2

Networks and complexity

Undeniably, the visualizations of the Internet or the airport network convey the notion of intricate, in some cases haphazard, systems of a very complicated nature. *Complexity*, however, is not the same as the addition of complicated features. Despite the fact that there is no unique and commonly accepted definition of complexity – it is indeed very unlikely to find two scientists sharing the same definition of *complex system* – we discuss from what perspectives many real-world networks can be considered as complex systems, and what are the peculiar features signaling this occurrence. In this chapter we review the basic topological and dynamical features that characterize real-world networks and we attempt to categorize networks into a few broad classes according to their observed statistical properties. In particular, self-organized dynamical evolution and the emergence of the small-world and scale-free properties of many networks are discussed as prominent concepts which have led to a paradigm shift in which the dynamics of networks have become a central issue in their characterization as well as in their modeling (which will be discussed in the next chapter). We do not aim, however, at an exhaustive exposition of the theory and modeling of complex networked structures since, as of today, there are reference textbooks on these topics such as those by Dorogovtsev and Mendes (2003), Pastor-Satorras and Vespignani (2004), and Caldarelli (2007), along with journal reviews by Albert and Barabási (2002), Dorogovtsev and Mendes (2002), Newman (2003c), and Boccaletti *et al.* ([2006]).

2.1 Real-world systems

In recent times, the increased power of computers and the informatics revolution have made possible the systematic gathering and handling of data sets on several large-scale networks, allowing detailed analysis of their structural and functional properties. As a first guide to the classification of the data obtained, we can provide a rudimentary taxonomy of real-world networks. Two main different classes

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

are infrastructure systems, and natural or living systems. Each of these classes can be further divided into different subgroups. Natural systems networks can be differentiated into the subgroups of biology, social systems, food webs and ecosystems, among others. For instance, biological networks refer to the complicated sets of interactions among genes, proteins and molecular processes which regulate biological life, while social networks describe the relations between individuals such as family links, friendships, and work relations. In particular, leading sociologists refer to our societies as networked societies and even if their analysis has been largely focused on small and medium-scale networks, social network studies play a key role in introducing basic definitions and quantities of network science. Indeed, biological and social networks are prominent research topics and much of network science is influenced by what has happened in these fields recently.

In turning our attention to infrastructure networks we can readily separate two main subcategories. The first contains virtual or cyber networks. These networks exist and operate in the digital world of cyberspace. The second subcategory includes physical systems such as energy and transportation networks. This is a rough classification since there are many interrelations and interdependencies existing among physical infrastructure networks, as well as between physical and digital networks. In the Internet, for instance, the cyber features are mixed with the physical features. The physical Internet is composed of physical objects such as routers – the main computers which allow us to communicate – and transmission lines, the cables which connect the various computers. On top of this physical layer lies a virtual world made of software that may define different networks, such as the World Wide Web (WWW), email networks, and Peer-to-Peer networks. These networks are the information transfer channels for hundreds of millions of users and, like the physical Internet, have grown to become enormous and intricate networks as the result of a self-organized growing process. Their dynamics are the outcomes of the interactions among the many individuals forming the various communities, and therefore are mixtures of complex socio-technical aspects. Infrastructure networks represent a combination of social, economic, and technological processes. Further examples can be found in the worldwide airport network (WAN) and power distribution networks where physical and technological constraints cooperate with social, demographic, and economic factors.

2.1.1 Networks everywhere

The various systems mentioned so far are characterized by the very different nature of their constitutive elements. It is therefore appropriate to offer a list of some network data sets prototypically considered in the literature, making clear to which property and elements their graph representation refers.

Social networks

The science of social networks is one of the pillars upon which the whole field of network science has been built. Since the early works of Moreno (1934) and the definition of the sociogram, social networks have been the object of constant analysis and study. Social networks represent the individuals of the population as nodes and the social ties or relations among individuals as links between these nodes. The links therefore may refer to very different attributes such as friendship among classmates, sexual relations among adults, or just the belonging to common institutions or work teams (collaborative interactions). The importance of these networks goes beyond social sciences and affects our understanding of a variety of processes ranging from the spreading of sexually transmitted diseases to the emergence of consensus and knowledge diffusion in different kinds of organizations and social structures.

Recently, the recording of social interactions and data in electronic format has made available data sets of unprecedented size. The e-mail exchanges in large corporate organizations and academic institutions make tracking social interactions among thousands of individuals possible in a precise and quantitative way (Ebel, Mielsch and Bornholdt, 2002; Newman, Forrest and Balthrop, 2002). Habits and shared interests may be inferred from web visits and file sharing. Professional communities have been analyzed from wide databases such as complex collaboration networks. Examples are the already classic collaboration network of film actors (Watts and Strogatz, 1998; Barabási and Albert 1999; Amaral *et al.*, 2000; Ramasco, Dorogovtsev and Pastor-Satorras, 2004), the company directors network (Newman, Strogatz and Watts, 2001; Davis, Yoo and Baker, 2003; Battiston and Catanzaro, 2004) and the network of co-authorship among scientists (Newman, 2001a; 2001b; 2001c). In these last examples we encounter bipartite networks (Wasserman and Faust, 1994), although the one mode projection is often used so that members are tied through common participation in one or more films, boards of directors, or academic papers. As an example, we report in Figure 2.1 an illustration of a construction of the scientific collaboration network (SCN). According to this construction, the authors are the nodes of the network and share an edge if they co-authored a paper. Similarly, for the actors' collaboration network, two actors are linked if they have co-starred in a movie. More information can be projected in the graph by weighting the intensity of the collaboration. A convenient definition of the weight is introduced by Newman (2001b), who considers that the intensity w_{ij} of the interaction between two collaborators i and j is given by

$$w_{ij} = \sum_p \frac{\delta_i^p \delta_j^p}{n_p - 1}, \tag{2.1}$$

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

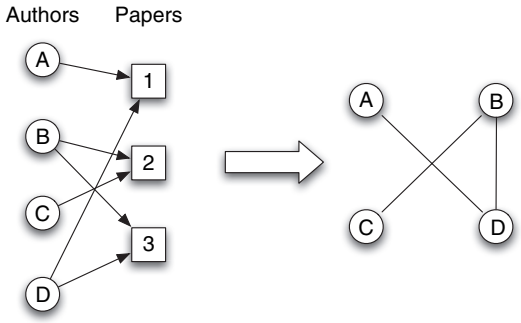


Fig. 2.1. A bipartite graph is defined as a graph whose nodes can be divided into two separate sets or classes such that every edge links a node of one set to a node of the other set. For a co-authorship network, authors and papers form two distinct classes and the edges representing the authorship just go from authors to papers. The one mode projection defines a unipartite graph, where the nodes of one class share an edge if they are connected to a common node of the other class. The one mode projection on the author set defines the co-authorship network in which authors share an edge if they co-authored at least one paper.

where the index p runs over all co-authored papers, n_p is the number of authors of the paper p, and δ_i^p is 1 if author i has contributed to paper p, and 0 otherwise. The strength of the interactions is therefore large for collaborators having many papers in common but the contribution to the weight introduced by any given paper is inversely proportional to the number of authors. Similarly, other definitions of weight may be introduced in order to measure the impact of collaborations in the scientific community. For instance, Börner *et al.* (2004) consider that the weight of each edge is also a function of the number of citations of each paper, providing a quantitative assessment of the impact of co-authorship teams. This is an example of social science merging with bibliometrics, another discipline which has recently benefited from the impact of the e-revolution and text digitalization on gathering large-scale network data sets. Just to cite a paramount example, the networks of citations (there is a directed link between two papers if one cites the other) among scientific papers of several databases for journals such as the *Proceedings of the National Academy of Sciences* (*PNAS*) or *Physical Review* contain thousands of nodes (Redner, 2005).

Transportation networks

Transportation systems such as roads, highways, rails, or airlines are crucial in our modern societies and will become even more important in a world where more than 50% of the population lives in urban areas.[1] The importance of this subject

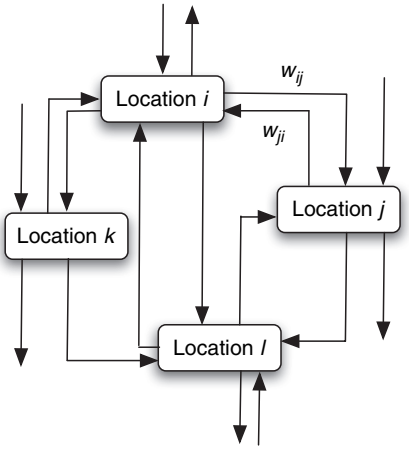
[1] Source: United Nations, population division `http://www.unpopulation.org`

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

justifies the huge literature published on these systems for at least the past 70 years. Studies cover a broad range of approaches from applied engineering to mathematic works on users' equilibrium (for an introductory book on the subject, see Sheffi [1985]). Geographers, regional science experts, and economists have long observed the existence of structures such as hierarchies or particular subgraphs present in the topology (see for example Yerra and Levinson [2005] and references therein). In the air traffic network, for instance, O'Kelly (1998) has discussed the existence of hubs and of particular motifs. Hierarchies present in a system of cities were also noticed a long time ago, and Christaller (1966) proposed his famous central place theory in which heterogeneous distributions of facilities and transportation costs are the cause of the emergence of hierarchies. More recently, Fujita, Krugman and Venables (1999) have proposed a model which explains the hierarchical formation of cities in terms of decentralized market processes. These different models are, however, very theoretical, and more thorough comparison with empirical data is needed at this stage. Only recently, large-scale data and extensive statistics have opened the path to better characterization of these networks and their interaction with economic and geographic features. In this context, the network representation operates at different scales and different representative granularity levels are considered.

The TRANSIMS project characterizes human flows at the urban level (Chowell *et al*., 2003; Eubank *et al*., 2004). This study concerns the network of locations in the city of Portland, Oregon, where nodes are city locations including homes, offices, shops, and recreational areas. The edges among locations represent the flow of individuals going at a certain time from a location to another. In Figure 2.2 we illustrate the network construction and we report an example of the type of data set used by Chowell, Hyman, Eubank and Castillo-Chavez (2003). Other relevant networks studied at this scale are subways (Latora and Marchiori, 2002), and roads and city streets (Cardillo *et al.*, 2006; Buhl *et al.*, 2006; Crucitti, Latora and Porta, 2006; Scellato *et al.*, 2006; Kalapala *et al.*, 2006).

At a slightly coarser scale, it is possible to map into a weighted network the commuting patterns among urban areas, municipalities, and counties (Montis *et al.*, 2007), as well as railway systems (Sen *et al.*, 2003). Finally, on the global scale, several studies have provided extensive analysis of the complete worldwide airport network and the relative traffic flows (Barrat *et al*., 2004a; Guimerà and Amaral, 2004; Guimerà *et al.*, 2005): this transportation system can be represented as a weighted graph where vertices denote airports, and weighted edges account for the passenger flows between the airports. The graphs resulting from large-scale transportation networks define the basic structure of metapopulation or patch models used in studies of epidemic spread. In this representation the



Person ID	Location ID	Location type	Arrival time	Departure time
116	4356	home	00:00	07:00
116	98135	work	08:00	11:00
116	71457	work	11:20	13:00
116	98135	work	13:20	17:00
116	4356	home	18:00	19:15
116	21343	social	19:30	21:00
116	4356	home	21:00	07:00
324	12679	home	00:00	07:00
324	431	school	08:00	14:00
324	12679	home	14:30	19:00
⋮	⋮	⋮	⋮	⋮

Fig. 2.2. Construction of the network between the various locations in the city of Portland as reported by Chowell *et al.* (2003). The table is a small sample of the data set, obtained through a TRANSIMS simulation, which records for each individual the successive locations visited during a day, with the arrival and departure time. The network between locations is constructed from this data set: an edge exists between location i and j whenever an individual has gone directly from i to j, and the weight w_{ij} gives the daily traffic between i and j.

network's nodes refer to different populations such as urban areas or cities, and an edge between two nodes denotes an exchange of individuals between the two populations (see also Chapter 9). These networks are usually directed and weighted in that the number of individuals going from one population to the other is not necessarily symmetric and the weight of each edge quantifies the number of exchanged individuals. In general, transportation networks naturally define such networked structures, with the edges' weights denoting the traffic flows (commuters, passengers etc.) between the different locations, and network theory appears as the natural tool for the study of these systems.

Internet

Characterizing how routers, computers, and physical links interconnect with each other in the global Internet is very difficult because of several key features of the network. One main problem is related to its exponential growth rate. The Internet, in fact, has already increased by five orders of magnitude since its birth. A second difficulty is the intrinsic *heterogeneity* of the Internet which is composed of networks engineered with considerable technical and administrative diversity. The different networking technologies are merged together by the TCP/IP architecture which provides connectivity but does not imply a uniform behavior. Also very important is the fact that the Internet is a *self-organizing* system whose properties

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

cannot be traced back to any global blueprint or chart. This means that routers and links are added by competing entities according to local economic and technical constraints, leading to a very intricate physical structure that does not comply with any globally optimized plan (Pastor-Satorras and Vespignani, 2004; Crovella and Krishnamurthy, 2006).

The combination of all these factors results in a general lack of understanding about the large-scale topology of the Internet, but in recent years, several research groups have started to deploy technologies and infrastructures in order to obtain a more global picture of this network. Efforts to obtain such maps have been focused essentially on two levels. First, the inference of router adjacencies amounts to a measure of the internet router (IR) level graph. The second mapping level concerns the Autonomous System (AS) graph of the Internet, referring to autonomously administered *domains* which to a first approximation correspond to internet service providers and organizations. Therefore, internet maps are usually viewed as undirected graphs in which vertices represent routers or Autonomous Systems and edges (links) represent the physical connections between them. Although these two graph representations are related, it is clear that they describe the Internet at rather different scales (see Figure 2.3). In fact, the collection of Autonomous Systems and inter-domain routing systems defines a coarse-grained picture of the Internet in which each AS groups many routers together, and links are the aggregations of all the individual connections between the routers of the corresponding AS.

Internet connectivity information at the level of Autonomous Systems can be retrieved by the inspection of routing tables and paths stored in each router (passive measurements) or by direct exploration of the network with a software probe (active measurements). Based on the first strategy, the *Oregon Route Views* project (RV) provides maps of the AS graph obtained from the knowledge of the routing

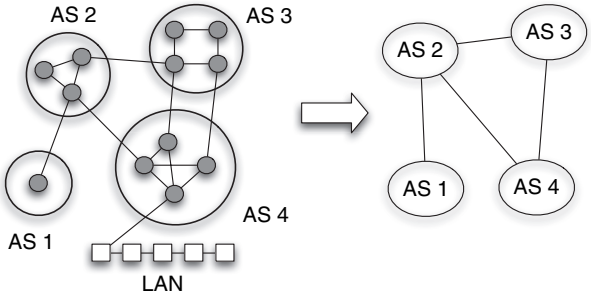


Fig. 2.3. Different granularity representations of the Internet. The hosts (squares) are included in the Local Area Networks (LAN) that connect to routers (shaded circles) and are excluded from the maps. Each Autonomous System is composed of many routers.

tables of several Border Gateway Protocol (BGP) peers.[2] This project has been running since 1997 and is one of the first projects to provide regular snapshots of the Internet's evolution. On the other hand, the most famous large infrastructure for active measurement has been implemented by the skitter project at CAIDA.[3] This project deployed several strategically placed probing monitors devoted to Internet mapping and measurement (Huffaker *et al*., 2002b). All the data are then centrally collected and merged in order to obtain large-scale Internet maps that minimize measurement biases. A different active strategy, the Distributed Internet Measurements and Simulations (DIMES) project (Shavitt and Shir, 2005), considers a distributed measurement infrastructure, based on the deployment of thousands of lightweight measurement agents around the globe. Several other projects have focused attention on different levels such as the maps of specific Internet providers or Internet regions.

World Wide Web

The World Wide Web (WWW) is probably the most famous virtual network living on the physical structure of the Internet. It is nowadays considered as a critical infrastructure in the sense that it has acquired a central role in the everyday functioning of our society, from purchasing airplane tickets to business teleconferences. The Web is so successful that its rapid and unregulated growth has led to a huge and complex network for which it is extremely difficult to estimate the total number of web pages, if possible at all.

The experiments aimed at studying the Web's graph structure are based on Web *crawlers* which explore connectivity properties by following the links found on each page. In practice, crawlers are special programs that, starting from a source page, detect and store all the links they encounter, then follow them to build up a set of pages reachable from the starting one. This process is then repeated for all pages retrieved, obtaining a second layer of vertices and so on, iterating for as many possible layers as allowed by the available storage capacity and CPU time. From the collected data it is then possible to reconstruct a graph representation of the WWW by identifying vertices with web pages and edges with the connecting hyperlinks.

Large crawls of the WWW have been constantly gathered for several years because of the importance of indexing and storing web pages for search engine effectiveness. This data availability has made the WWW the starting place for the study of large-scale networks (Albert, Jeong and Barabási, 1999; Broder

[2] University of Oregon Route Views Project. `http://www.routeviews.org/`

[3] Cooperative Association for Internet Data Analysis; `http://www.caida.org/tools/measurement/skitter/router_topology/`

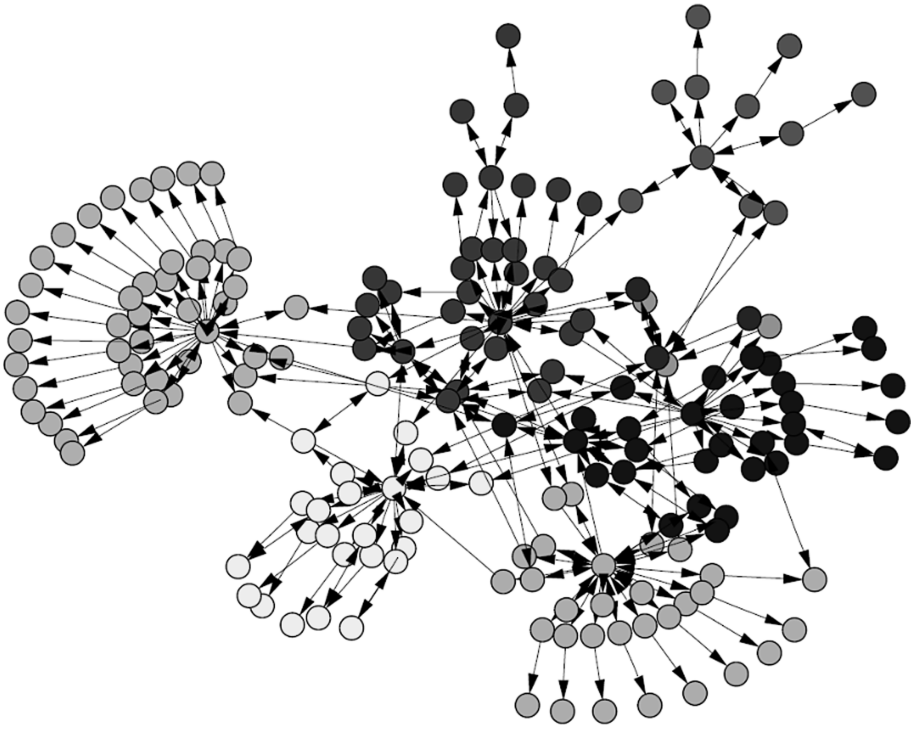


Fig. 2.4. Pages of a website and the directed hyperlinks between them. The graph represents 180 pages from the website of a large corporation. Different shades identify topics as identified with a community detection algorithm. Reprinted with permission from Newman and Girvan (2004). Copyright 2004 by the American Physical Society.

et al., 2000; Adamic and Huberman, 2001). The sheer size of the web graphs has also made possible the investigation of large-scale statistical properties and has led to the development of new measurement tools aimed at characterizing graphs with 10^7–10^8 vertices. In addition, the web graph is a prototypical *directed* graph, in which edges connect ordered pairs of vertices (see Chapter 1). Crawls count the number of each web page's outgoing hyperlinks, but in principle we know nothing about the incoming hyperlinks from other pages (see Figure 2.4). We can follow hyperlinks to reach pointed pages, but we cannot navigate backwards through the incoming hyperlinks. From this perspective, web graphs are often the gold standard used to address the discussion of paths, directionality, and component structure in large-scale directed networks.

Biological networks

It is fair to say that one of the reasons for the explosion of interest in networks lies in the possibility, due to high throughput experiments, of gathering large data

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

collections on the interactions or relations of entire proteomes or genomes, also called the "omic" revolution. Networks now pervade the biological world at various levels ranging from the microscopic realm of biological chemistry, genetics, and proteomics to the large scale of food webs.

At the microscopic level, many relevant aspects of biological complexity are encapsulated in the structure and dynamics of the networks emerging at different organizational levels, ranging from intracellular biochemical pathways to genetic regulatory networks or protein interaction networks (Barabási and Oltvai, 2004; Alon, 2003). A prominent example in this area is provided by protein interaction networks (PIN) of various organisms which can be mathematically represented as graphs with nodes representing proteins and edges connecting pairs of interacting proteins (see Chapter 12). The importance of microscopic biological networks is clearly related to the biological significance of the network's topology, and analysis in this direction has indeed pointed out correlation signatures between gene knock-out lethality and the connectivity of the encoded protein, negative correlation between the evolution rate of a protein and its connectivity, and functional constraints in protein complexes.

At a larger scale, biological networks can describe individuals' interactions in various animal and human populations. In this area, biology may overlap with social science. A typical example is given by the network describing the web of sexual relations that is both of interest from a social point of view and of great concern in the epidemiology of sexually transmitted diseases (Liljeros *et al.*, 2001; Schneeberger *et al.*, 2004).

Finally, at the very large scale we find the networks describing the food web of entire ecosystems. Roughly speaking, food webs describe which species eat which other species. In this perspective, nodes represent species and links are antagonistic trophic interactions of the predator–prey type (see Dunne, Williams and Martinez [2002a]; Montoya, Pimm and Solé [2006] and Chapter 12).

2.1.2 Measurements and biases

In the previous section and throughout the rest of the book we have made a particular effort to discuss and cite some of the most recent network data sets available in the literature. Data gathering projects, however, are continuously making public new data on large-scale networks and very likely larger and more accurate samples than those mentioned here will be available by the time this book is published. In addition the list of systems for which it is possible to analyze network data is continually enlarging. Genomic and proteomic data of new organisms are constantly added to the various community data repositories. New and very large maps of the WWW are acquired along with a better knowledge of the physical Internet.

In addition, many other networks related to the cyberworld are explored, such as Peer-to-Peer or email networks. Data on large-scale infrastructures such as transportation, power-grid, freight and economic networks are also constantly enlarging our inventory of network data.

On the other hand, approaching network data requires great caution and a critical perspective informed about all the details of the data gathering process. Indeed, for many systems the acquisition of the complete network structure is impossible owing to time, resource, or technical constraints. In this case, network sampling techniques are applied to acquire the most reliable data set and minimize the error or biases introduced in the measurement. The aim is to obtain network samples that exhibit reliable statistical properties resembling those of the entire network. Network sampling and the relative discussion of statistical reliability of the data sets are therefore a major issue in the area of network theory, unfortunately not always carefully scrutinized (Willinger *et al.*, 2002).

Examples of sampling issues can be found in all of the areas discussed previously. Crawling strategies to gather WWW data rely on exhaustive searches by following hyperlinks. Internet exploration consists of sending probes along the computer connections and storing the physical paths of these probes. These techniques can be applied recursively or repeated from different vantage points in order to maximize the discovered portion of the network. In any case, it is impossible to know the actual fraction of elements sampled in the network (Broido and claffy, 2001; Barford *et al.*, 2001; Qian *et al.*, 2002; Huffaker *et al.*, 2000; Huffaker *et al.*, 2002a; 2002b). In biological networks, the sources of biases lie in the intrinsic experimental error that may lead to the presence of a false positive or negative on the presence of a node or edge. A typical example can be found in high throughput techniques in biological network measurements such as experiments for detecting protein interactions (Bader and Hogue, 2002; Deane *et al.*, 2002). For these reasons, a large number of model-based techniques, such as probabilistic sampling design (Frank, 2004) developed in statistics, provide guidance in the selection of the initial data sets. In addition, the explosion in data gathering has spurred several studies devoted to the bias contained in specific, large-scale sampling of information networks or biological experiments (Lakhina *et al.*, 2002; Petermann and De Los Rios, 2004a; Clauset and Moore, 2005; Dall'Asta *et al.*, 2005, 2006a; Viger *et al.*, 2007).

2.2 Network classes

The extreme differences in the nature of the elements forming the networks considered so far might lead to the conclusion that they share few, if any, features.

The only clear commonalities are found in the seemingly intricate haphazard set of points and connections that their graph representations produce, and in their very large size. It is the latter characteristic that, making them amenable to large-scale statistical analysis, allowed the scientific community to uncover shared properties and ubiquitous patterns which can be expressed in clear mathematical terms.

2.2.1 Small-world yet clustered

Let us first consider the size of networks and the distance among elements. Even though graphs usually lack a metric, it is possible to define the distance between any pair of vertices as the number of edges traversed by the shortest connecting path (see Chapter 1). A first general evidence points toward the so-called small-world phenomenon. This concept describes in simple words that it is possible to go from one vertex to any other in the system passing through a very small number of intermediate vertices. The small-world effect, sometimes referred to as the "six degrees of separation" phenomenon, has been popularized in the sociological context by Milgram (1967) who has shown that a short number of acquaintances (on average six) is enough to create a connection between any two people chosen at random in the United States. Since then, the small-world effect has been identified as a common feature in a wide array of networks, in particular infrastructure networks where the small average distance is crucially important to speed up communications. For instance, if the Internet had the shape of a regular grid, its characteristic distance would scale with the number of nodes N as $\langle \ell \rangle \sim N^{1/2}$; with the present Internet size, each Internet Protocol packet would pass through 10^3 or more routers, drastically depleting communication capabilities. The small-world property is therefore implicitly enforced in the network architecture which incorporates hubs and backbones connecting different regional networks, thus strongly decreasing the value of $\langle \ell \rangle$.

To be more precise, the small-world property refers to networks in which $\langle \ell \rangle$ scales logarithmically, or more slowly, with the number of vertices. In many cases, data on the same network at different sizes are not available. The small-world property is then measured by inspecting the behavior of the quantity $M(\ell)$ defined as the average number of nodes within a distance less than or equal to ℓ from any given vertex. While in regular lattices $M(\ell)$ is expected to increase as a power law with the distance ℓ, in small-world networks this quantity follows an exponential or faster increase. In Figure 2.5 we report this quantity for several prototypical networks. At first sight the small-world feature appears a very peculiar property. By itself, however, it is not the signature of any special organizing principle, and it finds its explanation in the basic evidence that randomness appears as a major

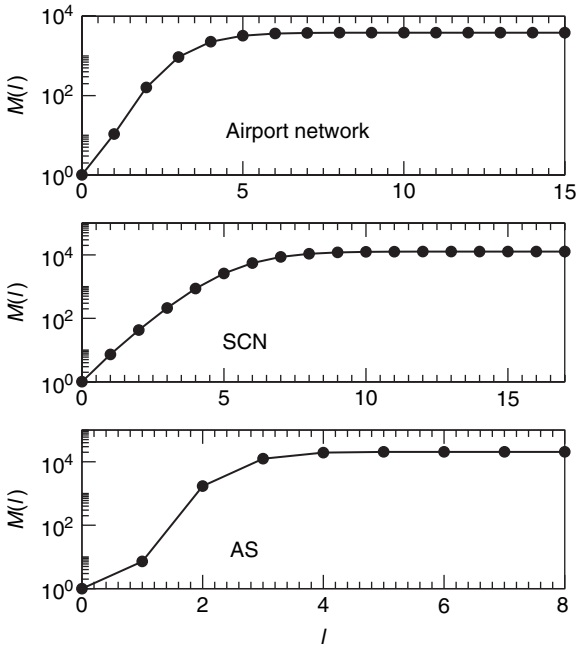


Fig. 2.5. Hop-plot for several prototypical networks. In all cases the number of nodes at distance l or less grows exponentially fast as shown by the behavior on a linear–log scale, before saturating because of the finite size of the network. The data sets considered are, from top to bottom: the worldwide airport network (data from the International Air Transport Association IATA, see Barrat *et al.* (2004a) and `http://www.iata.org`), the scientific collaboration network (SCN), see Newman (2001a; 2001b; 2001c) and `http://www-personal.umich.edu/~mejn/netdata/`, and the map of the Autonomous Systems of the Internet (AS) as obtained by the DIMES project (`http://www.netdimes.org`).

ingredient in the shaping of large-scale networks. Indeed, by looking at the intricacy and apparent absence of order of networks, the simplest assumption one can make is that the creation of a connection between two elements is a random event determined by the sum of a very large number of unpredictable circumstances. This is the guiding principle that defines the Erdős–Rényi model and the general formulation of random graph models and algorithms presented in Chapter 3. In brief, all of these models assume that the presence of an edge between two vertices is a random process occurring with the same probability, independent of the vertex characteristics. This simple construction suffices in yielding small-world graphs. It can be shown rigorously that the average shortest path distance between any two vertices in the graph increases just as the logarithm of the number N of vertices

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

considered. In other words, the small-world feature can be explained with the simple inclusion of randomness. This readily explains the ubiquity of this property since in all natural systems we have to allow for the presence of some level of noise and stochastic behavior in the dynamics at the origin of the networks.

More interesting evidence is that, in many social and technological networks, the small-world effect goes along with a high level of clustering (Watts and Strogatz, 1998; Watts, 1999). The clustering coefficient characterizes the local cohesiveness of the networks as the tendency to form groups of interconnected elements (see Chapter 1). It is easy to perceive that a random graph model cannot achieve high clustering in that no organizing principle is driving the formation of groups of interconnected elements. Interconnections happen only by chance, randomness being the only force shaping the network, and therefore in large random graphs the clustering coefficient becomes very small (see Chapter 3). In other words, random graphs feature the small-world effect but are not clustered, while regular grids tend to be clustered but are not small-world. This means that the high clustering is a memory of a grid-like ordering arrangement that is not readily understandable in a purely random construction. This puzzle has been addressed by the famous small-world model of Watts and Strogatz (1998) (see Chapter 3) which is able to capture both properties at the same time.

2.2.2 Heterogeneity and heavy tails

The evidence for the presence of more subtle structural organizations in real networks is confirmed by the statistical analysis of centrality measures. The functional form of the statistical distributions characterizing large-scale networks defines two broad network classes. The first refers to the so-called statistically *homogeneous* networks. The distributions characterizing the degree, betweenness, and weighted quantities have functional forms with fast decaying or "light" tails such as Gaussian or Poisson distributions. The second class concerns networks with statistically *heterogeneous* connectivity and weight patterns usually corresponding to skewed and heavy-tailed distributions. In order to better understand the basic difference between these two classes, let us focus for the moment on the degree distribution $P(k)$. The evidence for the high level of heterogeneity of many networks is simply provided by the fact that many vertices have just a few connections, while a few hubs collect hundreds or even thousands of edges. For instance, this feature is easily seen in the airport networks (Figure 2.6), showing the "hub" policy that almost all airlines have adopted since deregulation in 1978. The same arrangement can easily be perceived in many other networks where the presence of "hubs" is a natural consequence of different factors such as popularity, strategies,

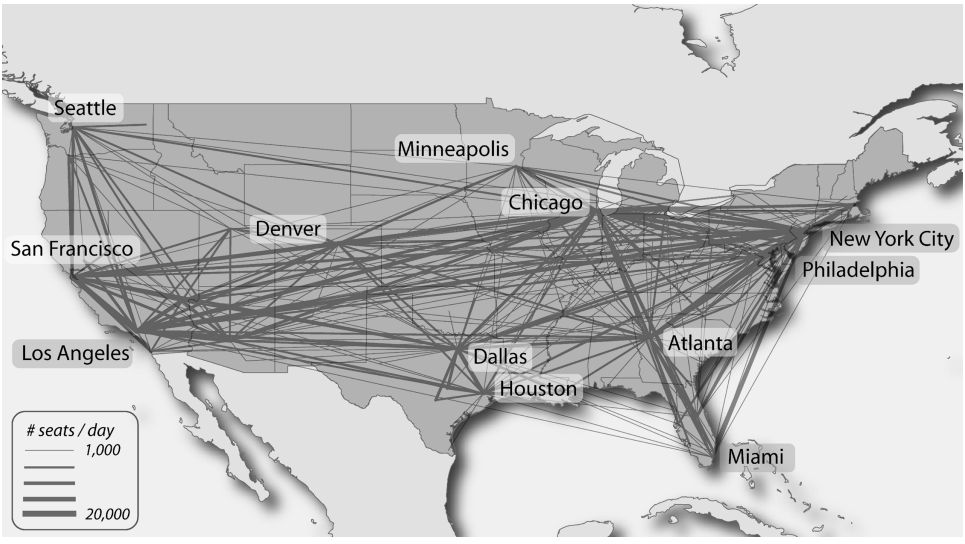


Fig. 2.6. Main air travel routes for the United States. Only the connections with more than 1000 available seats per day are shown. The presence of hubs is clearly visible. Figure courtesy of V. Colizza.

and optimization. For instance, in the WWW some pages become hugely popular and are pointed to by thousands of other pages, while in general most documents are almost unknown.

The presence of hubs and connectivity ordering turns out to have a more striking manifestation than initially thought, yielding in many cases a degree distribution $P(k)$ with heavy tails (Barabási and Albert, 1999). In Figure 2.7 we show the degree distribution resulting from the analysis of several real-world networks. In all cases the distribution is skewed and highly variable in the sense that degrees vary over a broad range, spanning several orders of magnitude. This behavior is very different from the case of the bell-shaped, exponentially decaying distributions and in several cases the heavy tail can be approximated by a power-law decay $P(k) \sim k^{-\gamma}$,[4] which results in a linear behavior on the double logarithmic scale. In distributions with heavy tails, vertices with degree much larger than the average $\langle k \rangle$ are found with a non-negligible probability. In other words, the average behavior of the system is not typical. In the distributions shown in Figure 2.7, the vertices will often have a small degree, but there is an appreciable probability of finding vertices with very large degree values. Yet all intermediate values are present and the average degree does not represent any special value for the distribution. This is

[4] Power-law distributions are in many cases also referred to as Pareto distributions.

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

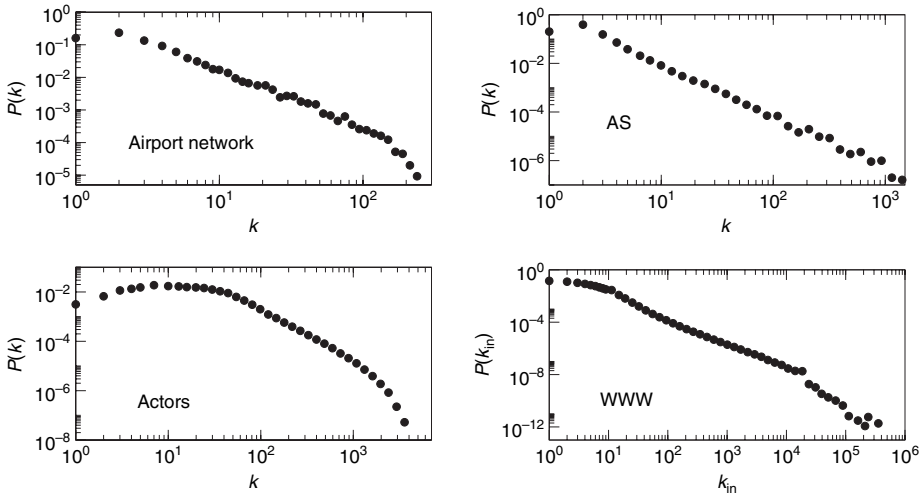


Fig. 2.7. Degree distribution $P(k)$ of real-world networks described in Section 2.1. Worldwide airport network (IATA data, see `http://www.iata.org/`); actors' collaboration network (Internet Movie database at `http://www.imdb.com/` and `http://www.nd.edu/~networks/resources.htm`); map of the Autonomous Systems of the Internet obtained by the DIMES mapping project (`http://www.netdimes.org`); map of the WWW collected in 2003 by the WebBase project (`http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/`). For the WWW we report the distribution of the in-degree k_{in}. Data courtesy of M. A. Serrano.

in strong contrast to the democratic perspective representation offered by homogeneous networks with bell-shaped distributions and fast decaying tails. The average value here is very close to the maximum of the distribution which corresponds to the most probable value in the system. The contrast between these types of distributions is illustrated in Figure 2.8, where we compare a Poisson and a power-law distribution with the same average degree.

In more mathematical terms, the heavy-tail property translates to a very large level of degree fluctuations. The significance of the heterogeneity contained in heavy-tailed distributions can be understood by looking at the first two moments of the distribution. We can easily compute the average value that the degree assumes in the network as

$$\langle k \rangle = \int_m^\infty k P(k)\, \mathrm{d}k, \tag{2.2}$$

where $m \geq 1$ is the lowest possible degree in the network. Here for the sake of simplicity we consider that k is a continuous variable but the same results hold in the discrete case, where the integral is replaced by a discrete sum. By computing the above integral, it is possible to observe that in any distribution with a power-law

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

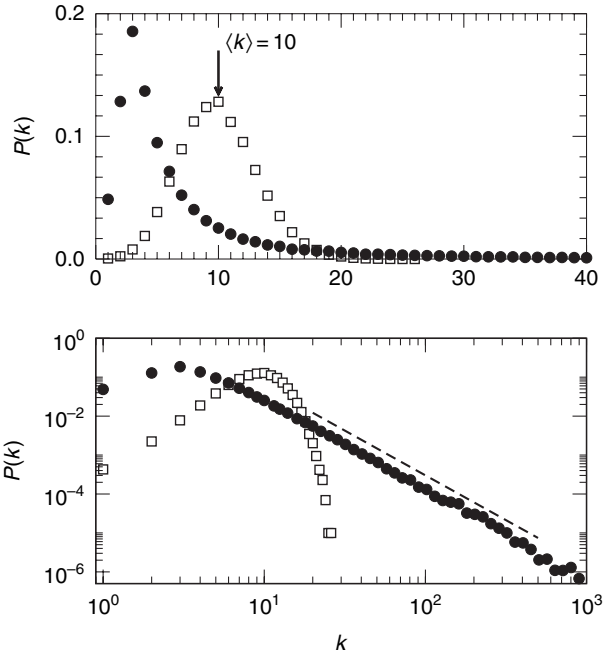


Fig. 2.8. Comparison of a Poisson and power-law degree distribution on a linear scale plot (top) and a double logarithmic plot (bottom). The two distributions have the same average degree $\langle k \rangle = 10$. The dashed line in the bottom figure corresponds to the power law $k^{-\gamma}$, where $\gamma = 2.3$.

tail with exponent $2 < \gamma \leq 3$, the average degree is well defined and bounded. On the other hand, a measure of the typical error we make if we assume that $\langle k \rangle$ is the typical degree value of a vertex is given by the normalized variance of the distribution $\sigma^2/\langle k \rangle^2$ that expresses the statistical fluctuations present in our system. The variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$ is dominated by the second moment of the distribution

$$\langle k^2 \rangle \sim \int k^2 P(k)\, \mathrm{d}k \sim \int_m^{k_c} k^{2-\gamma}\, \mathrm{d}k \sim k_c^{3-\gamma}. \tag{2.3}$$

In the asymptotic limit of infinite network sizes, the cut-off k_c corresponding to the largest possible degree value diverges ($k_c \to \infty$) (see Appendix 1), so that $\langle k^2 \rangle \to \infty$: fluctuations are unbounded and depend on the system size.[5] The absence of any intrinsic scale for the fluctuations implies that the average value is not a characteristic scale for the system. In other words, we observe a *scale-free* network as far as the degree of the vertices is concerned. This reasoning can

[5] In many cases we will refer to the infinite size limit as the *thermodynamic limit*.

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

be extended to values of $\gamma \leq 2$, since in this case even the first moment is unbounded.

The absence of an intrinsic characteristic scale in a power-law distribution is also reflected in the self-similarity properties of such a distribution; i.e. it looks the same at all length scales. This means that if we look at the distribution of degrees by using a coarser scale in which $k \to \lambda k$, with λ representing a magnification/reduction factor, the distribution would still have the same form. This is not the case if a well-defined characteristic length is present in the system. From the previous discussion, it is also possible to provide a heuristic characterization of the level of heterogeneity of networks by defining the parameter

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{2.4}$$

Indeed, fluctuations are denoted by the normalized variance which can be expressed as $\kappa/\langle k \rangle - 1$, and scale-free networks are characterized by $\kappa \to \infty$, whereas homogeneous networks have $\kappa \sim \langle k \rangle$. For this reason, we will generally refer to scale-free networks as all networks with heterogeneity parameter $\kappa \gg \langle k \rangle$.[6] We will see in the following chapters that κ is a key parameter for all properties and physical processes that are affected by the degree fluctuations. It is also important to note that, in the case of uncorrelated networks, $\kappa = k_{nn}(k)$ (see Equation (1.26)), so that there is a link between the divergence of fluctuations and the average degree of nearest neighbors.

The evidence of heavy-tail distributions is not found only in the connectivity properties. In general, betweenness distributions are also broad and exhibit scale-free behavior. Even more striking is the evidence for the heavy-tail character of weight and strength distributions with values spanning up to eight or nine orders of magnitude. In Figures 2.9 and 2.10 we report some examples of the betweenness, weight, and strength distributions observed in real networks data. It is worth commenting at this point that several studies have been devoted to whether the observed distributions can or cannot be approximated by a power-law behavior. A thorough discussion of the issues found in the measurement and characterization of power-law behavior is contained in the extensive review by Newman (2005b). While the presence or absence of power-law functional form is a well-formulated statistical question, it is clear that we are now describing real-world systems, for which in most cases the statistical properties are affected by noise, finite size, and other truncation effects. The presence of such effects, however, should not be considered a surprise. For instance, the heavy-tail truncation is the natural effect of the upper limit of the distribution, which must be necessarily present in every real world system. Indeed, bounded power laws (power-law distributions with a cut-off)

[6] Obviously, in the real world κ cannot be infinite since it is limited by the network size.

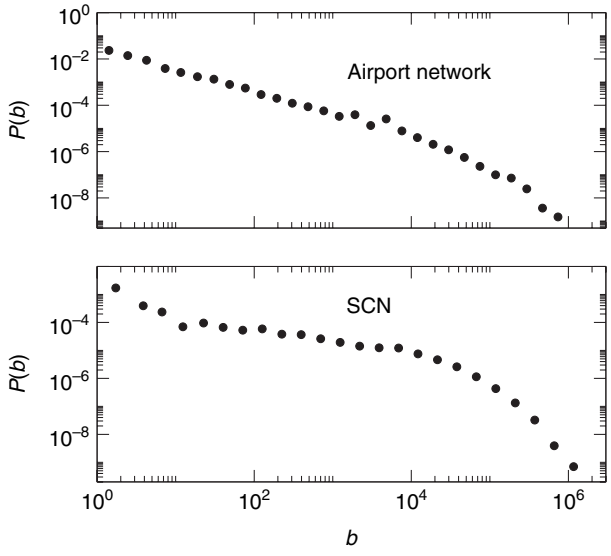


Fig. 2.9. Betweenness centrality distributions in the worldwide airport network (top, http://www.iata.org) and in the scientific collaboration network (SCN, bottom, http://www-personal.umich.edu/~mejn/netdata/).

are generally observed in real networks (Amaral *et al.*, 2000) and different mechanisms have been proposed to account for the presence of large degree truncations. In such a context, fitting to a power-law form can yield different results, depending on the range of values actually considered for the fit. In this sense, the exact value of the exponents or the precise analytical form of the observed behavior is of secondary importance, at least given the amount of noise often encountered in the data sets. In some other situations the behavior is surely more complicated than a pure power law, exponential cut-offs, or other functional forms being present. The crucial issue is that the observation of heavy-tailed, highly variable distributions provides statistical fluctuations which are extremely large and therefore cannot be neglected.

From this perspective, scale-free networks refer to all those systems in which fluctuations are orders of magnitude larger than expected values. Table 2.1 summarizes the numerical properties of some of the heavy-tailed probability distributions analyzed so far. The scale-free behavior is especially clear from the values of the heterogeneity parameter κ and the wide ranges spanned by the variables. The differences introduced by large fluctuations in many properties of the graph are too significant to be ignored: as we will see in the next chapters they have an enormous impact in the properties of the dynamical and physical processes occurring on these networks.

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

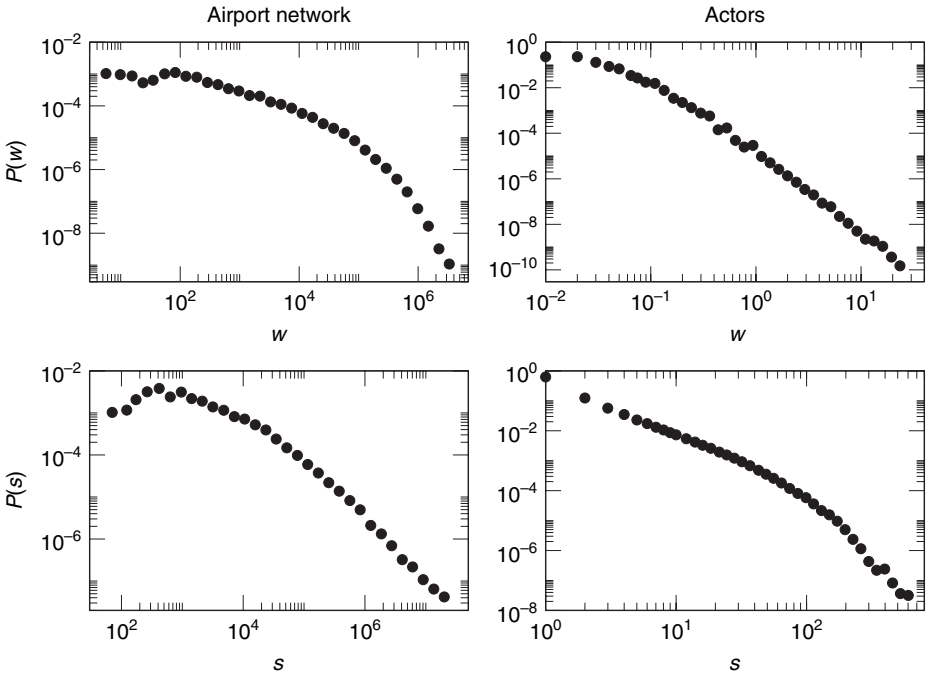


Fig. 2.10. Weight (top row) and strength (bottom row) distributions in the worldwide airport network (left column) and in the actors' collaboration network (right column). In the airport network (`http://www.iata.org`), the weight of a link corresponds to the average number of available seats per year between the two airports connected by this link (Barrat *et al*., 2004a). The strength therefore represents the average number of passengers (traffic) handled by an airport. For the actors, collaboration network (`http://www.imdb.com/`) the weight of a link connecting two actors i and j is given by Equation (2.1): the contribution of each co-starred movie is inversely proportional to the number of actors in the movie. The strength of an actor corresponds therefore to the number of movies in which she/he has appeared. All these quantities are clearly broad in distribution.

2.2.3 Higher order statistical properties of networks

A full account of the structural and hierarchical ordering calls for the analysis and study of higher order statistical indicators and other local and non-local graph measures. The list of interesting features and their specific measures is obviously strongly dependent on the specific network analyzed. Any study of real-world networks cannot neglect the specific nature of the system in the choice of the relevant measures and quantities to analyze. While we do not want to provide an extensive discussion of each specific network, it is worth turning our attention to the multipoint degree correlation functions and the clustering properties. These quantities, as will be apparent in the next chapters, characterize

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

Table 2.1 Numerical values characterizing the probability distributions described in this chapter, for various real-world networks.

Variable x	Sample	$\langle x \rangle$	$x_{\max}$	2σ	$\kappa = \langle x^2 \rangle / \langle x \rangle$
Degree k	WAN	9.7	318	41.4	53.8
	SCN	6.3	97	12.8	12.8
	Actors	80	3956	328	418
	AS	5.3	1351	70.8	242
In-degree k_{in}	WWW	24.1	378 875	843.2	7414.9
Betweenness b	WAN	6655	929 110	67 912	179 924
	SCN	37 087	3 098 700	235 584	411 208
	AS	14 531	9.3×10^6	439 566	3.3×10^6
Strength s	WAN	725 495	5.4×10^7	6×10^6	1.3×10^7
	SCN	3.6	91	9.4	9.8
	Actors	3.9	645	21	32

Networks included are the worldwide airport network (WAN), the scientific collaboration network (SCN, see `http://www-personal.umich.edu/~mejn/ netdata/`), the Actors' collaboration network (`http://www.imdb.com/`), the map of the Internet at the AS level, obtained by the DIMES project in May 2005 (`http://www.netdimes.org`), and the map of the WWW collected in 2003 by the WebBase project (`http://dbpubs.stanford.edu: 8091/~testbed/doc2/WebBase/`). The quantity $x_{\max}$ is the maximum value of the variable observed in the sample. The parameter $\kappa = \langle x^2 \rangle / \langle x \rangle$ and the mean square root deviation $\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$ estimate the level of fluctuations in the sample. The quantity 2σ is usually considered as a 95% confidence interval in measurements. It is also possible to appreciate that all heavy-tailed distributions show a maximum value of the variable $x_{\max} \gg \langle x \rangle$.

structural properties which are affecting the dynamical processes occurring on the network.

In Figure 2.11 we report the correlation and clustering spectra of several real-world networks. The figure clearly shows that most networks exhibit non-trivial spectra with a high level of variability. The correlation spectrum generally has one of the two well-defined mixing patterns presented in the previous chapter. This has led to the definition of the two broad classes of assortative and disassortative networks depending on whether their average nearest neighbors degree is an increasing or decreasing function of k (see Chapter 1). In most cases the clustering spectra also indicate significant variability of the cohesiveness properties as a function of the degree. These features have been extensively used in the attempt to formulate a conceptual and modeling understanding of the hierarchical and modular properties of networks (Ravasz *et al.*, 2002; Ravasz and Barabási, 2003). While many problems related to the structural ordering and hierarchical arrangement of large-scale networks remain open, the data of Figure 2.11 strengthen the evidence

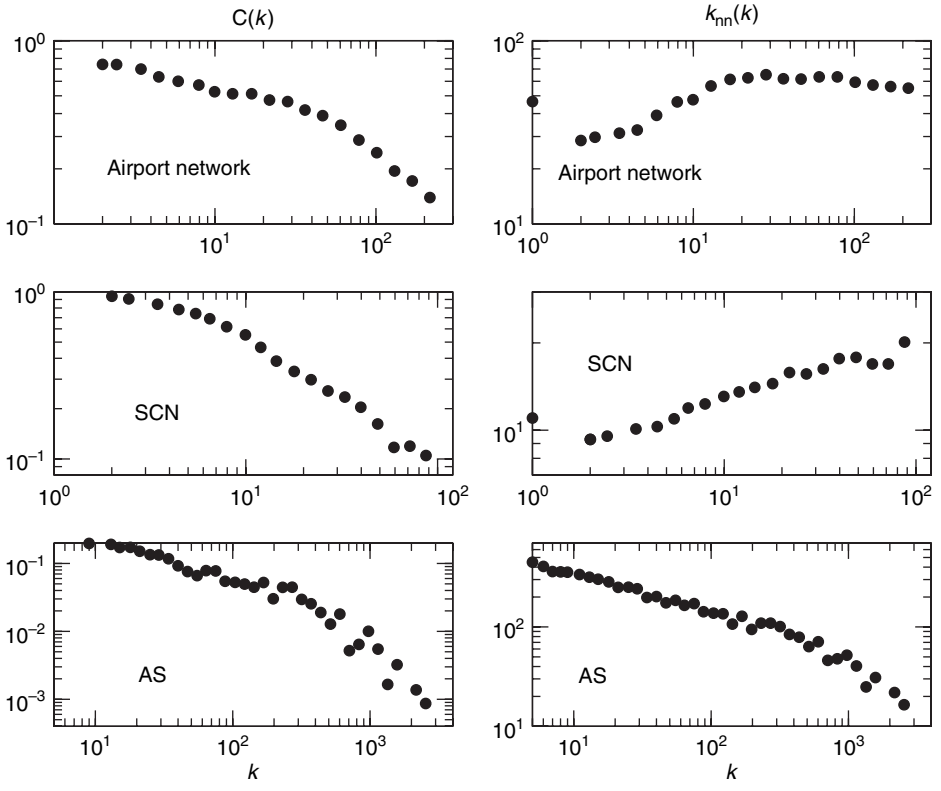


Fig. 2.11. Clustering spectrum (left column) and degree correlation spectrum (right column) for various real-world networks. The networks considered are, from top to bottom, the worldwide airport network (`http://www.iata.org`), the scientific collaboration network (SCN, see `http://www-personal.umich.edu/~mejn/netdata/`), and the map of the Internet at the AS level, obtained by the DIMES project (`http://www.netdimes.org`).

for considerable heterogeneity of networks and the lack of a typical scale in their description. At each degree value, different statistical correlation and clustering properties are found. These properties are highly variable, defining a *continuum* of hierarchical levels, non-trivially interconnected. In particular, there is no possibility of defining any degree range representing a *characteristic* hierarchical level.

A final note concerns the correlations usually present among weighted and topological properties. It appears in many cases that the weights of edges and the strengths of vertices are not trivially related to the connectivity properties of the graph. For instance, a good indication is provided by the dependence of the strength s_i on the degree k_i of vertices. In Figure 2.12 we show as an example the average

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

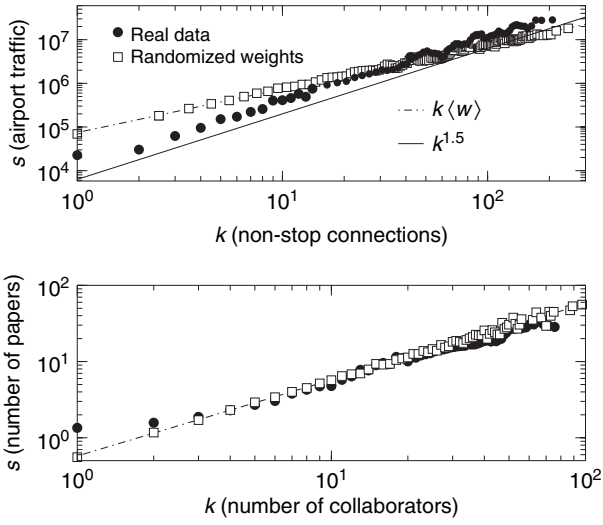


Fig. 2.12. Average strength $s(k)$ as a function of the degree k in the worldwide airport network (top, `http://www.iata.org`) and the scientific collaboration network (bottom, `http://www-personal.umich.edu/~mejn/netdata/`). In the airport network, the average strength represents the average number of passengers (traffic) handled by an airport with k direct connections. For the scientific collaboration network it gives the average number of papers of an author with k collaborators. We report the behavior obtained for both the real weighted networks and their randomized versions, which are generated by a random re-shuffling of the actual weights on the existing topology of the network.

strength $s(k)$ of vertices with degree k in the scientific collaboration network and in the worldwide airport network (Barrat *et al.*, 2004a). It is no surprise that in both cases the strength grows as a function of the degree. On the other hand, in the absence of correlations between the weight of edges and the degree of vertices, the strength of a vertex would be on average simply proportional to its degree. While this is the case for the scientific collaboration network, in general it is possible to observe a non-linear behavior of the form $s(k) \sim k^{\beta}$, with exponent $\beta \neq 1$, as in the worldwide airport network where $\beta \simeq 1.5$. Such behavior implies that the weights of edges belonging to highly connected vertices tend to have higher values than the ones corresponding to an uncorrelated assignment of weights. The analysis of the weighted quantities and the study of the correlations between weights and topology offer a complementary perspective on the structural organization of the network that might be undetected by quantities based only on topological information. In particular the presence of non-linear associations between topological and weighted quantities indicates a non-trivial interplay between the dynamical and structural properties of networks. Also, in this case, the evidence for heterogeneity,

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

variability and the absence of typical scale provide the signatures that generally characterize complex systems.

2.3 The complicated and the complex

In order to understand where complex networks can be found, and why they are defined as "complex," it is necessary to clarify the distinction between what is "complex" and what is merely complicated. This distinction is a critical one because the characteristic features and the behavior of complex systems in general differ significantly from those of merely complicated systems (Amaral and Barthélemy, 2003; Amaral and Ottino, 2004; Barabási, 2005).

The intricate appearance of large-scale graphs naturally evokes the idea of complicated systems in which a large number of components work together to perform a function. On the other hand, a computer or an airplane is also a very complicated system made by the assembly of millions of elements. Even a standard house is generally made of more than 20,000 different pieces, all of them performing different functions and tasks that follow a precise project. It is thus natural to ask what could distinguish complex systems from complicated ones. While a precise definition would certainly be very subjective, we can identify a few basic features typically characterizing complex systems.

A first point which generally characterizes complex systems is that they are emergent phenomena in the sense that they are the spontaneous outcome of the interactions among the many constituent units. In other words, complex systems are not engineered systems put in place according to a definite blueprint. Indeed, loosely speaking, complex systems consist of a large number of elements capable of interacting with each other and their environment in order to organize in specific emergent structures. In this perspective, another characteristic of complex systems is that decomposing the system and studying each subpart in isolation does not allow an understanding of the whole system and its dynamics, since the self-organization principles reside mainly in the collective and unsupervised dynamics of the many elements. It is easy to realize that the WWW, the Internet, and the airport network are all systems which grow in time by following complicated dynamical rules but without a global supervision or blueprint. The same can be said for many social and biological networks. All of these networks are self-organizing systems, which at the end of their evolution show an emergent architecture with unexpected properties and regularities. For example, in many cases complex systems can adapt to, evolve, and resist random removals of their components. It is clear that random removal of components of a computer will rapidly lead to malfunction or even to complete failure. In contrast, this is not the case for complex

systems as illustrated, for instance, by the Internet, for which the failure of a few routers will not prevent its normal functioning at the global level.

In simple terms, the Internet is a physical system that can be defined as a collection of independently administered computer networks, each one of them (providers, academic and governmental institutions, private companies, etc.) having its own administration, rules, and policies. There is no central authority overseeing the growth of this network-of-networks, where new connection lines (links) and computers (nodes) are being added on a daily basis. It is clear that the Internet is subject to technical constraints and, to a certain extent, engineered when we go down to the level of Local Area Networks and within single administered domains. On the other hand, the appearance and disappearance of computers and internet providers did not follow any global blueprint and was dictated by complicated dynamics in which economic, technical, and social reasons all played a role, along with the random variable of many individual decisions. In Figure 2.13 we show the monthly number of new and deleted vertices (in the Internet at the AS level) from November 1998 to November 2000 (Qian *et al.*, 2002). The plot clearly indicates that the overall growth of the Internet is the outcome of the net balance of a birth/death dynamic involving large fractions of the system. Despite this complex dynamic and the growth of vertices and edges, the statistical behavior in time of the various metrics characterizing internet graphs shows much smaller variations (Pastor-Satorras and Vespignani, 2004). The Internet has self-organized itself in a growing structure in which the complex statistical properties have reached a stationary state.

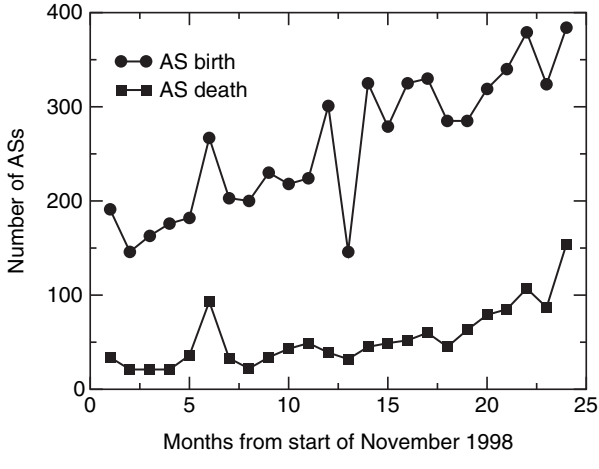


Fig. 2.13. Monthly number of new and dead autonomous systems in the period November 1998 to November 2000. Data from Qian *et al.* (2002).

https://doi.org/10.1017/CBO9780511791383.003 Published online by Cambridge University Press

Another main feature characterizing many complex systems is the presence of complications on all scales possible within the physical constraints of the system. In other words, when facing complex systems we are in the presence of structures whose fluctuations and heterogeneities extend and are repeated at all scales of the system. A typical example of this complexity is represented by fractal objects in which the same level of details and complications appears at whatever resolution we observe the object. Another clear example is provided by critical phenomena where infinitesimal localized perturbations can trigger macroscopic rearrangements across the entire system. Indeed, the long range correlations among the elements of the system may generate cascades of microscopic events disrupting the system at all scales. In the case of networks, the all-scale complication is statistically encoded in the heavy-tail distributions characterizing the structural properties. The larger the size of a system, the larger its heterogeneity and the variability of its properties. The absence of a typical degree or strength is, indeed, an indication that fluctuations extend over all the orders of magnitude allowed by the network size. Moreover, many other properties also change continuously over several orders of magnitude, with virtually infinite fluctuations. It is then impossible to define a typical scale in which an average description would be reliable.

As we have seen previously, heterogeneity and heavy-tail properties appear to be common characteristics of a large number of real-world networks, along with other complex topological features, such as hierarchies and communities. Analogously, most of these networks are dynamically evolving in time according to local processes that define the collective properties of the system. The evidence that a complex topology is the ubiquitous outcome of the evolution of networks can be hardly considered as incidental. It raises the question of the existence of some general organizing principles that might explain the emergence of this architecture in very different contexts. This consideration leads naturally to a shift of modeling focus to place more emphasis on the microscopic processes that govern the appearance and disappearance of vertices and links. As we will see in the next chapter, this approach has triggered the development of new classes of models which aim to predict the large-scale properties and behavior of the system from the dynamical interactions among its constituent units.