1

Preliminaries: networks and graphs

In this chapter we introduce the reader to the basic definitions of network and graph theory. We define metrics such as the shortest path length, the clustering coefficient, and the degree distribution, which provide a basic characterization of network systems. The large size of many networks makes statistical analysis the proper tool for a useful mathematical characterization of these systems. We therefore describe the many statistical quantities characterizing the structural and hierarchical ordering of networks including multipoint degree correlation functions, clustering spectrum, and several other local and non-local quantities, hierarchical measures and weighted properties.

This chapter will give the reader a crash course on the basic notions of network analysis which are prerequisites for understanding later chapters of the book. Needless to say the expert reader can freely skip this chapter and use it later as a reference if needed.

1.1 What is a network?

In very general terms a network is any system that admits an abstract mathematical representation as a graph whose nodes (vertices) identify the elements of the system and in which the set of connecting links (edges) represent the presence of a relation or interaction among those elements. Clearly such a high level of abstraction generally applies to a wide array of systems. In this sense, networks provide a theoretical framework that allows a convenient conceptual representation of interrelations in complex systems where the system level characterization implies the mapping of interactions among a large number of individuals.

The study of networks has a long tradition in graph theory, discrete mathematics, sociology, and communication research and has recently infiltrated physics and biology. While each field concerned with networks has introduced, in many cases, its own nomenclature, the rigorous language for the description of networks

is found in mathematical graph theory. On the other hand, the study of very large networks has spurred the definitions of new metrics and statistical observables specifically aimed at the study of large-scale systems. In the following we provide an introduction to the basic notions and notations used in network theory and set the cross-disciplinary language that will be used throughout this book.

1.2 Basic concepts in graph theory

Graph theory – a vast field of mathematics – can be traced back to the pioneering work of Leonhard Euler in solving the Königsberg bridges problem (Euler, 1736). Our intention is to select those notions and notations which will be used throughout the rest of this book. The interested reader can find excellent textbooks on graph theory by Bergé (1976), Chartrand and Lesniak (1986), Bollobás (1985, 1998) and Clark and Holton (1991).

1.2.1 Graphs and subgraphs

An undirected graph G is defined by a pair of sets $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a non-empty countable set of elements, called *vertices* or *nodes*, and $\mathcal{E}$ is a set of *unordered* pairs of different vertices, called *edges* or *links*. Throughout the book we will refer to a vertex by its order i in the set $\mathcal{V}$. The edge (i, j) joins the vertices i and j, which are said to be *adjacent* or *connected*. It is also common to call connected vertices *neighbors* or *nearest neighbors*. The total number of vertices in the graph (the cardinality of the set $\mathcal{V}$) is denoted as N and defines the order of the graph. It is worth remarking that in many biological and physical contexts, N defines the physical size of the network since it identifies the number of distinct elements composing the system. However, in graph theory, the size of the graph is identified by the total number of edges E. Unless specified in the following, we will refer to N as the size of the network.

For a graph of size N, the maximum number of edges is $\binom{N}{2}$. A graph with $E = \binom{N}{2}$, i.e. in which all possible pairs of vertices are joined by edges, is called a *complete N-graph*. Undirected graphs are depicted graphically as a set of dots, representing the vertices, joined by lines between pairs of vertices, representing the corresponding edges.

An interesting class of undirected graph is formed by hierarchical graphs where each edge (known as a child) has exactly one parent (node from which it originates). Such a structure defines a *tree* and if there is a parent node, or *root*, from which the whole structure arises, then it is known as a rooted tree. It is easy to prove that the number of nodes in a tree equals the number of edges plus one, i.e.,

$N = E + 1$ and that the deletion of any edge will break a tree into two disconnected trees.

A directed graph D, or digraph, is defined by a non-empty countable set of vertices $\mathcal{V}$ and a set of *ordered* pairs of different vertices $\mathcal{E}$ that are called directed edges. In a graphical representation, the directed nature of the edges is depicted by means of an arrow, indicating the direction of each edge. The main difference between directed and undirected graphs is represented in Figure 1.1. In an undirected graph the presence of an edge between vertices i and j connects the vertices in both directions. On the other hand, the presence of an edge from i and j in a directed graph does not necessarily imply the presence of the reverse edge between j and i. This fact has important consequences for the connectedness of a directed graph, as will be discussed in more detail in Section 1.2.2.

From a mathematical point of view, it is convenient to define a graph by means of the *adjacency matrix* $\mathbf{X} = \{x_{ij}\}$. This is a $N \times N$ matrix defined such that

$$x_{ij} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{if } (i, j) \notin \mathcal{E} \end{cases}. \tag{1.1}$$

For undirected graphs the adjacency matrix is symmetric, $x_{ij} = x_{ji}$, and therefore contains redundant information. For directed graphs, the adjacency matrix is not
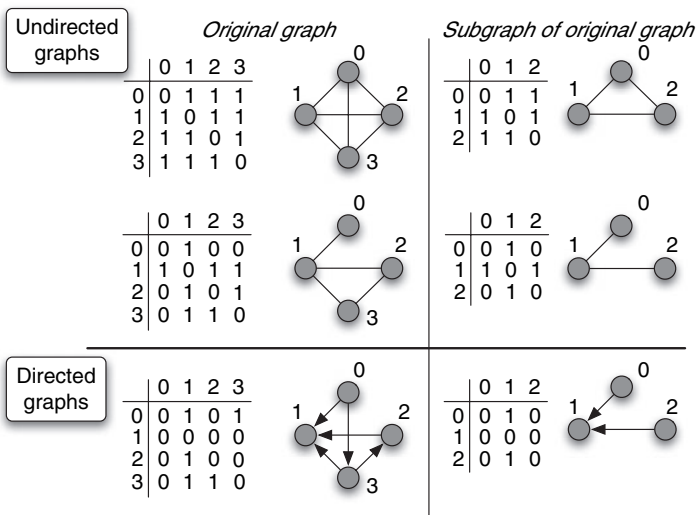

Fig. 1.1. Adjacency matrix and graphical representation of different networks. In the graphical representation of an undirected graph, the dots represent the vertices and pairs of adjacent vertices are connected by a line (edge). In directed graphs, adjacent vertices are connected by arrows, indicating the direction of the corresponding edge.

symmetric. In Figure 1.1 we show the graphical illustrations of different undirected
and directed graphs and their corresponding adjacency matrices.

An important feature of many graphs, which helps in dealing with their structure,
is their *sparseness*. The number of edges E for a connected graph (i.e., with no
disconnected parts) ranges from $N - 1$ to $\binom{N}{2}$. There are different definitions of
sparseness, but we will adopt the convention that when the number of edges scales
as $E \sim N^{\alpha}$ with $\alpha < 2$, the graph is said to be *sparse*. In the case where $E \sim N^2$,
the corresponding graph is called *dense*. By defining the *connectance* or *density* of
a graph as the number of existing edges divided by the maximal possible number
of edges $\mathcal{D} = E/[N(N - 1)/2]$, a graph is then sparse if $\mathcal{D} \ll 1$. This feature
implies, in the case of large graphs, that the adjacency matrix is mostly defined by
zero elements and its complete representation, while costly, does not contain much
relevant information. With large graphs, it is customary to represent the graph in
the compact form defined by the adjacency lists $\ell(i, v \in \mathcal{V}(i))$, where the set of all
neighbors of a fixed vertex i is called the neighborhood (set) of i and is denoted by
$\mathcal{V}(i)$. The manipulation of these lists is obviously very convenient in computational
applications because they efficiently store large sparse networks.

In many cases, we are also interested in subsets of a graph. A graph $G' = (\mathcal{V}', \mathcal{E}')$
is said to be a *subgraph* of the graph $G = (\mathcal{V}, \mathcal{E})$ if all the vertices in $\mathcal{V}'$ belong to $\mathcal{V}$
and all the edges in $\mathcal{E}'$ belong to $\mathcal{E}$, i.e. $\mathcal{E}' \subset \mathcal{E}$ and $\mathcal{V}' \subset \mathcal{V}$. A *clique* is a complete
n-subgraph of size $n < N$. In Figure 1.1 we provide the graphical and adjacency
matrix representations of subgraphs in the undirected and directed cases. The abun-
dance of given types of subgraphs and their properties are extremely relevant in the
characterization of real networks.[1] Small, statistically significant, coherent sub-
graphs, called motifs, that contribute to the set-up of networks have been identified
as relevant building blocks of network architecture and evolution (see Milo *et al.*
[2002] and Chapter 12).

The characterization of local structures is also related to the identification of
communities. Loosely speaking, communities are identified by subgraphs where
nodes are highly interconnected among themselves and poorly connected with
nodes outside the subgraph. In this way, different communities can be traced back
with respect to varying levels of cohesiveness. In directed networks, edge direc-
tionality introduces the possibility of different types of local structures. A possible
mathematical way to account for these local cohesive groups consists in examin-
ing the number of bipartite cliques present in the graph. A bipartite clique $K_{n,m}$
identifies a group of n nodes, all of which have a direct edge to the same m

---

[1] Various approaches exist to determine the structural equivalence, the automorphic equivalence, or the regular
equivalence of subnetworks, and measures for structural similarity comprise correlation coefficients, Euclidean
distances, rates of exact matches, etc.

nodes. The presence of subgraphs and communities raises the issue of modularity in networks. Modularity in a network is determined by the existence of specific subgraphs, called *modules* (or communities). Clustering techniques can be employed to determine major clusters. They comprise non-hierarchical methods (e.g., single pass methods or reallocation methods), hierarchical methods (e.g., single-link, complete-link, average-link, centroid-link, Ward), and linkage based methods (we refer the interested reader to the books of Mirkin (1996) and Banks *et al.* (2004) for detailed expositions of clustering methods). Non-hierarchical and hierarchical clustering methods typically work on attribute value information. For example, the similarity of social actors might be judged based on their hobbies, ages, etc. Non-hierarchical clustering typically starts with information on the number of clusters that a data set is expected to have and sorts the data items into clusters such that an optimality criterion is satisfied. Hierarchical clustering algorithms create a hierarchy of clusters grouping similar data items. Connectivity-based approaches exploit the topological information of a network to identify dense subgraphs. They comprise measures such as betweenness centrality of nodes and edges (Girvan and Newman, 2002; Newman, 2006), superparamagnetic clustering (Blatt, Wiseman and Domany, 1996; Domany, 1999), hubs and bridging edges (Jungnickel, 2004), and others. Recently, a series of sophisticated overlapping and non-overlapping clustering methods has been developed, aiming to uncover the modular structure of real networks (Reichardt and Bornholdt, 2004; Palla *et al.*, 2005).

1.2.2 Paths and connectivity

A central issue in the structure of graphs is the *reachability* of vertices, i.e. the possibility of going from one vertex to another following the connections given by the edges in the network. In a connected network every vertex is reachable from any other vertex. The connected components of a graph thus define many properties of its physical structure.

In order to analyze the connectivity properties let us define a *path* $\mathcal{P}_{i_0,i_n}$ in a graph $G = (\mathcal{V}, \mathcal{E})$ as an ordered collection of $n+1$ vertices $\mathcal{V}_\mathcal{P} = \{i_0, i_1, \ldots, i_n\}$ and n edges $\mathcal{E}_\mathcal{P} = \{(i_0, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n)\}$, such that $i_\alpha \in \mathcal{V}$ and $(i_{\alpha-1}, i_\alpha) \in \mathcal{E}$, for all α. The path $\mathcal{P}_{i_0,i_n}$ is said to connect the vertices i_0 and i_n. The *length* of the path $\mathcal{P}_{i_0,i_n}$ is n. The number $\mathcal{N}_{ij}$ of paths of length n between two nodes i and j is given by the ij element of the nth power of the adjacency matrix: $\mathcal{N}_{ij} = (\mathbf{X}^n)_{ij}$.

A *cycle* – sometimes called a *loop* – is a closed path ($i_0 = i_n$) in which all vertices and all edges are distinct. A graph is called *connected* if there exists a path connecting any two vertices in the graph. A *component* $\mathcal{C}$ of a graph is defined

as a connected subgraph. Two components $\mathcal{C}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{C}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ are disconnected if it is impossible to construct a path $\mathcal{P}_{i,j}$ with $i \in \mathcal{V}_1$ and $j \in \mathcal{V}_2$.

It is clear that for a given number of nodes the number of loops increases with the number of edges. It can easily be shown (Bergé, 1976) that for any graph with p disconnected components, the number of independent loops, or *cyclomatic number*, is given by

$$\Gamma = E - N + p. \tag{1.2}$$

It is easy to check that this relation gives $\Gamma = 0$ for a tree.

A most interesting property of random graphs (Section 3.1) is the distribution of components, and in particular the existence of a *giant component* $\mathcal{G}$, defined as a component whose size scales with the number of vertices of the graph, and therefore diverges in the limit $N \to \infty$. The presence of a giant component implies that a macroscopic fraction of the graph is connected.

The structure of the components of directed graphs is somewhat more complex as the presence of a path from the node i to the node j does not necessarily guarantee the presence of a corresponding path from j to i. Therefore, the definition of a giant component needs to be adapted to this case. In general, the component structure of a directed network can be decomposed into a giant weakly connected component (GWCC), corresponding to the giant component of the same graph in which the edges are considered as undirected, plus a set of smaller disconnected components, as sketched in Figure 1.2. The GWCC is itself composed of several parts because of the directed nature of its edges: (1) the giant strongly connected
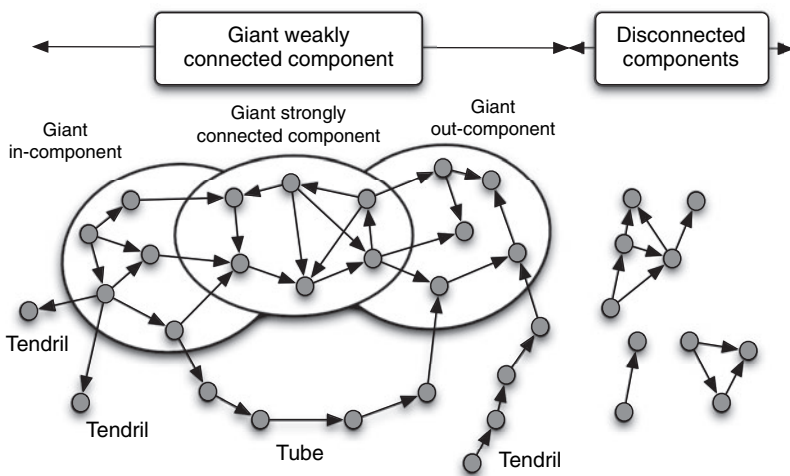

Fig. 1.2. Component structure of a directed graph. Figure adapted from Dorogovtsev *et al.* (2001a).

component (GSCC), in which there is a directed path joining any pair of nodes; (2) the giant in-component (GIN), formed by the nodes from which it is possible to reach the GSCC by means of a directed path; (3) the giant out-component (GOUT), formed by the nodes that can be reached from the GSCC by means of a directed path; and (4) the tendrils containing nodes that cannot reach or be reached by the GSCC (among them, the tubes that connect the GIN and GOUT), which form the rest of the GWCC.

The concept of "path" lies at the basis of the definition of distance among vertices. Indeed, while graphs usually lack a metric, the natural distance measure between two vertices i and j is defined as the number of edges traversed by the shortest connecting path (see Figure 1.3). This distance, equivalent to the chemical distance usually considered in percolation theory (Bunde and Havlin, 1991), is called the *shortest path length* and denoted as ℓ_{ij}. When two vertices belong to two disconnected components of the graph, we define $\ell_{ij} = \infty$. While it is a symmetric quantity for undirected graphs, the shortest path length ℓ_{ij} does not coincide in general with ℓ_{ji} for directed graphs.
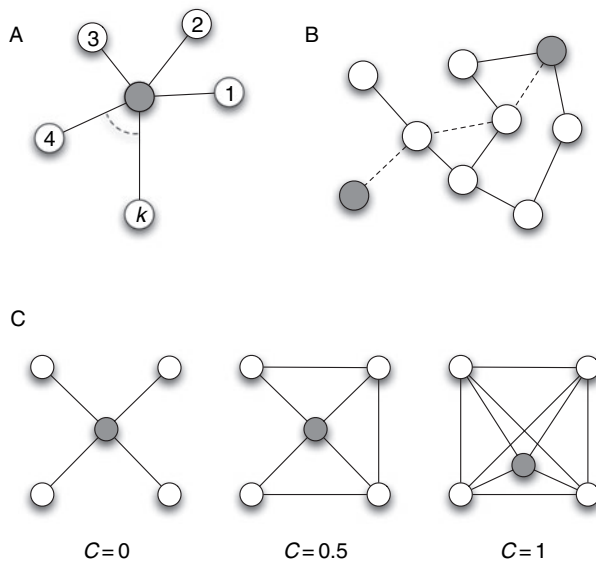

Fig. 1.3. Basic metrics characterizing a vertex i in the network. A, The degree k quantifies the vertex connectivity. B, The shortest path length identifies the minimum connecting path (dashed line) between two different vertices. C, The clustering coefficient provides a measure of the interconnectivity in the vertex's neighborhood. As an example, the central vertex in the figure has a clustering coefficient $C = 1$ if all its neighbors are connected and $C = 0$ if no interconnections are present.

By using the shortest path length as a measure of distance among vertices, it is possible to define the diameter and the typical size of a graph. The *diameter* is traditionally defined as

$$d_G = \max_{i,j} \ell_{ij}. \tag{1.3}$$

Another effective definition of the linear size of the network is the *average shortest path length*,[2] defined as the average value of ℓ_{ij} over all the possible pairs of vertices in the network

$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_{ij} \ell_{ij}. \tag{1.4}$$

By definition $\langle \ell \rangle \leq d_G$, and in the case of a well-behaved and bounded shortest path length distribution, it is possible to show heuristically that in many cases the two definitions behave in the same way with the network size.

There are also other measures of interest which are related to the characterization of the linear size of a graph. The eccentricity of a vertex i is defined by $ec(i) = \max_{j \neq i} \ell_{ij}$, and the radius of a graph G by $\mathrm{rad}_G = \min_i ec(i)$. These different quantities are not independent and one can prove (Clark and Holton, 1991) that the following inequalities hold for any graph

$$\mathrm{rad}_G \leq d_G \leq 2\,\mathrm{rad}_G. \tag{1.5}$$

Simple examples of distances in graphs include the complete graph where $\langle \ell \rangle = 1$ and the regular hypercubic lattice in D dimensions composed by N vertices for which the average shortest path length scales as $\langle \ell \rangle \sim N^{1/D}$. In most random graphs (Sections 2.2 and 3.1), the average shortest path length grows logarithmically with the size N, as $(\langle \ell \rangle \sim \log N)$ – a much slower growth than that found in regular hypercubic lattices. The fact that any pair of nodes is connected by a small shortest path constitutes the so-called *small-world effect*.

1.2.3 Degree and centrality measures

When looking at networks, one of the main insights is provided by the importance of their basic elements (Freeman, 1977). The importance of a node or edge is commonly defined as its centrality and this depends on the characteristics or specific properties we are interested in. Various measurements exist to characterize the centrality of a node in a network. Among those characterizations, the most

---

[2] It is worth stressing that the average shortest path length has also been referred to in the physics literature as another definition for the diameter of the graph.

commonly used are the degree centrality, the closeness centrality, or the betweenness centrality of a vertex. Edges are frequently characterized by their betweenness centrality.

Degree centrality

The degree k_i of a vertex i is defined as the number of edges in the graph incident on the vertex i. While this definition is clear for undirected graphs, it needs some refinement for the case of directed graphs. Thus, we define the *in-degree* $k_{\text{in},i}$ of the vertex i as the number of edges arriving at i, while its *out-degree* $k_{\text{out},i}$ is defined as the number of edges departing from i. The degree of a vertex in a directed graph is defined by the sum of the in-degree and the out-degree, $k_i = k_{\text{in},i} + k_{\text{out},i}$. In terms of the adjacency matrix, we can write

$$k_{\text{in},i} = \sum_j x_{ji}, \qquad k_{\text{out},i} = \sum_j x_{ij}. \tag{1.6}$$

For an undirected graph with a symmetric adjacency matrix, $k_{\text{in},i} = k_{\text{out},i}$. The degree of a vertex has an immediate interpretation in terms of centrality quantifying how well an element is connected to other elements in the graph. The *Bonacich power index* takes into account not only the degree of a node but also the degrees of its neighbors.

Closeness centrality

The *closeness centrality* expresses the average distance of a vertex to all others as

$$g_i = \frac{1}{\sum_{j \neq i} \ell_{ij}}. \tag{1.7}$$

This measure gives a large centrality to nodes which have small shortest path distances to the other nodes.

Betweenness centrality

While the previous measures consider nodes which are topologically better connected to the rest of the network, they overlook vertices which may be crucial for connecting different regions of the network by acting as bridges. In order to account quantitatively for the role of such nodes, the concept of betweenness centrality has been introduced (Freeman, 1977; Newman, 2001a): it is defined as the number of shortest paths between pairs of vertices that pass through a given vertex. More precisely, if σ_{hj} is the total number of shortest paths from h to j and $\sigma_{hj}(i)$ is the number of these shortest paths that pass through the vertex i, the betweenness of i is defined as

$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}}. \tag{1.8}$$

A similar quantity, the load or stress centrality, does not discount the multiplicity of equivalent paths and reads as $L_i = \sum_{h \neq j \neq i} \sigma_{hj}(i)$. The above definitions may include a factor $1/2$ to avoid counting each path twice in undirected networks. The calculation of this measure is computationally very expensive. The basic algorithm for its computation would lead to a complexity of order $\mathcal{O}(N^2 E)$, which is prohibitive for large networks. An efficient algorithm to compute betweenness centrality is reported by Brandes (2001) and reduces the complexity to $\mathcal{O}(NE)$ for unweighted networks.

According to these definitions, central nodes are therefore part of more shortest paths within the network than less important nodes. Moreover, the betweenness centrality of a node is often used in transport networks to provide an estimate of the traffic handled by the vertices, assuming that the number of shortest paths is a zero-th order approximation to the frequency of use of a given node. Analogously to the vertex betweenness, the betweenness centrality of edges can be calculated as the number of shortest paths among all possible vertex couples that pass through the given edge. Edges with the maximum score are assumed to be important for the graph to stay interconnected. These high-scoring edges are the "bridges" that interconnect clusters of nodes. Removing them frequently leads to unconnected clusters of nodes. The "bridges" are particularly important for decreasing the average path length among nodes in a network, for speeding up the diffusion of information, or for increasing the size of the part of the network at a given distance from a node. However, networks with many such bridges are more fragile and less clustered.

1.2.4 Clustering

Along with centrality measures, vertices are characterized by the structure of their local neighborhood. The concept of *clustering*[3] of a graph refers to the tendency observed in many natural networks to form cliques in the neighborhood of any given vertex. In this sense, clustering implies the property that, if the vertex i is connected to the vertex j, and at the same time j is connected to l, then with a high probability i is also connected to l. The clustering of an undirected graph can be quantitatively measured by means of the *clustering coefficient* which measures the local group cohesiveness (Watts and Strogatz, 1998). Given a vertex i, the clustering $C(i)$ of a node i is defined as the ratio of the number of links between the neighbors of i and the maximum number of such links. If the degree of node i is k_i and if these nodes have e_i edges between them, we have

$$C(i) = \frac{e_i}{k_i(k_i - 1)/2}, \qquad (1.9)$$

[3] Also called *transitivity* in the context of sociology (Wasserman and Faust, 1994).