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

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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

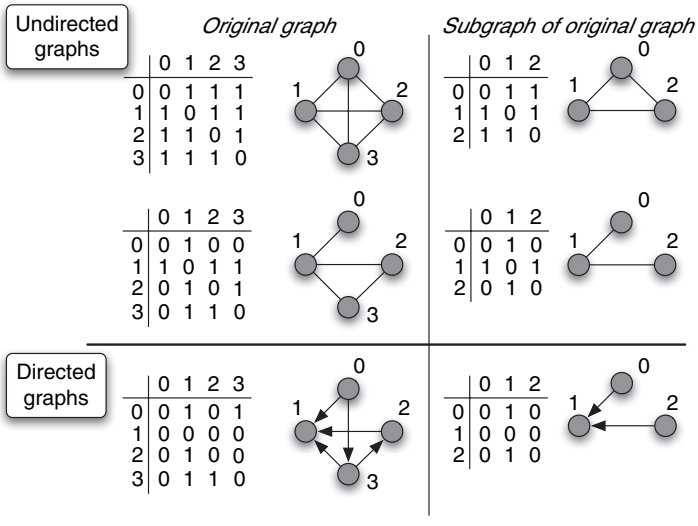

Fig. 1.1. Adjacency matrix and graphical representation of different networks. In the graphical representation of an undirected graph, the dots represent the vertices and pairs of adjacent vertices are connected by a line (edge). In directed graphs, adjacent vertices are connected by arrows, indicating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

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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 abundance of given types of subgraphs and their properties are extremely relevant in the characterization of real networks.[1] Small, statistically significant, coherent subgraphs, 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 directionality introduces the possibility of different types of local structures. A possible mathematical way to account for these local cohesive groups consists in examining 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

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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

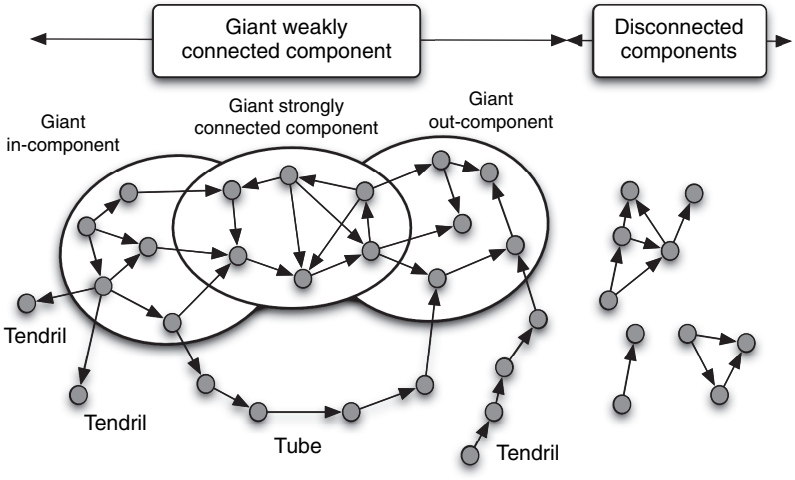

Fig. 1.2. Component structure of a directed graph. Figure adapted from Dorogovtsev *et al.* (2001a).

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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.

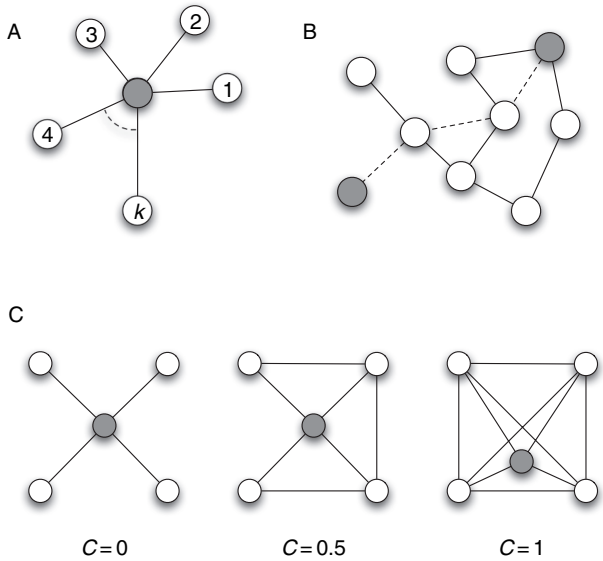

Fig. 1.3. Basic metrics characterizing a vertex i in the network. A, The degree k quantifies the vertex connectivity. B, The shortest path length identifies the minimum connecting path (dashed line) between two different vertices. C, The clustering coefficient provides a measure of the interconnectivity 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.

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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 $\text{rad}_G = \min_i ec(i)$. These different quantities are not independent and one can prove (Clark and Holton, 1991) that the following inequalities hold for any graph

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

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.

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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

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

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

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

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

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

where it is worth remarking that this measure of clustering only has a meaning for $k_i > 1$. For $k_i \leq 1$ we define $C(i) \equiv 0$. Given the definition of e_i, it is easy to check that the number of edges among the neighbors of i can be computed in terms of the adjacency matrix $\mathbf{X}$ as

$$e_i = \frac{1}{2} \sum_{jl} x_{ij} x_{jl} x_{li}. \tag{1.10}$$

In Figure 1.3, we provide an illustration of some simple examples of the clustering of vertices with a given neighborhood. The average clustering coefficient of a graph is simply given by

$$\langle C \rangle = \frac{1}{N} \sum_{i} C(i). \tag{1.11}$$

It is worth noting that the clustering coefficient has been defined in a number of similar ways, for instance as a function of triples in the network (triples are defined as subgraphs which contain exactly three nodes) and reversing the order of average and division in Equation (1.11)

$$C_{\Delta} = \frac{3 \times \text{number of fully connected triples}}{\text{number of triples}}, \tag{1.12}$$

where the factor 3 is due to the fact that each triangle is associated with three nodes. This definition corresponds to the concept of the fraction of transitive triples introduced in sociology (Wasserman and Faust, 1994). Different definitions give rise to different values of the clustering coefficient for a given graph. Hence, the comparison of clustering coefficients among different graphs must use the very same measure. In any case, both measures are normalized and bounded to be between 0 and 1.

1.3 Statistical characterization of networks

One of the elements that has fostered the recent development of network science can be found in the recent possibility of systematic gathering and handling of data sets on several large-scale networks. Indeed, in large systems, asymptotic regularities cannot be detected by looking at local elements or properties. In other words, one has to shift the attention to statistical measures that take into account the global behavior of these quantities.

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

1.3.1 Degree distribution

The degree distribution $P(k)$ of undirected graphs is defined as the probability that any randomly chosen vertex has degree k. It is obtained by constructing the normalized histogram of the degree of the nodes in a network. In the case of directed graphs, one has to consider instead two distributions, the in-degree $P(k_{\text{in}})$ and out-degree $P(k_{\text{out}})$ distributions, defined as the probability that a randomly chosen vertex has in-degree k_{in} and out-degree k_{out}, respectively. The average degree of an undirected graph is defined as the average value of k over all the vertices in the network,

$$\langle k \rangle = \frac{1}{N} \sum_i k_i = \sum_k k P(k) \equiv \frac{2E}{N}, \tag{1.13}$$

since each edge end contributes to the degree of a vertex. For a directed graph, the average in-degree and out-degree must be equal,

$$\langle k_{\text{in}} \rangle = \sum_{k_{\text{in}}} k_{\text{in}} P(k_{\text{in}}) = \langle k_{\text{out}} \rangle = \sum_{k_{\text{out}}} k_{\text{out}} P(k_{\text{out}}) \equiv \frac{\langle k \rangle}{2}, \tag{1.14}$$

since an edge departing from any vertex must arrive at another vertex. Analogously to the average degree, it is possible to define the nth moment of the degree distribution,

$$\langle k^n \rangle = \sum_k k^n P(k). \tag{1.15}$$

A *sparse* graph has an average degree that is much smaller than the size of the graph, $\langle k \rangle \ll N$. In the next chapters we will see that the properties of the degree distribution will be crucial to identify different classes of networks.

1.3.2 Betweenness distribution

Analogously to the degree, it is possible to introduce the probability distribution $P_b(b)$ that a vertex has betweenness b, and the average betweenness $\langle b \rangle$ defined as

$$\langle b \rangle = \sum_b b \; P_b(b) \equiv \frac{1}{N} \sum_i b_i. \tag{1.16}$$

For this quantity it is worth showing its relation with the average shortest path length $\langle \ell \rangle$. By simply reordering the sums in the betweenness definition we have

$$\sum_i b_i = \sum_i \sum_{h,j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}} = \sum_{h \neq j} \frac{1}{\sigma_{hj}} \sum_{i \neq h,j} \sigma_{hj}(i). \tag{1.17}$$

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

A simple topological reasoning gives $\sum_{i\neq h,j}\sigma_{hj}(i) = \sigma_{hj}(\ell_{hj} - 1)$. Plugged into the previous equation, this yields $\sum_i b_i = N(N-1)(\langle\ell\rangle - 1)$, and therefore

$$\langle b\rangle = (N-1)(\langle\ell\rangle - 1). \tag{1.18}$$

As this formula shows, it is easy to realize that the betweenness usually takes values of the order $\mathcal{O}(N)$ or larger. For instance, in a star graph, formed by $N-1$ vertices with a single edge connected to a central vertex, the betweenness takes a maximum value $(N-1)(N-2)$ at the central vertex (the other peripheral vertices having 0 betweenness). For this reason, in the case of very large graphs ($N \to \infty$) it is sometimes convenient to define a rescaled betweenness $\tilde{b} \equiv N^{-1}b$.

Finally, it is clear that in most cases the larger the degree k_i of a node, the larger its betweenness centrality b_i will be. More quantitatively, it has been observed that for large networks and for large degrees there is a direct association of the form (Goh, Kahng and Kim, 2001, 2003; Barthélemy, 2004)

$$b_i \sim k_i^{\eta}, \tag{1.19}$$

where η is a positive exponent depending on the network. Such associations are extremely relevant as they correspond to the fact that a large number of shortest paths go through the nodes with large degree (the hubs). These nodes will therefore be visited and discovered easily in processes of network exploration (see Chapter 8). They will also typically see high traffic which may result in congestion (see Chapter 11). Of course, fluctuations are also observed in real and model networks, and small-degree nodes may also have large values of the betweenness centrality if they connect different regions of the network, acting as bridges. Such low-degree nodes with high centrality appear, for instance, in networks where spatial constraints limit the ability of hubs to deploy long links (Guimerà and Amaral, 2004; Barrat, Barthélemy and Vespignani, 2005).

1.3.3 Mixing patterns and degree correlations

As a discriminator of structural ordering of large-scale networks, the attention of the research community has initially been focused on the degree distribution, but it is clear that this function is only one of the many statistics characterizing the structural and hierarchical ordering of a network. In particular, it is likely that nodes do not connect to each other irrespective of their property or type. On the contrary, in many cases it is possible to collect empirical evidence of specific mixing patterns in networks. A typical pattern known in ecology, epidemiology, and social science as "assortative mixing" refers to the tendency of nodes to connect to other nodes with similar properties. This is common to observe in the social context where people prefer to associate with others who share their interests. Interesting observations about

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

assortative mixing by language or race are abundant in the literature. Likewise, it is possible to define a "disassortative mixing" pattern whenever the elements of the network prefer to share edges with those who have a different property or attribute. Mixing patterns have a profound effect on the topological properties of a network as they affect the community formation and the detailed structural arrangements of the connections among nodes.

While mixing patterns can be defined with respect to any type or property of the nodes (Newman, 2003a), in the case of large-scale networks the research community's interest has focused on the mixing by vertex degree. This type of mixing refers to the likelihood that nodes with a given degree connect with nodes with similar degree, and is investigated through the detailed study of multipoint degree correlation functions. Most real networks do exhibit the presence of non-trivial correlations in their degree connectivity patterns. Empirical measurements provide evidence that high or low degree vertices of the network tend, in many cases, to preferentially connect to other vertices with similar degree. In this situation, correlations are named assortative. In contrast, connections in many technological and biological networks are more likely to attach vertices of very different degree. Correlations are then referred to as disassortative. The correlations, although other possibilities could be considered, are characterized through the conditional probabilities $P(k', k'', \ldots, k'^{(n)} \mid k)$ that a vertex of degree k is simultaneously connected to a number n of other vertices with corresponding degrees $k', k'', \ldots, k'^{(n)}$. Such quantities might be the simplest theoretical functions that encode degree correlation information from a local perspective. A network is said to be uncorrelated when the conditional probability is structureless, in which case the only relevant function is just the degree distribution $P(k)$.

In order to characterize correlations, a more compact quantity is given by the Pearson assortativity coefficient r (Newman, 2002a)

$$r = \frac{\sum_e j_e k_e / E - \left[\sum_e (j_e + k_e)/(2E)\right]^2}{\left[\sum_e (j_e^2 + k_e^2)/(2E)\right] - \left[\sum_e (j_e + k_e)/(2E)\right]^2}, \tag{1.20}$$

where j_e and k_e denote the degree of the extremities of edge e and E is the total number of edges. This quantity varies from -1 (disassortative network) up to 1 (perfectly assortative network). However, such a measure can be misleading when a complicated behavior of the correlation functions (non-monotonous behavior) is observed. In this case the Pearson coefficient gives a larger weight to the more abundant degree classes, which in many cases might not express the variations of the correlation function behavior.

More details on the degree correlations are provided by the two-point conditional probability $P(k' \mid k)$ that any edge emitted by a vertex with degree k is connected to a vertex with degree k'. Even in this simple case, however, the direct evaluation

of the function from empirical data is a rather cumbersome task. In general, two nodes' degree correlations can be represented as the three-dimensional histograms of $P(k' \mid k)$ or related quantities[4] (Maslov and Sneppen, 2002). On the other hand, such a histogram is highly affected by statistical fluctuations and, thus, it is not a good candidate when the data set is not extremely large and accurate. A more practical quantity in the study of the network structure is given by the average nearest neighbors degree of a vertex i

$$k_{\mathrm{nn},i} = \frac{1}{k_i} \sum_{j \in \mathcal{V}(i)} k_j, \tag{1.21}$$

where the sum is over the nearest neighbors vertices of i. From this quantity a convenient measure to investigate the behavior of the degree correlation function is obtained by the average degree of the nearest neighbors, $k_{\mathrm{nn}}(k)$, for vertices of degree k (Pastor-Satorras, Vázquez and Vespignani, 2001; Vázquez, Pastor-Satorras and Vespignani, 2002)

$$k_{\mathrm{nn}}(k) = \frac{1}{N_k} \sum_{i/k_i = k} k_{\mathrm{nn},i}, \tag{1.22}$$

where N_k is the number of nodes of degree k. This last quantity is related to the correlations between the degrees of connected vertices since on average it can be expressed as

$$k_{\mathrm{nn}}(k) = \sum_{k'} k' P(k'|k). \tag{1.23}$$

If degrees of neighboring vertices are uncorrelated, $P(k'|k)$ is only a function of k' and thus $k_{\mathrm{nn}}(k)$ is a constant. In the presence of correlations, the behavior of $k_{\mathrm{nn}}(k)$ identifies two general classes of networks (see Figure 1.4). If $k_{\mathrm{nn}}(k)$ is an increasing function of k, vertices with high degree have a larger probability of being connected

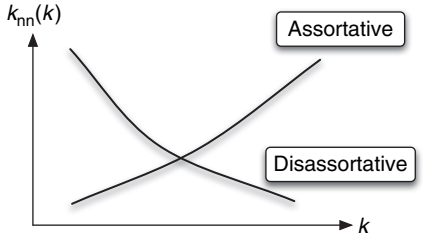

Fig. 1.4. Pictorial representation of the assortative (disassortative) mixing property of networks as indicated by the behavior of the average degree of the nearest neighbors, $k_{\mathrm{nn}}(k)$, for vertices of degree k.

[4] For instance, the joint degree distribution $P(k', k)$ that defines the probability that a randomly chosen edge connects two vertices of degrees k and k'.

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

with large degree vertices. This corresponds to an assortative mixing (Newman, 2002a). On the contrary, a decreasing behavior of $k_{\mathrm{nn}}(k)$ defines a disassortative mixing, in the sense that high degree vertices have a majority of neighbors with low degree, while the opposite holds for low degree vertices (Newman, 2002a).

It is important to stress, however, that given a certain degree distribution, a completely degree–degree uncorrelated network with finite size is not always realizable owing to structural constraints. Indeed, any finite-size random network presents a structural cut-off value k_{c} over which the requirement of the lack of dangling edges introduces the presence of multiple and self-connections and/or degree–degree correlations (Boguñá, Pastor-Satorras and Vespignani, 2004; Moreira, Andrade and Amaral, 2002). Networks with bounded degree distributions and finite second moments $\langle k^2 \rangle$ present a maximal degree $k_{\max}$ that is below the structural one k_{c}. However, in networks with heavy-tailed degree distribution (see Section 2.1), this is not always the case and k_{c} is generally smaller than $k_{\max}$. In this instance, structural degree–degree correlations and higher order effects, such as the emergence of large cliques (Bianconi and Marsili, 2006a; Bianconi and Marsili, 2006b), are present even in maximally random networks.[5] Structural correlations are genuine correlations, which is not surprising since they are just imposed by topological constraints and not by a special ordering or evolutionary principle shaping the network. A more detailed discussion on topological constraints and the properties of random networks can be found in Appendix 1.

In the case of random uncorrelated networks, it is possible to obtain an explicit form for the conditional probability $P(k' \mid k)$. In this case $P(k' \mid k)$ does not depend on k and its functional form in terms of k' can be easily obtained by calculating the probability that any given edge is pointing to a node of degree k'. The probability that one edge is wired to one node of degree k' is just the total number of edges departing from nodes of degree k' divided by the number of all edges departing from nodes of any degree. Since each one of the $N_{k'}$ nodes emanates k' edges, we obtain

$$P_{\mathrm{unc}}(k' \mid k) = \frac{k' N_{k'}}{\sum_{k''} k'' N_{k''}}, \tag{1.24}$$

where considering that $P(k) = N_k / N$, finally yields

$$P_{\mathrm{unc}}(k' \mid k) = \frac{1}{\langle k \rangle} k' P(k'). \tag{1.25}$$

This expression states that even in an uncorrelated network, the probability that any edge points to a node of a given degree k' is not uniform but proportional to

[5] Operatively, the maximally random network can be thought of as the stationary ensemble of networks visited by a process that, at any time step, randomly selects a couple of links of the original network and exchanges two of their ending points (automatically preserving the degree distribution).

the degree itself. In other words, by following any edge at random it is more likely you will end up in a node with large degree: the more connected you are, the easier it is to find you. This result is particularly relevant in the case of networks where degree values may be very different, and will prove to be particularly useful in several analytical calculations where the heterogeneous nature of the network is affecting equilibrium and non-equilibrium processes even in the uncorrelated case. The behavior of $k_{\mathrm{nn}}(k)$ in the uncorrelated case can be easily derived from Equation (1.25) and reads as

$$k_{\mathrm{nn}}^{\mathrm{unc}}(k) = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{1.26}$$

As we stated previously, in a random uncorrelated network the average nearest neighbor degree does not depend on k and has a constant value determined by the two first moments of the degree distribution.

1.3.4 Clustering spectrum

Correlations among three vertices can be measured by means of the probability $P(k', k'' \mid k)$ that a vertex of degree k is simultaneously connected to two vertices with degrees k' and k''. As previously indicated, the conditional probabilities $P(k', k'' \mid k)$ or $P(k'' \mid k)$ are difficult to estimate directly from real data, so other assessments have been proposed. All of them are based on the concept of clustering coefficient $C(i)$ as expressed in Equation (1.9), which refers to the tendency to form triangles in the network (see Section 1.2.4). A measure customarily used in graph characterization is the average clustering coefficient $\langle C \rangle = N^{-1} \sum_i C(i)$ which expresses the statistical level of cohesiveness measuring the global density of interconnected vertex triplets in the network. Although statistical scalar measures are helpful as a first indication of clustering, it is always more informative to work with quantities that explicitly depend on the degree. As in the case of two vertices' correlations, a uniparametric function is defined by the average clustering coefficient $C(k)$ restricted to classes of vertices with degree k (Vázquez *et al.*, 2002; Ravasz *et al.*, 2002)

$$C(k) = \frac{1}{N_k} \sum_{i\,/\,k_i = k} C(i) \tag{1.27}$$

where N_k is the number of vertices with degree k. A functional dependence of the local clustering on the degree can be attributed to the presence of a complex structure in the three-vertex correlation pattern. Indeed, it has been observed that $C(k)$ exhibits, in many cases, a non-trivial dependence on k that is supposed to partly

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

encode the hierarchical structure of the network (Vázquez *et al.*, 2002; Ravasz *et al.*, 2002).

1.3.5 Rich-club phenomenon

Several other statistical measures have been defined in the case of large-scale networks as simple proxies for their architectures and many of them are specifically devised for certain types of graphs, as seen in the density of bipartite cliques $K_{n,m}$ in directed graphs. Analogously, the "rich-club" phenomenon has been discussed in both social and computer sciences (de Solla Price, 1986; Wasserman and Faust, 1994; Zhou and Mondragon, 2004; Pastor-Satorras and Vespignani, 2004), and refers to the tendency of high degree nodes, the hubs of the network (the *rich nodes*), to be very well connected to each other, forming well-interconnected subgraphs (*clubs*) more easily than low degree nodes. Zhou and Mondragon (2004) have introduced a quantitative measure of this tendency through the rich-club coefficient, expressed as

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}, \tag{1.28}$$

where $E_{>k}$ is the number of edges among the $N_{>k}$ nodes with degree larger than k, and $N_{>k}(N_{>k}-1)/2$ represents the maximum possible number of edges among these $N_{>k}$ nodes, so that $\phi(k)$ measures the fraction of edges actually connecting those nodes out of the maximum number of edges they might possibly share. A growing behavior as a function of k indicates that high degree nodes tend to be increasingly connected among themselves.[6] On the other hand, a monotonic increase of $\phi(k)$ does not necessarily imply the presence of a rich-club organizing principle. Indeed, even in the case of the Erdős–Rényi graph – a completely random network – there is an increasing rich-club coefficient. This implies that the increase of $\phi(k)$ is a natural consequence of the fact that vertices with large degree have a larger probability of sharing edges than low degree vertices. This feature is therefore imposed by construction and does not represent a signature of any particular organizing principle or structure. For this reason, it is appropriate to measure the rich-club phenomenon as $\rho_{\text{ran}}(k) = \phi(k)/\phi_{\text{ran}}(k)$, where $\phi_{\text{ran}}(k)$ is the rich-club coefficient of the maximally random network with the same degree distribution $P(k)$ as the network under study (Colizza *et al.*, 2006b). In this case an actual rich-club ordering is denoted by a ratio $\rho_{\text{ran}}(k) > 1$. In other words, $\rho_{\text{ran}}(k)$ is a

[6] It is also worth stressing that the rich-club phenomenon is not trivially related to the mixing properties of networks described in Section 1.3.3, which permit the distinction between assortative networks, where large degree nodes preferentially attach to large degree nodes, and disassortative networks, showing the opposite tendency (Colizza *et al.*, 2006b).

normalized measure which discounts the structural correlations due to unavoidable connectivity constraints, providing a better discrimination of the actual presence of the rich-club phenomenon due to the ordering principles shaping the network. This example makes explicit the need to consider the appropriate null hypotheses when measuring correlations or statistical properties. Such properties might either be simply inherent to the connectivity constraints present in networks, or be the signature of real ordering principles and structural properties due to other reasons.

1.4 Weighted networks

Along with a complex topological structure, real networks display a large heterogeneity in the capacity and intensity of their connections: the weights of the edges. While the topological properties of a graph are encoded in the adjacency matrix x_{ij}, weighted networks are similarly described by a matrix w_{ij} specifying the weight on the edge connecting the vertices i and j ($w_{ij} = 0$ if the nodes i and j are not connected). The weight w_{ij} may assume any value and usually represents a physical property of the edge: capacity, bandwidth, traffic. A very significant measure of a network's properties in terms of the actual weights is also obtained by looking at the vertex *strength* s_i defined as (Yook *et al.*, 2001; Barrat *et al.* 2004a)

$$s_i = \sum_{j \in \mathcal{V}(i)} w_{ij}, \tag{1.29}$$

where the sum runs over the set $\mathcal{V}(i)$ of neighbors of i. The strength of a node integrates the information about its degree and the importance of the weights of its links and can be considered as the natural generalization of the degree. When the weights are independent of the topology, the strength typically grows linearly with the degree, i.e. with the number of terms in the sum (1.29): $s \simeq \langle w \rangle k$ where $\langle w \rangle$ is the average weight. In the presence of correlations we obtain in general $s \simeq Ak^{\beta}$ with $\beta = 1$ and $A \neq \langle w \rangle$, or $\beta > 1$. Statistical measures for weighted networks are readily provided by the probability distributions $P(w)$ and $P(s)$ that any given edge and node have weight w and strength s, respectively.

In general, topological measures do not take into account that some edges are more important than others. This can easily be understood with the simple example of a network in which the weights of all edges forming triples of interconnected vertices are extremely small. Even for a large clustering coefficient, it is clear that these triples have a minor role in the network's dynamics and organization, and the network's clustering properties are definitely overestimated by a simple topological analysis. Similarly, high degree vertices could be connected to a majority of low degree vertices while concentrating the largest fraction of their strength only on the vertices with high degree. In this case the topological information

would point to disassortative properties while the network could be considered as effectively assortative, since the more relevant edges in terms of weight are linking high degree vertices. In order to solve these incongruities, it is possible to introduce specific definitions that explicitly consider the weights of the links and combine the topological information with the weight distribution of the network. Many different clustering coefficient definitions have been introduced in the literature (Barrat *et al.,* 2004; Onnela *et al.,* 2005; Serrano, Boguñá and Pastor-Satorras, 2006; Saramäki *et al.,* 2007). A convenient *weighted clustering coefficient* is defined as

$$C^{\mathrm{w}}(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} x_{ij} x_{ih} x_{jh}. \tag{1.30}$$

The quantity $C^{\mathrm{w}}(i)$ is a count of the weight of the two participating edges of the vertex i for each triple formed in the neighborhood of i. This definition not only considers the number of closed triangles in the neighborhood of a vertex but also considers their total relative edge weights with respect to the vertex's strength. The factor $s_i(k_i - 1)$ is a normalization factor that ensures that $0 \leq C^{\mathrm{w}}(i) \leq 1$. Consistently, the $C^{\mathrm{w}}(i)$ definition recovers the topological clustering coefficient in the case that $w_{ij} = \text{const}$. It is customary to define C^{w} and $C^{\mathrm{w}}(k)$ as the weighted clustering coefficient averaged over all vertices of the network and over all vertices with degree k, respectively. For a large randomized network (without any correlations between weights and topology), it is easy to see that $C^{\mathrm{w}} = C$ and $C^{\mathrm{w}}(k) = C(k)$. In real weighted networks, however, we can face two opposite cases. If $C^{\mathrm{w}} > C$, we observe a network in which the interconnected triples are more likely formed by edges with larger weights (see Figure 1.5). In contrast, $C^{\mathrm{w}} < C$ signals a network in which the topological clustering is generated by edges with low weight. In the latter, it is explicit that the clustering has a minor effect in the organization of the network since the largest part of the interactions (traffic, frequency of the relations, etc. . .) occurs on edges not belonging to interconnected triples. In order to obtain a more detailed knowledge of the structure of the network, the variations of $C^{w}(k)$ with respect to the degree class k can be analyzed and compared with those of $C(k)$.

Similarly, the *weighted average nearest neighbors degree* is defined as (Barrat *et al*., 2004a)

$$k^{\mathrm{w}}_{\mathrm{nn},i} = \frac{1}{s_i} \sum_{j=1}^{N} x_{ij} w_{ij} k_j. \tag{1.31}$$

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

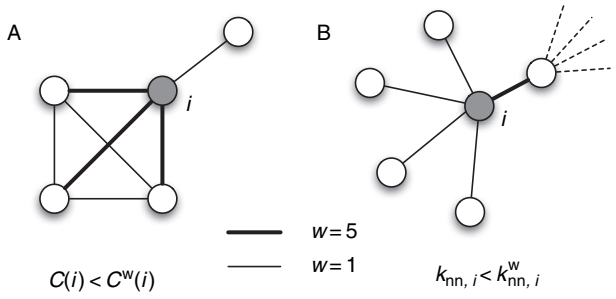

Fig. 1.5. A, The weighted clustering of a node i is larger than its topological counterpart if the weights of the links (i, j) are concentrated on the cliques in which i participates. B, Example of a node i with small average nearest neighbors degree but large *weighted* average nearest neighbors degree: i is mostly connected to low-degree nodes but the link with largest weight points towards a well-connected hub.

This quantity (see also Serrano *et al.*, 2006) is a simple generalization of the average nearest neighbors degree (Pastor-Satorras *et al.*, 2001)

$$k_{\mathrm{nn},i} = \frac{1}{k_i} \sum_{j=1}^{N} x_{ij} k_j \tag{1.32}$$

and performs a local weighted average of the nearest neighbor degree according to the normalized weight of the connecting edges, w_{ij}/s_i. This definition implies that $k^{\mathrm{w}}_{\mathrm{nn},i} > k_{\mathrm{nn},i}$ if the edges with the larger weights point to the neighbors with larger degree and $k^{\mathrm{w}}_{\mathrm{nn},i} < k_{nn,i}$ for the opposite (see Figure 1.5). Thus, $k^{\mathrm{w}}_{\mathrm{nn},i}$ measures the effective *affinity* to connect with high or low degree neighbors according to the magnitude of the actual interactions. Also, the behavior of the function $k^{w}_{nn}(k)$ (defined as the average of $k^{w}_{nn,i}$ over all vertices with degree k) marks the weighted assortative or disassortative properties considering the actual interactions among the system's elements.

A final note should be made concerning the local heterogeneities introduced by weights. The strength of a node i is the sum of the weights of all links in which i participates. The same strength can, however, be obtained with very different configurations: the weights w_{ij} may be either of the same order s_i/k_i or heterogeneously distributed among the links. For example, the most heterogeneous situation is obtained when one weight dominates over all the others. A simple way to measure this "disparity" is given by the quantity Y_2 introduced in other contexts (Herfindal, 1959; Hirschman, 1964; Derrida and Flyvbjerg, 1987; Barthélemy, Gondran and Guichard, 2003).

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

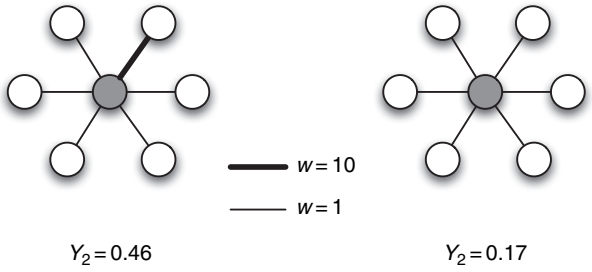

Fig. 1.6. If there is a small number of important connections around a node, the quantity Y_2 is of order $1/m$ with m of order unity. In contrast, if all the connections are of the same order, Y_2 is small and of order $1/k$ where k is the node's degree.

$$Y_2(i) = \sum_{j \in \mathcal{V}(i)} \left[\frac{w_{ij}}{s_i} \right]^2 . \tag{1.33}$$

If all weights are of the same order then $Y_2 \sim 1/k_i$ (for $k_i \gg 1$) and if a small number of weights dominate then Y_2 is of the order $1/m$ with m of order unity (see Figure 1.6).

A similar finding is encoded in the local entropy, defined for nodes of degree larger than 2 as

$$f(i) = -\frac{1}{\ln k_i} \sum_{j \in \mathcal{V}(i)} \frac{w_{ij}}{s_i} \ln \left[\frac{w_{ij}}{s_i} \right] . \tag{1.34}$$

This quantity goes from 0 if the strength of i is fully concentrated on one link to the maximal value 1 for homogeneous weights: it can thus be used as an alternative or complement to the disparity Y_2 to investigate the local heterogeneity of the weights.

It is important to stress regarding weighted networks that we have emphasized some of the measures that are customarily used to analyze large-scale complex systems. In particular, most of them have been adopted in the context of the recent analysis of infrastructure and communication networks, and will be used to evaluate the effect of heterogeneities and complexity in dynamical phenomena affected by the weighted nature of the connections. Several other quantities related to weighted properties are, however, defined and used in graph theory. In particular, weighted distances between two nodes are defined by assigning an effective length (depending on the weight) to each edge and summing these lengths along the path followed. The minimum spanning tree corresponds to the way of connecting all the vertices together with the smallest possible global weight. In standard graph theory, mostly *flows* have been studied, which represent a particular case of weighted networks where the weights must satisfy the conservation equation at each node, stating that the total amount of flow into a node is equal to the amount of flow

https://doi.org/10.1017/CBO9780511791383.002 Published online by Cambridge University Press

going out of it. Although this problem is very important in many instances (such as electrical networks, fluids in pipes, etc.), the conservation equation is not satisfied in all real-world networks and we refer the reader interested in flow problems (such as the maximum flow) to graph theory books (Clark and Holton, 1991; Cormen *et al.*, 2003; Jungnickel, 2004).