys. Rev. E **64** 26118 (2001).

[81] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman and U. Alon. Cond. Mat. 0312028 (2003).
[82] W. Aiello, F. Chung, and L. Lu, *Proceedings of the thirty-second annual ACM symposium on Theory of computing* (Portland, Oregon, ACM Press, 2000), pp. 171–180.
[83] F. Chung and L. Lu, Proc. Nat. Acad. Sci. USA **99** 15879 (2002).
[84] R. Cohen and S. Havlin, Phys. Rev. Lett. **90** 58701 (2003).
[85] M.T. Gastner and M.E.J. Newman, The European Physical Journal B, **49** 247 (2006).
[86] R. Albert, I. Albert and G.L. Nakarado, Phys. Rev. E **69** 025103 (2004).
[87] R. Kinney, P. Crucitti, R. Albert and V. Latora, Eur. Phys. J. B **46** 101 (2005).
[88] A. Barrat, M. Barthélemy, R. Pastor-Satorras and A. Vespignani, Proc. Nat. Acad. Sci. USA **101** 3747 (2004).
[89] R. Guimerà, S. Mossa, A. Turtschi and L.A.N. Amaral, Proc. Nat. Acad. Sci. USA **102** 7794 (2005).
[90] Y. Hayashi, Physics 0512011 (2005).
[91] V. Latora and M. Marchiori, Physica A **314** 109 (2002).
[92] O. Sporns, Complexity 8 (2002).
[93] M. Kaiser and C.C. Hilgetag, Phys. Rev. E **69** 036103 (2004).
[94] V. Latora and M. Marchiori, Phys. Rev. Lett. **87** 198701 (2001).
[95] R. Guimerà, A. Díaz-Guilera, F. Vega-Redondo, A. Cabrales and A. Arenas, Phys. Rev. Lett. **89** 248701 (2002).
[96] V. Gol'dshtein, G.A. Koganov and G.I. Surdutovich, Cond. Mat. 0409298 (2004).
[97] V. Latora and M. Marchiori, Phys. Rev. E **71** 015103R (2005).
[98] M.E.J. Newman, Phys. Rev. E **64** 016131 (2001).
[99] J.-P. Onnela, J. Saramäki, J. Kertész and K. Kaski, Phys. Rev. E **71** 065103(R) (2005).
[100] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67** 026112 (2003).
[101] S.N. Soffer and A. Vázquez, Phys. Rev. E **71** 057101 (2005).
[102] H.J. Kim and J.M. Kim, Phys. Rev. E **72** 036109 (2005).
[103] G. Caldarelli, R. Pastor-Satorras and A. Vespignani, Eur. Phys. J. B **38** 183 (2004).
[104] H.D. Rozenfeld, J.E. Kirk, E.M. Bollt and D. ben Avraham, J. Phys. A: Math. Gen. **38** 4589 (2005).
[105] P.G. Lind, M.C. Gonzalez and H.J. Herrmann, Phys. Rev. E **72** 056127 (2005).
[106] K. Klemm and P.F. Stadler, Cond. Mat. 0506493 (2005).
[107] G. Bianconi and A. Capocci, Phys. Rev. Lett. **90** 078701 (2003).
[108] G. Bianconi, G. Caldarelli and A. Capocci, Phys. Rev. E **71** 066116 (2005).
[109] G. Bianconi and M. Marsili, J. Stat. Mech.: Theory Exper. P06005 (2005).
[110] G. Bianconi and M. Marsili, Phys. Rev. E **73** 066127 (2006).
[111] V. Colizza, A. Flammini, M.A. Serrano and A. Vespignani, Nature Phys. **2** 110 (2006).
[112] S. Zhou and R.J. Mondragon, Commun. Lett. IEEE **8** 180 (2004).
[113] S.N. Dorogovtsev and J.F.F. Mendes, Cond. Mat. 0404593 (2004).
[114] S. Maslov and K. Sneppen, Science **296** 910 (2002).
[115] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **66** 047104 (2002).
[116] R. Pastor-Satorras, A. Vazquez and A. Vespignani, Phys. Rev. Lett. **87** 258701 (2001).
[117] M.E.J. Newman, Phys. Rev. Lett. **89** 208701 (2002).
[118] M. Catanzaro, G. Caldarelli and L. Pietronero, Phys. Rev. E **70** 037101 (2004).
[119] J. Park and M.E.J. Newman, Phys. Rev. E **68** 026112 (2003).
[120] J. Berg, M. Lässig and A. Wagner, BMC Evolut. Biol. **4** 51 (2004).
[121] M. Brede and S. Sinha, Cond. Mat. 0507710 (2005).
[122] M. di Bernardo, F. Garofalo and F. Sorrentino, Cond. Mat. 0506236 (2005).
[123] M. di Bernardo, F. Garofalo and F. Sorrentino, Int. J. Bifurc. Chaos, in press (2006).
[124] N. Madar, T. Kalisky, R. Cohen, D. ben Avraham and S. Havlin, Eur. Phys. J. B **38** 269 (2004).
[125] T. Zhou, Z.-Q. Fu and B.-H. Wang, Prog. Nat. Sci. **16** 452 (2006).
[126] S. Gupta, R.M. Anderson and R.M. May, AIDS **03** 807 (1989).
[127] M.E.J. Newman, Phys. Rev. E **67** 026126 (2003).
[128] P. Holme, F. Liljeros, C.R. Edling and B.J. Kim, Phys. Rev. E **68** 056107 (2003).
[129] E. Estrada and J.A. Rodríguez-Veláquez, Phys. Rev. E **72** 046105 (2005).

[130] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, London, 1965).
[131] L. Brillouin, *Science and Information Theory* (Dover Phoenix Editions, 2004).
[132] L.E. Reichl, *A Modern Course in Statistical Physics* (Wiley-Interscience, 1998).
[133] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois, 1963).
[134] B. Wang, H. Tang, C. Guo and Z. Xiu, Cond. Mat. 0506725 (2005).
[135] L. Demetrius and T. Manke, Physica A **346** 682 (2004).
[136] R.V. Solé and S. Valverde, *Lecture Notes in Phyics* (Springer, Berlin, 2004), pp. 169–190.
[137] K. Sneppen, A. Trusina and M. Rosvall, Europhys. Lett. **69** 853 (2005).
[138] M. Rosvall and K. Sneppen, Cond. Mat. 0604036 (2006).
[139] M. Rosvall, A. Trusina, P. Minhagen and K. Sneppen, Phys. Rev. Lett. **94** 028701 (2005).
[140] A. Trusina, M. Rosvall and K. Sneppen, Phys. Rev. Lett. **94** 238701 (2005).
[141] M. Rosvall, A. Grönlund, P. Minnhagen and K. Sneppen, Phys. Rev. E **72** 046117 (2005).
[142] G. Bianconi, Cond. Mat. 0606365 (2006).
[143] L.C. Freeman, Sociometry **40** 35 (1977).
[144] A. Arenas, A. Cabrales, A. Díaz-Guilera, R. Guimerà and F. Vega-Redondo, *Statistical Mechanics of Complex Networks*, volume 625 of *Lecture Notes in Physics* (Springer, Berlin, 2003).
[145] M.E.J. Newman, Soci. Networ. **27** 39 (2005).
[146] D. Koschützki, K.A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl and O. Zlotowski, Lecture Notes in Computer Science, 3418 (2005).
[147] I.J. Farkas, I. Derenyi, A.-L. Barabási and T. Vicsek, Phys. Rev. E **64** 026704 (2001).
[148] K.-I. Goh, B. Kahng and D. Kim, Phys. Rev. E **64** 051903 (2001).
[149] V. Rosato and F. Tiriticco, Eur. Lett. **66** 471 (2004).
[150] M.L. Mehta, *Random Matrices* (Academic Press, London, 1991).
[151] A.J. Seary and W.D. Richards, *Dynamic Social Network Modeling and Analysis* (National Academy Press, 2003), pp. 209–228.
[152] A. Arenas, L. Danon, A. Díaz-Guilera, P.M. Gleiser and R. Guimerà, Eur. Phys. J. B **38** 373 (2004).
[153] P.M. Gleiser and L. Danon, Adv. Complex Syst. **6** (2003).
[154] R. Guimerà and L.A.N. Amaral, Nature **433** 895 (2005).
[155] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Proc. Nat. Acad. Sci. USA **101** 2658 (2004).
[156] J. Reichardt and S. Bornholdt, Cond. Mat. 0603718 (2006).
[157] J. Reichardt and S. Bornholdt, Cond. Mat. 0606220 (2006).
[158] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93** 218701 (2004).
[159] M.E.J. Newman and M. Girvan, Phys. Rev. E **69** 026113 (2004).
[160] G. Schlosser and G.P. Wagner, *Modularity in Development and Evolution* (University of Chicago Press, Chicago, 2004).
[161] E. Ziv, M. Middendorf and C.H. Wiggins, Physical Review E **71** 046117 (2005).
[162] M.E.J. Newman, Eur. Phys. J. B **38** 321 (2004).
[163] L. Danon, J. Duch, A. Arenas and A. Díaz-Guilera, J. Statist. Mech.: Theory Exper. P09008 (2005).
[164] A.J. Seary and W.D. Richards, *Proceedings of the International Conference on Social Networks*, volume 1 (1995).
[165] M.E.J. Newman, Physics 0605087 (2006).
[166] M Fiedler, Czechosl. Math. J. **23** 298 (1973).
[167] A. Pothen, H. Simon and K.P. Liou, SIAM J. Matrix Anal. Appl. **11** 430 (1990).
[168] A. Capocci, V.D.P. Servedio, G. Caldarelli and F. Colaiori, Phys. A **352** 669 (2005).
[169] M.E.J. Newman, Proc. Nat. Acad. Sci. USA **103** 8577 (2006).
[170] J.R. Tyler, D.M. Wilkinson and B.A. Huberman, *Proceedings of the First International Conference on Communities and Technologies* (2003).
[171] L. da F. Costa, Phys. Rev. E **70** 056106 (2004).
[172] M.R. Anderberg, *Cluster analysis for applications* (Academic Press, London, 1973).

[173] A.K. Jain and R.C. Dubes, *Algorithms for clustering data* (Prentice Hall, New York, 1988).
[174] H.C. Romesburg, *Cluster analysis for researchers* (Robert E. Krieger, London, 1990).
[175] J. Hopcroft, O. Khan, B. Kulis and B. Selman. Proc. Nat. Acad. Sci. USA **101** 5249 (2004).
[176] A. Clauset, M.E.J. Newman and C. Moore, Phys. Rev. E **70** 066111 (2004).
[177] L. Danon, A. Díaz-Guilera and A. Arenas, Physics. 0601144 (2006).
[178] J. Duch and A. Arenas, Phys. Rev. E **72** 027104 (2005).
[179] M.E.J. Newman, Phys. Rev. E **69** 026113 (2004).
[180] J.P. Bagrow and E.M. Bollt, Phys. Rev. E **72** 046108 (2005).
[181] A. Clauset, Phys. Rev. E **72** 026132 (2005).
[182] S.N. Dorogovtsev, A.V. Goltsev and J.F.F. Mendes, Phys. Rev. Lett. **96** 40601 (2006).
[183] J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat and A. Vespignani, cs.NI/0504107 (2005).
[184] S.N. Dorogovtsev, J.F.F. Mendes, A.M. Povolotsky and A.N. Samukhin, Phys. Rev. Lett. **95** 195701 (2005).
[185] A.V. Goltsev, S.N. Dorogovtsev and J.F.F. Mendes, Phys. Rev. E **73** 056101 (2006).
[186] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt and E. Shir, cs.NI/0607080 (2006).
[187] S. Wuchty and E. Almaas, Proteomics **5** 444 (2005).
[188] M. Altaf-Ul-Amin, K. Nishikata, T. Koma, T. Miyasato, Y. Shinbo, M. Arifuzzaman, C. Wada, M. Maeda, T. Oshima, H. Mori and S. Kanaya, Genome Inform. **14** 498 (2003).
[189] S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Nature Genet. **31** 64 (2002).
[190] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, Science **298** 824 (2002).
[191] M. Middendorf, E. Ziv and C.H. Wiggins, Proc. Nat. Acad. Sci. **102** 3192 (2005).
[192] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer and U. Alon, Science **303** 1538 (2004).
[193] E. Estrada and J.A. Rodríguez-Velázquez, Phys. Rev. E **71** 056103 (2005).
[194] L. Vincent, Sig. Process. **16** 365 (1989).
[195] E.R. Dougherty and R.A. Lotufo, *Hands-on Morphological Image Processing* (SPIE Press, New York, 2003).
[196] H. Heijmans, P. Nacken, A. Toet and L. Vincent, J. Visual Commun. Image Repres. **3** 24 (1990).
[197] M.P. Viana and L. da F. Costa, Cond. Mat. 0504346 (2005).
[198] L. da F. Costa, Phys. Rev. Lett. **93** 098702 (2004).
[199] L. da F. Costa and L.E.C. da Rocha, Eur. Phys. J. B **50** (2006).
[200] L. da F. Costa and F.N. Silva, J. Stat. Phys. in press (2004).
[201] C. Song, S. Havlin and H.A. Makse, Nature **433** 392 (2005).
[202] B. Machta and J. Machta, Phys. Rev. E **71** 026704 (2005).
[203] B. Codenotti and M. Leoncini, *Introduction to Parallel Processing* (Addison-Wesley, London, 1993).
[204] H. Meyer-Ortmanns, Cond. Mat. 0311109 (2003).
[205] J.C. Claussen, q-bio, MN/0410024 (2004).
[206] D. Garlaschelli and M.I. Loffredo, Phys. Rev. Lett. **93** 268701 (2004).
[207] M. Kaiser and C.C. Hilgetag, Biol. Cybern. **90** 311 (2004).
[208] L. da F. Costa, M. Kaiser and C. Hilgetag, Physics 0607272 (2006).
[209] P.L. Krapivsky and S. Redner, J. Phys. A: Math. Gen. **35** (2002).
[210] A.L. Edwards, *An Introduction to Linear Regression and Correlation* (W.H. Freeman and Co, San Francisco, 1993).
[211] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, London, 2004).
[212] K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, New York, 1990).
[213] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice Hall, New York, 2002).

[214] P.R.R. Prado, F.F. Franco, M.H. Manfrin and L. da F. Costa, *Proceedings of the Third Brazilian Symposium of Mathematical and Computacional Biology I* (E-papers publishing, Rio de Janeiro, 2004) pp. 329–340.
[215] V. Batagelj and A. Mrvar, *Pajek datasets*. University of Ljubljana, Slovenia, http://vlado.fmf.uni-lj.si/pub/networks/data (2006).
[216] R. Guimerà and L.A.N. Amaral, Eur. Phys. J. B **38** 381 (2004).
[217] H. Jeong, S.P. Mason, A.-L. Barabási and Z.N. Oltvai, Nature **411** 41 (2001).
[218] S.H. Yook, H. Jeong and A.-L. Barabási, Proc. Nat. Acad. Sci. USA **99** 13382 (2002).
[219] D. Stoyan, W.S. Kendall and J. Mecke, *Stochastic Geometry and Its Applications* (John Wiley and Sons, London, 1996).
[220] J.F. Hair, R.E. Anderson, R.L. Tatham and W.C. Black, *Multivariate Data Analysis* (Prentice-Hall Int. Inc., New Jersey, 1998).