Classical random graphs

2

In this lecture we give an insight into the simplest and most studied random networks—classical random graphs.

2.1 Two classical models

In 1951–1952, applied mathematicians Ray J. Solomonoff and Anatol Rapoport published a series of papers in the *Bulletin of Mathematical Biophysics*, which did not attract much attention at the time [165, 166]. It is in these papers that the $G_{N,p}$ or Gilbert model, as it is now known, of a random graph (see Fig. 1.8) was introduced.[1] Later this basic model was rediscovered by mathematician E. N. Gilbert (1959). The notation $G_{N,p}$ indicates that this is a statistical ensemble of networks, G, with two fixed parameters: a given number of nodes N in each ensemble member and a given probability p that two nodes have an interconnecting link.

In the second half of the 1950s outstanding Paul Erdős and Alfréd Rényi introduced another random graph model and actually established random graph theory as a field of mathematics [87, 88]. The *Erdős–Rényi random graph* is a statistical ensemble whose members are all possible labelled graphs of given numbers of nodes N and links L, and all these members have equal statistical weight. This random graph is also called the $G_{N,L}$ model—a statistical ensemble G of graphs with two fixed parameters for each member of the ensemble: (i) a given number of nodes N and (ii) a given number of links L.[2] This is a special case of a general construction extensively exploited in the science of networks. The idea of this basic construction is to build an ensemble whose members are all possible graphs satisfying given restrictions, and all these members are realized with equal probability—uniformly randomly. One can say, this is the maximally random network that is possible under the given restrictions. One can also say that a network of this kind satisfies a given constraint but is otherwise random. Figure 2.1 shows an example—a small Erdős–Rényi network of 3 nodes and 1 link. Compare this small ensemble with that of the $G_{N=3,p}$ model in Fig. 1.8 and note a clear difference between these two ensembles. Remarkably, this difference turns out to be negligibly small in sufficiently large sparse networks.

Let us discuss this point in more detail. In simple terms, statistical mechanics is based on two kinds of ensembles—*canonical and grand canonical ensembles*. In all members of the first ensemble, the number of particles is equal and fixed. In the second, grand canonical ensemble, the chemical potential is fixed, and the numbers of particles are

[1] The terms Bernoulli or binomial random graph are also relevant. The term 'binomial' is explained by the binomial form of statistical weights in this model, see Fig. 1.8.

[2] Mathematicians usually use the notations $G_{n,p}$ and $G_{n,M}$ for the ensembles $G_{N,p}$ and $G_{N,L}$, respectively.

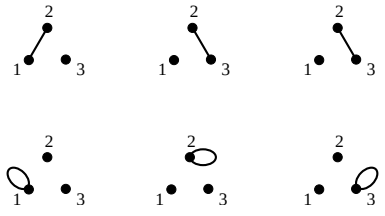

Fig. 2.1 The Erdős–Rényi random graph of 3 nodes and 1 link, which is $G_{N=3,L=1}$. All six graphs have equal statistical weight. Draw all configurations of the $G_{N=3,L=2}$ random graph.

different in different ensemble members. In statistical physics, these two ensembles—in our case two ways to define a random system—usually become equivalent as the number of particles approaches infinity. In random graphs links actually play the role of particles. So the $G_{N,L}$ and $G_{N,p}$ models correspond to the canonical and grand canonical ensembles, respectively. Suppose that $N \to \infty$, and the networks are sparse. Then one can show that these two random networks approach each other if $p = L/[N(N-1)/2]$.[3] In that sense, these two models are so close that they are called together 'classical random graphs' or even simply 'random graphs'.[4] Moreover, the term 'Erdős–Rényi model' sometimes refers to both of these ensembles. The reader may be surprised that in contrast to the Gilbert model, the Erdős–Rényi graph contains multiple connections and 1-loops, see Fig. 2.1. Then how can these models be equivalent? The explanation is that multiple connections and 1-loops in a large sparse Gilbert graph are not important—we will see that there are few of them.

Physicists know that analysis of the grand canonical ensembles is technically easier than for the canonical ones. In this respect the Gilbert model has an advantage. Let us obtain the degree distribution of this network using intuitive arguments. A node in this random graph can be connected to each of the other $N-1$ nodes with probability p. Then combinatorics immediately results in the binomial form of the probability that q of these $N-1$ links are present:

$$P(q) = C_{N-1}^{q} p^q (1-p)^{N-1-q}\,, \tag{2.1}$$

which is the degree distribution of the finite graph. Here $C_n^q = n!/[q!(n-q)!]$ is the binomial coefficient. One can obtain this exact formula strictly, averaging over the ensemble. The resulting mean degree of a node is $\langle q \rangle = p(N-1)$. When the network is large ($N \to \infty$) while $\langle q \rangle$ is finite (i.e. $p \to \text{const}/N$), the binomial distribution (2.1) approaches the Poisson one:

$$P(q) = e^{-\langle q \rangle} \langle q \rangle^q \, \frac{1}{q!}\,. \tag{2.2}$$

In this limit the Gilbert and Erdős–Rényi models are equivalent, and so this Poisson distribution is valid for all classical random graphs. The extremely rapid decay of the distribution is determined by the factorial denominator. We have mentioned that the degree distributions of practically all interesting real-world networks decay much slower.

Importantly, the degree distributions in these random graphs completely describe their architectures. Indeed, links in these networks are distributed independently. The only restriction is the fixed mean degree of a node. Therefore a node in a classical random graph 'does not know' the statistics of connections of any other node. In that sense, even connected nodes are statistically independent. Random graphs of this kind are called *uncorrelated networks*, which is one of the basic notions in this field.[5] We will consider general uncorrelated networks in detail in the following lectures.

[3] More precisely, the statistical characteristics of typical members of these two ensembles converge. In particular, a large majority of members in the Gilbert ensemble have only relatively small deviations in the numbers of links from $L = pN(N-1)/2$. Note that $N(N-1)/2$ is the total number of links in the complete graph of N nodes.

[4] One can even say that classical random graphs are maximally random networks with a given mean degree $\langle q \rangle$ of a node.

[5] In particular, the absence of correlation means factorization of the joint distributions of random variables. For example, let $P(q, q')$ be the probability that one node has degree q and another one has degree q'. Then in an uncorrelated network, $P(q, q') = f(q)f(q')$, where the function $f(q)$ is completely determined by the degree distribution, see Lecture 5.

2.2 Loops in classical random graphs

As was mentioned, the large sparse classical random graphs have few loops. What does this mean? Let us first discuss small loops. It is easy to find the clustering coefficient and so the total number of loops of length 3 in a large classical random graph. Let the graph be of N vertices with a mean degree $\langle q\rangle$. Recall that the clustering coefficient of a node is the probability that two nearest neighbours of the node are themselves nearest neighbours. In our case this probability is $\langle q\rangle/(N-1) \cong \langle q\rangle/N$. So the clustering is

$$C = \overline{C} = \frac{\langle q\rangle}{N}\,. \tag{2.3}$$

Note the equality $C = \overline{C}$ and the independence of the clustering coefficient of a node from its degree. This is, of course, the case for any uncorrelated network. So the clustering coefficient for an infinite sparse classical random graph approaches zero. That is, clustering in these networks is only a finite size-effect. Let us compare the clustering coefficients of real-world networks with those of the classical random graphs with the same numbers of nodes and links. For example, these days (2009) the map of routers in the Internet contains about half a million nodes—routers—or even more. The Internet grows exponentially, and so this number changes rapidly. The mean number of connections of a node in this network is about 10. The clustering coefficient is about 0.1. For a classical random graph of the same size, with the same mean degree, formula (2.3) gives $C \sim 10^{-5}$, which is four orders of magnitude lower than in the Internet! In the 1990s, when exploration of real-world networks was in its early stages, many empirical studies highlighted this tremendous difference, thus showing how far classical random graph models stray from reality.

Formula (2.3) allows us to find the total number of triangles in a classical random graph:[6]

$$\mathcal{N}_3 = \langle q\rangle^3/6\,. \tag{2.4}$$

That is, the number of triangles in a sparse classical random graph does not depend on its size. This number is finite even if these graphs are infinite. Similarly, one can find the number of loops of an arbitrary length L. This number also does not depend on L: $\mathcal{N}_L \cong \langle q\rangle^L/(2L)$ if L is smaller than the diameter of the network, which we expect to be of the order of $\ln N$. Thus any finite neighbourhood of a node in these random graphs almost certainly does not contain loops. In that sense, these networks are *locally tree-like*, which is a standard term. We will use this convenient feature extensively. On the other hand, there may be plenty of long loops of length exceeding $\ln N$; namely, $\ln \mathcal{N}_L \sim N$ if $L \gg \ln N$. Obviously, these loops cannot spoil the local tree-like character.

One may observe another important object, *cliques*. A clique is a fully connected subgraph. For example, a loop of length 3 provides us with a 3-clique. Since there are so few loops in these networks, one can see that

[6] This expression is obtained in the following way. $\mathcal{N}_3 = \frac{1}{3}C(\#\text{ of triples})$, see Section 1.9. The total number of connected triples of nodes is $N(\langle q^2\rangle - \langle q\rangle)/2$. Finally, we use the fact that for the Poisson distribution the second and first moments are related:

$$\langle q^2\rangle - \langle q\rangle = \langle q\rangle^2.$$

3-cliques are the maximum possible. The bigger cliques in large sparse classical random graphs are almost entirely absent.

There is a similar random graph construction, even more simple than the Erdős–Rényi model. This is the *random regular graph*. All N vertices of this graph have equal degrees. The construction is similar to the Erdős–Rényi model. The random regular graph is a statistical ensemble of all possible graphs with N vertices of degree q, where all the members are realized with equal probability. In other words, this is a maximally random regular graph.[7] This graph also has loops of various lengths, including loops of length 1, as in the Erdős–Rényi graph. Their number is $\mathcal{N}_L \cong (q-1)^L/(2L)$, and so these networks also have a locally tree-like structure. In that sense, an infinite random regular graph approaches the Bethe lattice with the same node degree.

[7] Note that a random regular graph does not belong to the category of classical random graphs.

2.3 Diameter of classical random graphs

Now we can immediately exploit the local tree-like character of simple random networks. We essentially repeat our derivation of the diameter of a Bethe lattice from Section 1.4. The only difference being that, in a random tree one should substitute a branching b in the derivation for the mean branching $\overline{b}$ of a link. So the number z_n of the n-th nearest neighbours of a node grows as $\overline{b}^{n-1}$. This exponential growth guarantees that the number of nodes S_n which are not farther than distance n from a given node is of the same order, $\overline{b}^n$. Of course, this tree ansatz fails when S_n is already close to the size of a network, that is when n is close to the diameter. It turns out, however, that if we ignore the presence of loops even in this range of n and estimate $\overline{b}^{\overline{\ell}} \sim N$, we get a non-essential error in the limit of large N. So the resulting mean internode distance is $\overline{\ell} \cong \ln N/\ln \overline{b}$ at large N. This result is valid for all uncorrelated networks.

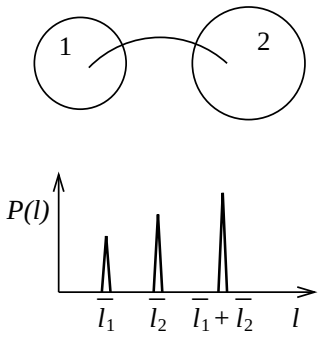

Fig. 2.2 A network consisting of two large modules—small worlds—interconnected by a single link has a wide distribution $\mathcal{P}(\ell)$ of internode distances. Indeed, let the mean separations between nodes inside modules 1 and 2 be $\overline{\ell}_1$ and $\overline{\ell}_2$, respectively. Then two nodes in different modules are separated by approximately $\overline{\ell}_1+\overline{\ell}_2$ links. So the distribution has three peaks: at $\overline{\ell}_1$, $\overline{\ell}_2$, and $\overline{\ell}_1+\overline{\ell}_2$. If, however, the number of connections between the modules is any finite (nonzero) fraction of the total number of links, then all three peaks merge into a single narrow peak.

The exponential growth, $z_n \sim \overline{b}^n$, has another remarkable consequence. Since $z_{\overline{\ell}} \sim N$, a finite fraction of nodes in these networks are at distance $\overline{\ell}$ from each other. That is, the width of distribution of internode distances $\mathcal{P}(\ell)$ in these networks is finite, even if these nets are infinitely large. In other words, nodes in infinite networks are almost surely mutually equidistant. The width $\delta\ell$ of the distribution $\mathcal{P}(\ell)$ is finite only in specific network models. However, the relatively narrow internode distance distributions, $\delta\ell \ll \overline{\ell}$, are typical for a wide range of small worlds.[8] (For a rare counterexample, a specific small world without this property, see Fig. 2.2.) This is why the mean internode distance in large networks is close to the diameter (the maximum separation between nodes). Note that mutual equidistance is realized only in very large networks, where the mean separation of nodes is much greater than 1. Even the largest real-world nets (in the WWW, for example, $\overline{\ell} \sim 10$, which is not that great) still have rather wide internode distance distributions.

[8] For comparison, one can easily estimate that for d-dimensional lattices, $\delta\ell/\overline{\ell} \sim 1-2^{-d}$, and so this distribution is wide if d is finite.

For a random q-regular graph, branching equals $q-1$, and so readily

$\overline{\ell} \cong \ln N / \ln(q-1)$. To obtain the diameter of a classical random graph, we must first find the average branching of its links. For this purpose, we will use a remarkably general relation for networks. Consider an arbitrary undirected graph of N nodes—a single realization, not an ensemble. This graph may include bare nodes and be simple or multiple. Let the numbers of nodes of degree $q = 0, 1, 2, \ldots$, be $N(q)$. So $N = \sum_q N(q)$, and the frequency of occurrence of nodes of degree q is the ratio $N(q)/N$. The total degree of the graph (double the number of links) equals $\sum_q qN(q) = N\langle q \rangle$. Clearly, $N(q)$ nodes of degree q attract $qN(q)$ stubs ('halves of links') of total number $N\langle q \rangle$. Then the frequency with which end nodes of a randomly chosen link have degree q, is $qN(q)/(N\langle q \rangle)$. In other words, select a link at random, choose at random one of its end nodes, then the probability that it has q connections equals $qN(q)/(N\langle q \rangle)$.

For random networks, this important statement is formulated as follows. *In a random network with a degree distribution $P(q)$, the degree distribution of an end node of a randomly chosen link is $qP(q)/\langle q \rangle$*, see Fig. 2.3. Therefore the connections of end nodes of links are organized in a different way from those of randomly chosen nodes. We will use this fact extensively in the following sections. In particular we have $\langle q^2 \rangle / \langle q \rangle$ for the average degree of an end of a randomly chosen link, which is greater than the mean degree of a node $\langle q \rangle$ in the network. The mean branching is the average value of the number of connections of an end of a randomly chosen link, minus one—the link itself. So it is $\overline{b} = (\langle q^2 \rangle / \langle q \rangle) - 1$. Recall that for the Poisson distribution, $\langle q^2 \rangle - \langle q \rangle = \langle q \rangle^2$. Therefore in classical random graphs, an average branching is equal to a mean degree, $\overline{b} = \langle q \rangle$. Finally we arrive at the famous formula

$$\overline{\ell} \cong \frac{\ln N}{\ln \langle q \rangle} \tag{2.5}$$

for the mean separation between nodes in a classical random graph and for its diameter. Interestingly, this formula provides a reasonably good estimate for mean internode distance in numerous real-world networks, even if they differ strongly from classical random graphs. For example, for the map of routers, $N \sim 2 \times 10^5$ and $\langle q \rangle \sim 3$, formula (2.5) gives $\overline{\ell} \approx 11$, which is close to an empirical value.

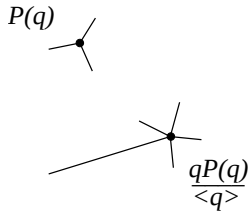

Fig. 2.3 End nodes of a randomly chosen link in a network have different statistics of connections from the degree distribution of this network.

2.4 The birth of a giant component

In our discussion of diameters we missed one thing. In general, a graph may consist of different, separate parts—clusters. All nodes inside each of these parts—*connected components*—have connecting paths running within them.[9] Deriving formula (2.5) for a diameter we actually supposed that a random graph consists of a single connected component. This is certainly not true when a mean degree $\langle q \rangle$ approaches zero and the random graph consists of bare nodes. Already in 1951 Ray Solomonoff and Anatol Rapoport had discovered that when $\langle q \rangle$ exceeds 1,

[9] We assume here that a network is undirected.

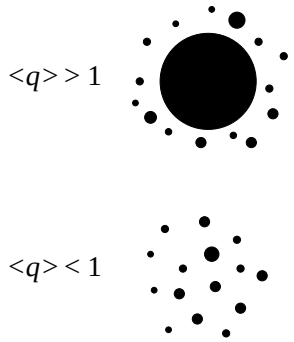

Fig. 2.4 The organization of a classical random graph. The filled circles show connected components. When mean degree $\langle q \rangle$ is above 1, the graph contains a giant connected component, which is absent if $\langle q \rangle < 1$.

[10] In physics, if there is no jump of an order parameter (in our case, the relative size of a giant component) at the critical point, then it is a continuous phase transition.

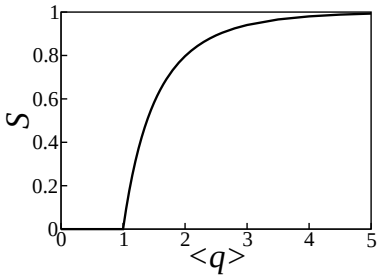

Fig. 2.5 The relative size of a giant connected component in a classical random graph versus the mean degree of its nodes. Near the birth point, $S \cong 2(\langle q \rangle - 1)$.

infinite classical random graphs contain a large connected component including a finite fraction of nodes. This is a so-called *giant connected component*. Physicists, working in the field of condensed matter, are familiar with a close analogy—a *percolation cluster* [169]. Remove at random a fraction of nodes from an infinite conducting lattice, so that a fraction p of nodes are retained. Then below some critical value of p the lattice is split into a set of finite disconnected clusters, and the system is isolating. On the other hand, a current can flow—'percolate'—from one border of the lattice to another if p is above this critical concentration—a percolation threshold p_c. The current flows on an infinite percolation cluster of connected nodes, which is quite similar to a giant connected component. The situation in classical random graphs is as follows, see Fig. 2.4. When a classical random graph has many links, $\langle q \rangle > 1$, it consists of a single giant connected component and, also, numerous 'finite connected components'. On the other hand, if $\langle q \rangle < 1$, the giant connected component is absent and there are only plenty of finite connected components.

This qualitative picture is generic for networks. The general properties of a network are primarily determined by whether or not a giant connected component is present. So the first question about any network should be about the presence of this component. Strictly speaking, a giant connected component is well defined for infinite networks. What about finite nets? We must inspect the dependence on network size, N. If the number of nodes in a connected component grows proportionally to N, then it is treated as 'giant'. In contrast, finite connected components practically do not grow with N.

Solomonoff and Rapoport found that a giant connected component in classical random graphs emerges exactly when a mean degree $\langle q \rangle$ surpasses 1. This happens without a jump, see Fig. 2.5. In that sense the birth of a giant connected component is a continuous phase transition, where $\langle q \rangle = 1$ is the critical point.[10] This is the main structural transition in a network, where network characteristics dramatically change. Note that all these changes take place in the regime of a sparse network, in which the number of connections is low compared to the maximal possible number. Furthermore, Fig. 2.5 demonstrates that a giant connected component approaches the size of a classical random graph still being in a sparse regime. In particular, the relative size of a giant connected component is already above 99% at $\langle q \rangle = 5$. Thus main qualitative changes in the architecture of networks are in the narrow region $\langle q \rangle \ll N$. Remarkably, most studied real-world networks are sparse.

Returning to formula (2.5) for the diameter, we may conclude that it certainly fails close to the birth point of a giant connected component. We will discuss this special point in detail in the following lectures. On the other hand, when a giant component contains a reasonably large fraction of nodes in a classical random graph, deviations from relation (2.5) are negligible.

2.5 Finite components

After Erdős and Rényi, mathematicians spent a few decades studying the statistics and structure of connected components in the classical random graphs. In simple terms, the overall picture looks to be as follows. Let the size N of a classical random graph tend to infinity. We already know about a giant component. What about 'finite' ones? Note these inverted commas. The point is that this standard term may be confusing. It turns out that some 'finite connected components' still grow with N but extremely slowly, much slower than a giant connected component (whose size is proportional to N). Let us first stay away from the critical point $\langle q\rangle = 1$ in either of the two phases: in the 'normal phase', without a giant component, or in the phase with a giant component. Then the biggest 'finite' connected component, the second biggest, the third, the i-th biggest, with i being any finite number, all of these components have sizes of the order of $\ln N$. The total number of components grows with N, and so most of the components are really finite as $N \to \infty$. In any case, the sizes of all of these components are much smaller than that of the giant one.

Let us now move to the the critical point, where a giant component is still absent. At this point, the biggest connected component, the second, the third, the i-th biggest, with i being any finite number, all of these components are of the order of $N^{2/3}$.[11] It is important that this size, $N^{2/3}$, is much smaller than N but much bigger than $\ln N$.

The statistics of connected components are remarkably different at the critical point and away from it. In the normal phase and in the phase with a giant component, these distributions have a rapid exponential decay. In contrast, at the critical point, the size distribution of components $\mathcal{P}(s)$ decays slowly, as a power law:

$$\mathcal{P}(s) \sim s^{-5/2}\,. \tag{2.6}$$

The sum $\sum_s s\mathcal{P}(s) \sim \sum_s s s^{-5/2}$ converges. Thus the mean size $\langle s\rangle$ of a finite connected component is finite at any value of $\langle q\rangle$, including the birth point of a giant connected component. Figure 2.6 (see the lower curve) shows that the finite components are mostly very small—one or two nodes.

We see from Fig. 2.6 that the dependence of $\langle s\rangle$ on $\langle q\rangle$ has only a not particularly impressive cusp at the critical point. On the other hand, another, related average demonstrates a real critical singularity. Choose at random a node in a network. What is the probability that this node is in a connected component of s nodes? Clearly, this probability $\mathcal{P}'(s)$ is proportional to the product $s\mathcal{P}(s)$. At the birth point of a giant connected component, $\mathcal{P}'(s) \sim s^{-3/2}$, so it has an infinite first moment. Thus the average size $\langle s\rangle'$ of a finite connected component to which a (randomly chosen) node belongs diverges at the critical point. The upper curve in Fig. 2.6 shows the full dependence of $\langle s\rangle'$ on $\langle q\rangle$. In the critical region, $\langle s\rangle' \cong 1/|\langle q\rangle - 1|$, which strongly resembles the famous Curie–Weiss law for susceptibility in physics.[12]

[11] Rigorously speaking, this is valid within a so-called scaling window, where the deviation from the critical point, $|\langle q\rangle - 1|$, is less than, say, $N^{-1/3}$, Bollobás (1984) [39].

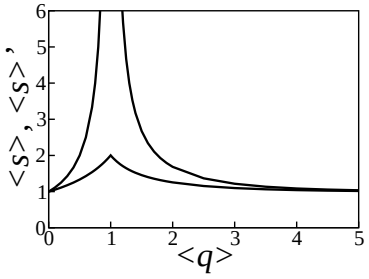

Fig. 2.6 The lower curve: the mean size (number of nodes) $\langle s\rangle$ of a finite component in a classical random graph versus the mean degree of its nodes. The upper curve: the mean size $\langle s\rangle'$ of a finite component to which a randomly chosen node belongs as a function of mean degree. $\langle s\rangle'$ diverges at the critical point as $1/|\langle q\rangle - 1|$, while $\langle s\rangle$ is finite.

[12] The Curie–Weiss law works for interacting systems with a second-order phase transition provided that the mean-field theory is valid. For example, near a critical temperature in ferromagnets the magnetization

$$M(T, H = 0) \propto \sqrt{T_c - T}$$

in a zero applied field. A susceptibility $\chi(T, H)$ characterizes the response of the magnetization to a small addition of an applied field: $H \to H + \delta H$. That is, $\chi(T, H) \equiv \partial M(T, H)/\partial H$. According to the Curie–Weiss law, the susceptibility at a zero applied field is

$$\chi(T, H = 0) \propto 1/|T - T_c|.$$

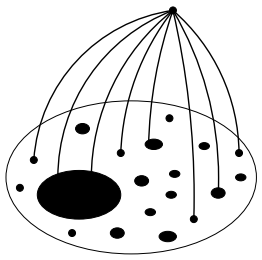

Fig. 2.7 An additional node attached to randomly chosen nodes in a network increases the giant connected component and so plays the role of an external field. The black spots are connected components in a network.

One may even say that the average $\langle s\rangle'$ plays the role of 'susceptibility' in graph theory and in percolation problems. Figure 2.7 explains this analogy. Let us add a node to an arbitrary network and link it to a number n of randomly chosen nodes. Suppose first that this number is a finite fraction of the total number of nodes N. Then, thanks to the connections of the 'external' node, we obtain a giant connected component even if this component is absent in the original network. Clearly, this specific attachment plays the role of an applied field which increases the size $M = SN$ of a giant connected component—similarly to the magnetic field changing magnetization in a ferromagnet. We intentionally use the same notation for the size of a giant connected component as for magnetization in physics. Similarly to magnetization (the ferromagnetic order parameter) the size of a giant connected component plays the role of an order parameter in networks. The 'susceptibility' may now be introduced as an increase in the size of a giant component in response to a small addition to the number of 'external' links $n \to n + \delta n$. The 'zero field susceptibility' is then $\partial M(n)/\partial n|_{n=0}$. One can easily see that this is exactly the average size $\langle s\rangle'$, since the attachments are to randomly selected nodes.

To conclude this lecture, we emphasize that the qualitative picture described here for classical random graphs is of a surprisingly general nature going even beyond graph theory and the science of networks. In this respect, this is a zero model of a random network, but it is not a toy model.

Small and large worlds

3

3.1 The world of Paul Erdős

The fantastic productivity of Paul Erdős and his travelling life (Erdős had a reputation as a mathematician-pilgrim) resulted, in particular, in an incredible number of coauthors. Over 500 mathematicians had the privilege of being coauthors with Erdős. So Erdős plays the role of a hub in the network of collaborations between mathematicians. Nodes in this net are the mathematicians (authors), and undirected connections are coauthorships in at least one publication. We have already mentioned that this is actually a one-mode projection of a bipartite collaboration network. In total, the network of collaborations between mathematicians includes about 337 000 authors, connected by 496 000 links.

It is a great honour to be Erdős' coauthor, but also it is a honour to be a coauthor of Erdős' coauthor (in reality, this honour is shared by a lot of mathematicians), and so on. These grades of the 'closeness' to Erdős are classified by *Erdős numbers*. The Erdős number of Erdős is 0, that of his coauthors is 1, of coauthors of Erdős' coauthors is 2, *etc.* In short, the Erdős number is the shortest path length between a mathematician and Erdős. Figure 3.1 shows the numbers of mathematicians with various Erdős numbers. One can easily see that this is actually the distribution of internode distances in this network. Note that this world of Erdős has a crucially different structure from the Erdős–Rényi model. Indeed, a rapidly decreasing Poisson degree distribution does not admit hubs with 500 in graphs with such a small number of connections. By using the formula for the Poisson distribution, the reader can find the probability that in a classical random graph with the same numbers of nodes and links as in the mathematics collaboration graph, at least one node has 500 connections. This probability is of the order of 10^{-934}.

One can see from Fig. 3.1 that the average distance of a mathematician from Erdős is about 5—only five steps/coauthorships from greatness! For comparison, formula (2.5) gives the mean internode distance 12 for a classical random graph with the same number of nodes and links. Thus this network is even more compact than the classical random graph, though the difference is not dramatic. Note that the distribution in this figure is not narrow. The diameter equals 15, which is three times bigger than the mean distance between nodes. This is in contrast to the mutual equidistance of nodes in infinite small worlds, which we discussed. This is not that surprising, since the mean internode distance is not much greater than 1.

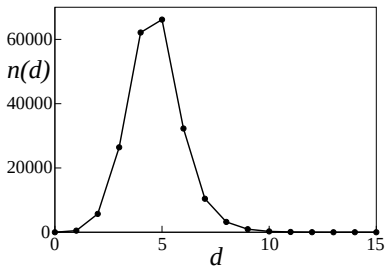

Fig. 3.1 The number of mathematicians with Erdős number d in 2001. This plot was made by using data from the web page of the Erdős Number Project, http://www.oakland.edu/grossman/erdoshp.html.

One should note that the described small world of Paul Erdős is a very typical collaboration network. All these networks are small worlds. Now, what about the largest directed net—the WWW?

3.2 Diameter of the Web

In 1999 Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabási, physicists from the University of Notre Dame, measured what they called 'the diameter' of the WWW [5]. In fact, this was the average length of, what is important, the directed path between two pages of the WWW. Albert, Jeong, and Barabási succeeded in collecting data from a small part of the WWW, namely the nd.edu domain of the University of Notre Dame. Then, how could such limited data be used for obtaining the mean internode distance of the large Web?

The researchers used the following approach. Although the complete map of the nd.edu domain contained only 325 729 pages and 1 469 680 hyperlinks, these numbers were large enough to expect that the organization of connections in the domain is close to that in the entire Web. The architecture of the nd.edu was approximately reproduced in a set of even smaller model networks of different sizes. These small networks could easily be generated and studied numerically. The only features of the real WWW they reproduced were the distributions of incoming and outgoing connections of nodes, but this was sufficient.[1] It was easy to measure the mean shortest-directed-path length for each of these mini-Webs. The resulting size dependence $0.35 + 0.89 \ln N$ was extrapolated to the estimated size of the Web in 1999, $N \approx 800\,000\,000$, which gave the mean internode distance 19. In the spirit of our calculations of $\overline{\ell}$ in classical random graphs, we can easily estimate the average length of the shortest directed path in a directed network: $\overline{\ell}_d \approx \ln N / \ln\langle q_{\rm o}\rangle$. Here $\langle q_{\rm o}\rangle$ is a mean out-degree, that is the mean number of outgoing links of a node.[2] Substituting the above numbers to this estimate we obtain $\overline{\ell}_d \approx 14$ for the network of 800 000 000 nodes, which is not very far from the result of Albert, Jeong, and Barabási.

[1] As a model of the WWW these authors used a maximally random network with a given distribution of incoming and outgoing links of nodes.

[2] By definition, the in-degree $q_{\rm i}$ of a node is the number of its incoming links, the out-degree $q_{\rm o}$ is the number of outgoing connections.

Later direct measurements made by a large group of computer scientists from AltaVista, IBM, and Compaq resulted in a close value, 16, for the part of the WWW (about 200 000 000 pages) observed by the search engine AltaVista [44]. In other words, a web navigator would have to make only about 20 'clicks' on average to reach a web page, following hyperlinks. Thus, in respect of internode distances, the large Web is a very small object—a small world.

3.3 Small-world networks

The collaboration networks and the WWW are only two important examples among an infinity of real networks showing the small-world feature. We also mentioned that the great majority of real-world networks have strong clustering. The values 0.1, 0.2, 0.3, etc. for a clustering coef-

ficient are very typical. This combination—the small-world phenomenon and the strong clustering—seemed to be completely incomprehensible in the late 1990s, when the only reference model widely used for comparison was a classical random graph. Unfortunately, classical random graphs have very weak clustering. Ten years later, it is clear that there exist plenty of easy ways to get a strongly clustered small world. The nature of strong clustering (and, in general, of numerous loops) is not considered to be a serious problem now. Nonetheless, it was the desire to understand the high values of clustering coefficient in the empirical data that inspired sociologist Duncan Watts and applied mathematician Steven Strogatz (1998) to propose an original model, interconnecting themes of networks and lattices [181]. This very popular model has significantly influenced the development of the field.

The basic idea of Watts and Strogatz was as follows. Suppose somebody, who knows only lattices and the Erdős–Rényi model, wants to construct a network with the small-world feature and numerous triangles. The classical random graphs demonstrate the small-world phenomenon but have few triangles. On the other hand, many lattices have numerous triangles (if their unit cells include triangles of bonds) but the lattices have no small-world feature. Then let us combine a lattice with many triangles and a classical random graph with the small-world feature. The combination of the 'geographically short-range' connections and long-range shortcuts has clear roots in real life: find examples from communications, sociology, etc. Technically, Watts and Strogatz connected pairs of randomly chosen nodes in a lattice by links—'shortcuts', see Fig. 3.2. Thanks to these long-range shortcuts, even nodes, widely separated geographically within the mother lattice, have a chance to become nearest neighbours. Clearly, the shortcuts make the resulting networks more compact than the original lattice. Watts and Strogatz called networks of this kind *small-world networks*. This term is used even if a mother lattice is disordered, or it has no triangles. The clustering here is of secondary importance. Indeed, what is so special about loops of length 3? Loops of length 4, 5, etc. are no less important.

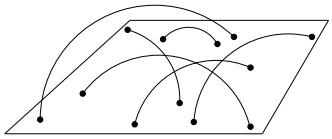

Fig. 3.2 The idea of a small-world network. Long-range shortcuts—links connecting randomly chosen nodes—are added to a d-dimensional lattice. Together with the nodes and connections within the mother lattice (which are not shown) these links form a small-world network.

Figure 3.3 (a) shows the original Watts–Strogatz network where randomly chosen links of a mother lattice were rewired to randomly selected nodes. The rewiring produces the same effect as the added long-range shortcuts, Fig. 3.3 (b). Figure 3.3 (b) also explains why we have superposition of a lattice and a classical random graph. Indeed, if we let the connections within the mother lattice be absent, then the shortcuts and the nodes form a classical random graph. The resulting small-world network is already 'a complex one'.

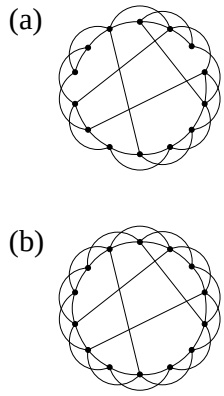

Fig. 3.3 The original Watts–Strogatz model (a) with rewiring of links [181], and its variation (b) with addition of shortcuts (b) [132]. Notice numerous 3-loops within the mother lattice.

The problem is: what does happen when the number of shortcuts increases? It is easy to see that even a single long-range shortcut sharply diminishes the average internode distance $\overline{\ell}$. For example, the first shortcut added to a one-dimensional lattice (as in Fig. 3.3) may reduce $\overline{\ell}$ by half! So, the influence of shortcuts on the shortest path lengths in a small-world network is dramatic. Let p be the relative number of shortcuts in a small-world network (with respect to the overall number of con-

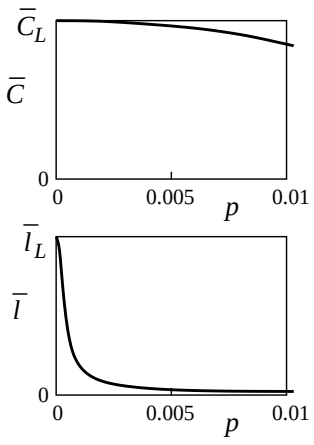

Fig. 3.4 A sketch of two typical dependences for small-world networks. Mean clustering $\overline{C}$ versus the relative number of shortcuts p and average internode distance $\overline{\ell}$ versus p, in a small-world network with a fixed finite number of nodes. $\overline{C}_L$ and $\overline{\ell}_L$ are the corresponding values for the mother lattice without shortcuts.

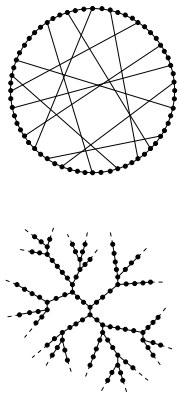

Fig. 3.5 This small-world network has locally a structure of a tree.

nections. Watts and Strogatz observed that even at a very low p, where the clustering is still nearly the same as in the mother lattice, the mean internode distance is tremendously reduced. Figure 3.4 shows schematically these typical dependences of clustering and the average distance between nodes on p in a finite small-world network. Thus even at small (but non-zero) p the network shows the small-world phenomenon.

One should stress that in these networks there is no sharp transition between two regimes: from large to small worlds. Rather, a smooth crossover is realized as the number of shortcuts grows. In particular, in the limit of the infinite number of nodes N, we obtain a 'large world' if the number of shortcuts N_s is finite, and a small world if the relative number of shortcuts p is finite. Let us estimate the average distance between nodes, $\overline{\ell}(p)$. We assume that $N \to \infty$ and p is finite, that is, the number $N_s \sim pN$ of shortcuts is large enough—the small-world regime. In this case the shortcuts determine the global architecture of the network, and the network resembles a classical random graph with N_s links and of the order of N_s 'supernodes'. These 'supernodes' are regions of the mother lattice, which contain neighbouring ends of shortcuts. To this graph, we can apply a classical formula (2.5), that is $\overline{\ell} \sim \ln N$, with two changes. In our estimation we ignore constant factors and for the sake of simplicity suppose that the network is based on a one-dimensional mother lattice—a chain. In the resulting network, the average distance between ends of different shortcuts is $\overline{d} \sim 1/p \sim N/N_s$. Then, (i) substitute N for N_s; (ii) take into account that the shortest paths in the network pass the mother lattice between ends of different shortcuts, so multiply the classical expression by $\overline{d}$. This gives the result:

$$\overline{\ell}(p) \sim \overline{d} \ln N_s \sim (1/p) \ln(Np)\,. \tag{3.1}$$

Thus the diameter of this small-world network is about $N_s/\ln(N_s) \gg 1$ times smaller than that of the mother lattice.

We can obtain relation (3.1) in a slightly different way. The point is that often, large small-world networks may be treated as locally tree-like. Look at Fig. 3.5 which shows a small-world net based on a simple chain. One can see that as $N \to \infty$, any finite environment of an arbitrary node is tree-like. The figure also shows this local structure. The reader can apply the arguments, which we used in Section 2.3, to this tree-like structure, estimate the mean branching, and rederive result (3.1).

3.4 Equilibrium versus growing trees

Up to now we have discussed equilibrium random networks although most real networks are non-equilibrium. The difference between networks from these two classes is sometimes great. For demonstration purposes we compare here two important random trees—equilibrium and growing. These two models are selected to be as close to each other as possible the only difference being: one tree is equilibrium and the second—growing.

Let us first introduce recursive networks. A recursive network grows in the following way. Add a new node and attach it to a number of already existing nodes. Then repeat again, and again, and again. Figure 3.6 (a) explains this growth process (the bubble denotes an existing network). The nodes for attachment are selected by rules defined in specific models. Notice that in this growth, new links cannot emerge between already existing nodes. This is, of course, a great simplification, which makes these networks easy for analysis. Most growth models that we will discuss are of this type. Let us now suppose that in the recursive process, each new node has only one link, and the growth starts from a single node, see Fig. 3.6 (b). Then this network is a recursive tree (has no loops). Here the attachment is also a sequential birth of new nodes by existing ones. One should specify how a node for attachment is selected. The simplest rule is to choose it from among the existing nodes with equal probability, that is uniformly at random. This gives *a random recursive tree* as it is known in graph theory. This construction provides us with a maximally random tree with a given number of nodes under only one restriction: this tree is obtained by sequential addition of labelled nodes—the causality restriction. The random recursive tree is used as a reference model for growing networks. We will consider this important random graph in more detail in Section 7.1. Note that its degree distribution significantly differs from that of classical random graphs. For large recursive trees, the degree distribution approaches the exponential form: $P(q) = 2^{-q}$.

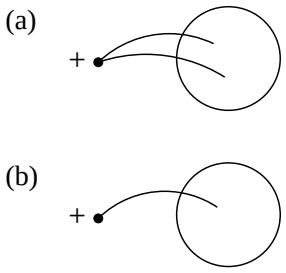

Fig. 3.6 How recursive networks grow. (a) The growth of a loopy recursive network. Each new node is attached to a few chosen existing nodes. Note that new connections are only between an added node and already existing nodes. (b) The growth of a recursive tree. (We assume that the growth starts from a single node.)

It is an easy exercise to find that the mean internode distance in the tree is $\overline{\ell}(N) \cong 2\ln N$, if the number N of nodes is large.[3] Thus this random tree is a small world. Let us compare the random recursive tree with an equilibrium tree. By definition, this is a maximally random tree with a given number of nodes N. Compare with the definition of the Erdős–Rényi model. In other terms, this statistical ensemble contains all possible trees of N nodes taken with equal statistical weight. Figure 3.7 shows the members of both of these ensembles for a few small values of N.

[3] Compare the total lengths of the shortest paths between all pairs of nodes in this network at 'times' N and $N+1$. These are $\overline{\ell}(N)N(N-1)/2$ and $\overline{\ell}(N+1)(N+1)N/2$, respectively. The difference due to the new attached node is $1+[\overline{\ell}(N)+1](N-1)$. Together these three terms give an equation for $\overline{\ell}(N)$, which can easily be solved.

One can see that in the random recursive tree some configurations

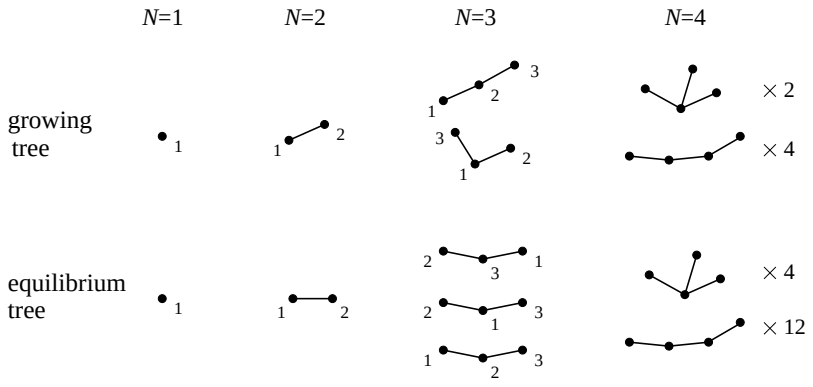

Fig. 3.7 The random recursive tree versus the equilibrium tree. The upper and lower rows shows the members of these two statistical ensembles. Each of them has a unit statistical weight. At $N = 4$ we indicate the number of configurations which differ from each other only by labels.

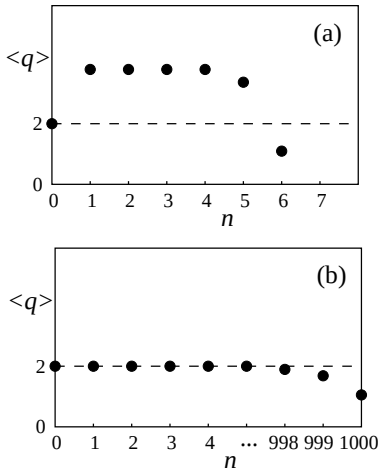

Fig. 3.8 The average degree of the n-th nearest neighbour of a randomly chosen node as a function of n for two networks: for a particular small world—a recursive tree (a), and for an equilibrium tree—a fractal (b).

are absent. For example, there is no configuration 2—3—1 (equivalent to 1—3—2). Already at $N = 4$, the equilibrium tree has a larger mean internode distance than that of the growing one (compare weights of more and less compact configurations in the figure). As N increases, this difference becomes vast. One can show that the diameter of the equilibrium random tree scales as $\sqrt{N}$ at large N. Therefore it is not a small world but rather a random fractal with Hausdorff dimension 2. Interestingly, the degree distribution of this network is close to classical random graphs. In the former case it is simply factorial: $P(q) \propto 1/(q-1)!$. Nonetheless, one of these networks is a small world and the second—a 'large' one.

In Fig. 3.8, the readers can see how different are the neighbourhoods of a randomly chosen node in a random recursive tree (a) and in an equilibrium tree (b). The figure shows the dependences of the average degree of the n-th nearest neighbour of a node on n. Clearly, $\langle q \rangle(n = 0)$ coincides with the average degree $\langle q \rangle$ of nodes in a network. This average degree approaches 2 in any infinite tree. A significant difference is for $n \geq 1$. In this range, for the small-world phenomenon the mean branching must be greater than 1. So in the recursive random trees $\langle q \rangle(n) > 2$. In fact, the dependence, schematically presented in Fig. 3.8 (a), is valid for any small world. Of course, $\langle q \rangle$ in loopy networks will be greater than 2. As for the equilibrium trees, the mean branching is close to 1 even in a very far neighbourhood of a node.

3.5 Giant connected component at birth is fractal

This 'fractal appearance' of the equilibrium trees has remarkable consequences even for loopy networks with the small-world property. The point is that we can find equilibrium trees within loopy small worlds. For an arbitrary connected graph, one can construct a tree which spans over the entire graph. This subgraph, the *spanning tree*, by definition, consists of all nodes of a given graph and of some of its links. Clearly, if a given graph is not tree, then it has a number of different spanning trees. In principle, this number may be very large. *A uniform (or random) spanning tree* of a given graph is a statistical ensemble whose members are all spanning trees of this graph, taken with equal probability. In simple terms, this is the maximally random spanning tree of a given graph. One may suppose that this random tree is equilibrium. Furthermore, it turns out that the uniform spanning trees of classical random graphs are fractals of Hausdorff dimension 2 [99]. In the preceding section we explained how to make a small world from a 'large one' by using long-range shortcuts. Contrastingly, the uniform spanning trees enable us to make a large world from a small one.

Even though the classical random graphs are loopy networks, they contain plenty of trees. Note that classical random graphs are only locally tree-like. So their finite components are (almost surely) trees,

but a giant connected component is loopy—it has long loops. Recall our estimation of the diameter of a classical random graph in Section 2.3. We emphasize here that the result $\overline{\ell} \approx \ln N / \ln\langle q\rangle$ was obtained for an equilibrium, locally tree-like network which, however, has many long loops. We mentioned that this estimate is valid only sufficiently far from the birth point of a giant connected component. Suppose that we are approaching the birth point $\langle q\rangle = 1$ from above, and so a giant component increasingly resembles finite ones. So, it has fewer and fewer loops. In the limit $\langle q\rangle \to 1$, the giant component becomes an equilibrium tree. One can show that it has a fractal architecture, and its Hausdorff dimension equals 2, as is natural for random trees of this kind. Thus in the limit $\langle q\rangle \to 1$, the classical random graphs are fractals!

We can easily reach this fractal state by removing at random nodes or links in a graph, until a giant connected component disappears. In a large but still finite random graph, one can also make the following operation. Choose at random a node within a giant connected component. Then remove this node, and all of its neighbours closer than distance n from it. When n is sufficiently small, we will get a hole in the original giant connected component. With growing n, this hole increases, and at some critical value, it splits a giant component into a set of finite ones, see Fig. 3.9. Close to this point, the vanishing giant component again turns out to be a two-dimensional fractal [161]. In that sense, the remote part of a classical random graph with a removed 'centre' has a fractal structure.

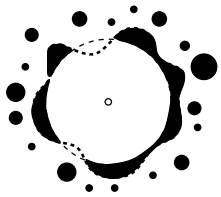

Fig. 3.9 The remnant of a random graph after the removal of a large environment of a node (open dot). A giant connected component on the verge of vanishing has a fractal architecture.

3.6 Dimensionality of a brush

We introduced small worlds as objects with an infinite fractal or Hausdorff dimension. Recall how the Hausdorff dimension was defined in terms of the size dependence of the mean internode distance: $\overline{\ell}(N) \sim N^{1/d}$, that is $d = \ln N / \ln \overline{\ell}$. One should note that this definition of the dimensionality of a system is not unique. Let us discuss an alternative way using a random walk on a network. A random walk here plays the role of a useful instrument which allows us to characterize the structure of a network. Consider a particle which at each time step, with equal probability, moves from a node to one of its nearest neighbours. After t steps of a random walk on a regular d-dimensional lattice, the particle will typically be found at a distance of the order of $\sqrt{t}$ from a starting point. So the 'area' of the region where the particle may be found at time t is about $t^{d/2}$. This is simply the number of nodes in the hyperball of radius $\sqrt{t}$. Then the probability that the particle will be found at the starting point after t steps is $p_0(t) \sim t^{-d/2}$. This can be used as another definition for the dimensionality. Remarkably, for some objects, this definition results in a dimension which differs from the Hausdorff one. So it is natural to introduce a special number, a *spectral dimension* d_s, defined by the relation $p_0(t) \sim t^{-d_s/2}$.[4] Actually, d_s shows that random walks in a given network are organized similarly to those on d_s-dimensional lattices.

[4] We will explain why it is called 'spectral' in future lectures.

A difference between spectral and Hausdorff dimensions may be illustrated by combs and brushes, see examples in Fig. 1.2 (c) and (d). Let a brush be based on a large d-dimensional lattice with long hairs—linear chains—growing from some of the nodes of a lattice. The Hausdorff dimension of this brush is $d+1$. As for its spectral dimension, it was found that $d_s = 1+d/2$ when d is not higher than 4, and $d_s = 3$ for $d \geq 4$ [107]. One can even use a network with the small-world feature and attach long linear chains to its nodes, and again, the spectral dimension is only 3.

From the Internet to cellular nets

4

A wide range of real-world networks from diverse areas have surprisingly similar architectures. In this lecture we discuss and compare structural properties of a few basic man-made and natural networks. Before starting, we emphasize the principal difference between two global networks—the Internet and the World Wide Web (WWW).

(i) The Internet is a global technological network of connected computers through which users can access data and programs from other computers. Links in this network can be wired or wireless.

(ii) The WWW is a global information network, an array of web documents (files of various formats), connected by hyperlinks. The hyperlinks are mutual references in web documents.

Even though most impressive, the WWW is only one of many Internet applications. Another one, for example, is email.

4.1 Levels of the Internet

The first distributed computer network ARPANET was constructed at the end of 1969.[1] Originally ARPANET linked only four nodes: the University of California at Los Angeles, Stanford Research Institute, the University of California at Santa Barbara, and the University of Utah. This pioneering, US national net afterwards grew into the Internet by interconnecting with other networks [53]. The key idea was to build the Internet as a federation of interconnected autonomous (independently managed), peer networks of very different types and architectures. The routing of data packets within the peer networks is maintained by their individual internal rules—protocols, while routing between these networks is performed by common internetwork routing protocols.[2] Thus, based on 'hierarchical routing', Internet technology was substantially formed by the middle of the 1980s. By 1991, the Internet included 700 000 host computers, and, in principle, approached the modern form.

The specific organization of the Internet as a network of numerous autonomous networks, without a central authority is apparently optimal and inevitable. 'The Internet is the first computational artifact that was not designed by one economic agent, but emerged from the distributed, uncoordinated, spontaneous interaction (and selfish pursuits) of many.

[1] The US Defense Advanced Research Projects Agency (DARPA) played a great role in the history of networking and initiated the 'Internetting' program in 1972. The organization of ARPA (1958, renamed DARPA in 1972) was initiated in response to the launch of Sputnik on October 4, 1957 by the Soviet Union.

[2] A protocol is an algorithm, a standard, a set of formal instructions and rules, see in more detail in Lecture 12.

[3] From the paper 'On a network creation game' by Fabrikant *et al.* (2003) [90].

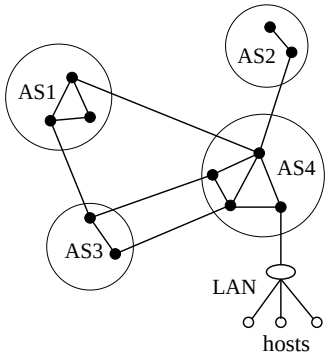

Fig. 4.1 Multilayer organization of the Internet. Open dots show host computers of users, LAN is a local area network, filled dots are routers, and ASs are autonomous systems.

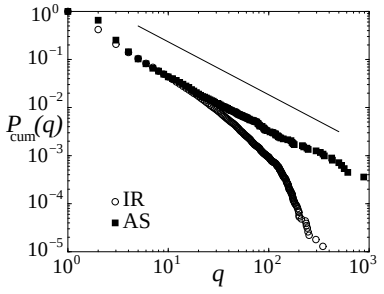

Fig. 4.2 The cumulative degree distributions of the two networks: AS level map measured in October/November 1999 and the Internet Router (IR) map of May 2001. Adapted from the paper of Vázquez, Pastor-Satorras, and Vespignani [177].

[4] The cumulative degree distribution is defined as $P_{\rm cum}(q) = \sum_{q' \geq q} P(q')$. If $P(q) \sim q^{-\gamma}$, then $P_{\rm cum}(q) \sim q^{1-\gamma}$. Researchers use these distributions to diminish inevitably strong fluctuations in the region of large degrees. Another method to smooth fluctuations is binning—accumulation or averaging data within some degree intervals.

Today's Internet consists of over 12,000 subnetworks ("autonomous systems"), of different sizes, engaged in various, and varying over time, degrees of competition and collaboration.'[3] This long quotation touches upon several key aspects of the evolving architecture of the Internet. Let us consider this multilayer architecture in more detail. The Internet includes hosts (computers of users), servers (computers or programs providing a network service), and routers, arranging the traffic. The total number of hosts (including handheld devices) in the Internet was about 570 million in July 2008 and will probably reach 3 billion (3 000 000 000) by 2011 [1]. Figure 4.1 shows schematically the multilayer architecture of the Internet. The complete structure of the Internet including all host computers has never been investigated. Routers with their undirected interconnections form the router level network in the Internet. The autonomously administered subnetworks (autonomous systems, AS) in this network are the nodes of the second, AS-level graph. Routing between autonomous systems is maintained by the common Border Gate Protocol. (A gateway is a system (software or a device) that joins two networks together.) In the AS graph, two autonomous systems are connected by a single undirected link if at least two of their routers are directly connected. Because of the exponentially rapid growth, it is even hard to estimate the present sizes of these two networks. Very approximately, the Internet contained 40–70 thousand autonomous systems in 2009. The number of routers was higher, say, by 15–25 times.

The routers in the Internet have geographical locations. It turns out, however, that usually it is hard to find their precise coordinates. As for the autonomous systems, some of them contain routers distributed all over the world, and any geographical mapping is impossible. This is why the majority of studies have to ignore the geographical factor in Internet architecture. The statistics of connections in the Internet remained unstudied until the end of 1990s. In 1999, three scientists, Faloutsos, Faloutsos, and Faloutsos, analysed the partial maps of the AS and router networks in the Internet and discovered that both of them had not classical, but scale-free architectures. They found that the degree distributions of these networks could be approximately fitted by power laws [91]. The networks that they investigated were small, so in Fig. 4.2, we reproduce empirical cumulative degree distributions obtained later for larger AS and router maps by Pastor-Satorras, Vázquez, and Vespignani [148,177].[4] Both these empirical cumulative distributions were fitted by power-law dependencies with exponent 1.1, and therefore the degree distribution exponent $\gamma = 2.1$.

The empirical distributions in Fig. 4.2 are very typical [10]. Note that the 'quality' of these 'power laws' is rather poor, especially for the router network, which is also a very typical situation. The AS network consisted of about 11 000 autonomous systems with, on average, 4.2 connections for each AS. The network of routers had about 230 000 nodes and the router's mean degree $\langle q \rangle \approx 2.8$. For smaller networks and higher values of exponent γ, empirical curves usually resemble even less a power law. Importantly, the Internet maps which empirical researchers analyse, are

very approximate and always incomplete. These maps were obtained by sending packet probes over the network from one or a few sources. Unfortunately, this technique may seriously distort the empirical degree distributions. Nodes with few connections have a higher chance of escaping the probe, and the measured degree distributions are more skewed than the real ones [151, 56]. Nonetheless, despite all these difficulties and restrictions, the empirical data for the Internet show remarkably skewed degree distributions. The biggest autonomous systems have a few thousand connections to other ASs.

Both the networks have small diameters. The average internode distance in the AS network is only 3–4 hops. In the router network this distance was found to be about 9. These were data obtained in 1999–2001 for Internet maps. The Internet grows exponentially with time, and both the AS and router networks evolve rapidly. It was theoretically suggested that these and many other networks become more densely connected with time. In other words, the growth is 'accelerated' in the sense that the number of links in a network grows more rapidly than the number of nodes [78]. Empirical observations have confirmed that this is indeed the case [120]. The growing average degree results in increasing clustering. Furthermore this densification leads to shrinking diameters of the networks. The networks exponentially grow but their diameters are constant or even decrease with time.[5]

One can show that the heavy-tailed degree distributions make strong clustering inevitable. The measured values were $\overline{C} \approx 0.03$ and $\overline{C} \approx 0.3$ for mean clustering in the router and AS networks, respectively [177]. Moreover, in contrast to classical random graphs, the clustering of a node in the router and AS networks was observed to depend strongly on the degree of this node, see Fig. 4.3. We explained that in this situation, the clustering coefficient C can essentially differ from $\overline{C}$. In reality, C in these networks is much smaller than $\overline{C}$.

Pastor-Satorras, Vázquez, and Vespignani also measured a quantity, which became one of the standard structural characteristics of complex networks. They investigated the average degree $\overline{q}_{\rm nn}(q)$ of the nearest neighbour of a node with q connections. This quantity is independent of q in the uniformly random, that is, uncorrelated networks; see the next lecture. In these uncorrelated networks, nodes 'know' nothing about the degrees of their neighbours. In the Internet, this is certainly not the case. Figure 4.4 demonstrates the sharp difference between Internet networks and uniform ones. Therefore, it is not only the skewed degree distribution that distinguishes these complex networks from classical random graphs. Subsequent studies have shown that the Internet is not an exception. On the contrary, very few real-world networks have no structural correlations.

[5] Interestingly, with time, many routers and autonomous systems disappear.

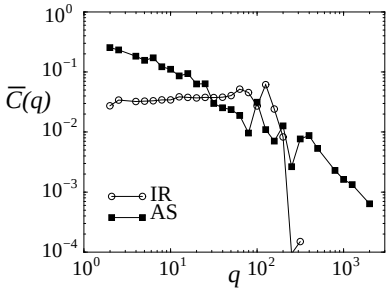

Fig. 4.3 The mean clustering of a node versus the degree of this node for the same two networks as above. Adapted from the paper of Vázquez, Pastor-Satorras, and Vespignani [177].

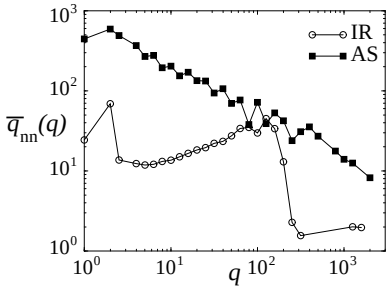

Fig. 4.4 The average degree of the nearest neighbour of a node of degree q for the same two networks as above. Adapted from the paper of Vázquez, Pastor-Satorras, and Vespignani [177].

4.2 The WWW

In essence, the WWW is simply a system for automated retrieving of information in the form of electronic documents (files of various formats). The original idea was to organize links between documents in a way, convenient for users. In the WWW, the linking is based on *hypertext*. The hypertext contains highlighted parts which cover the link to other documents. Clicking on highlighted text causes the fulfilment of the underlying link to the corresponding document and downloads this file to the user's computer. In 1989, Tim Berners-Lee proposed a hypertext system for CERN, the European Laboratory for Particle Physics in Geneva, Switzerland [25, 65]. In 1990, he wrote a program 'WorldWideWeb', which is a web browser editor, 'a program which provides access to the hypertext world', and the first web page was placed on the first web server in CERN. Hypertext documents in the WWW are written by using special programming language, Hypertext Markup Language (HTML) (look at any web page by using the option *View the Page Source* of your browser). The functioning of the WWW is based on the Hypertext Transfer Protocol (HTTP), which enables the flow and processing of web documents and requests in the Web. Thus, in 1990, the four required components—(i) the protocol of the Web, HTTP, (ii) the language of the Web, HTML (which are the two main standards of the Web), (iii) a web browser, and (iv) web servers—were created. The WWW was born.

On the 25 July 2008 the official Google blog announced: the Google index 'hit a milestone: 1 trillion (as in 1 000 000 000 000) unique URLs (Uniform Resource Locator) on the web at once!'[6] This number—a trillion pages in the Google index, however impressive, does not allow us to estimate the size of the exponentially growing WWW. The difficulty is that only a very small fraction of pages are accessible by search engines. These public pages, sufficiently static to be scanned, form the so-called *Surface Web*. The huge part of the WWW, (*the Deep Web*), are electronic documents in databases and archives with restricted public access and, also, rapidly varying pages (time tables, web calendars, and so on) with all their hyperlinks. A clear border between the Surface and Deep Webs is absent, and it is barely possible to reliably evaluate their sizes. What empirical researchers can study are sufficiently large pieces of the WWW. The analysis of these parts allows us to understand the architecture of the WWW.

[6] See http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

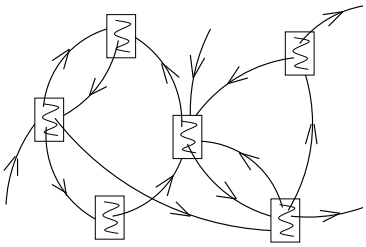

Fig. 4.5 Connections in the WWW. Notice reciprocal hyperlinks.

Figure 4.5 shows very schematically how hyperlinks connect web documents. Compare this schematic view of directed connections in the WWW with a more simple picture of connections in the network of citations in scientific papers, see Fig. 1.7. Notice reciprocal hyperlinks in the Web graph. Two Web documents can cite each other while scientific papers cannot. It turned out that the number of reciprocal hyperlinks in the WWW is surprisingly large, up to 60% of all connections [85, 138]. The abundance of reciprocal hyperlinks in the WWW becomes more clear if we look at the scheme of connections of a typical home page,

see Fig. 4.6. These figures also show longer loops. The directedness of the hyperlinks make makes these loops more diverse than in undirected networks. For example, there are six different loops of length 3 if we also take into account reciprocal links. In principle, the definition of clustering must be modified to account for this diversity. Usually, however, the clustering of the WWW network is measured ignoring the directedness of connections, which typically gives $C \approx 0.1$ and $\overline{C} \approx 0.3$.[7]

The directedness of hyperlinks determines a more rich and interesting organization of finite and giant components in the WWW than in undirected networks. According to a standard definition, a giant component obtained ignoring the directedness of connections is *a giant weakly connected component*. For a general directed network, this giant component is organized as is shown in Fig. 4.7. There is *a giant strongly connected component*, which consists of the nodes mutually reachable by directed paths. The nodes of *a giant out-component* are reachable from the strongly connected component by directed paths. *A giant in-component* contains all the nodes from which the strongly connected component is reachable. By this definition, the giant strongly connected component is the intersection of the giant in- and out-components. The remaining part of the giant weakly connected component is the mess of so-called tendrils. Apparently, the presence of a giant strongly connected component is vitally important for the function of the WWW. In 1999, Broder and coauthors measured the sizes of these components using a map of a sufficiently large part of the WWW (about 200 million pages) [44]. They found that the giant strongly connected component contained about 30% of the pages in the weakly connected component, while the tendrils contain about 25%. The average length of the shortest direct path between two web pages in these measurements was about 16. Remarkably, the maximum separation of nodes observed in this network was very large, namely, about 1000 clicks!

Similarly to the Internet, the density of connections in the WWW grows with time. Broder and coauthors first made their measurements in May 1999 and found that the average in- and out-degrees of a node are equal at 7.22. When they repeated the measurements in October 1999, the average in- and out-degrees were already 7.85.

The scale-free in- and out-degree distributions of the WWW, $P_{\rm i}(q_{\rm i}) \sim q_{\rm i}^{-\gamma_{\rm i}}$ and $P_{\rm o}(q_{\rm o}) \sim q_{\rm o}^{-\gamma_{\rm o}}$, respectively, were observed in 1999 by Albert, Jeong, and Barabási, who studied a relatively small nd.edu domain. Broder and coauthors (2000) obtained the degree distributions for a much larger network [44]. They fitted the empirical in- and out-degree distributions by power-law dependencies, with exponents $\gamma_{\rm i} = 2.1$ and $\gamma_{\rm o} = 2.7$, respectively. The power law for $P_{\rm i}(q_{\rm i})$ was observed in a wide range of in-degrees (about three orders of magnitude), and so the obtained value 2.1 is reliable. This is not the case for the out-degree distribution. The range of out-degrees in the WWW is much narrower than that of in-degrees, so there is even a doubt that a power law for out-degrees exists at all [71].[8] In any case, in respect of at least the in-degree distribution, the WWW is a scale-free network.

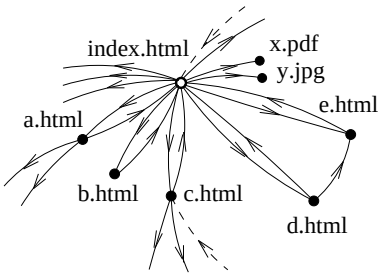

Fig. 4.6 Structure of a typical home page. The dashed arrows show hyperlinks coming from external web documents.

[7] These values are for the same nd.edu domain of the WWW, which we discussed in the previous lecture [5]. Surprisingly, taking into account the hyperlink directedness leads to even smaller numbers of triangles than one could expect [31].

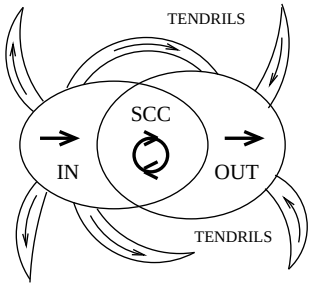

Fig. 4.7 Organization of a giant weakly connected component in a directed network. SCC, IN, and OUT are the giant strongly connected component and the giant in- and out-components. For more detail, see [80]. Compare with Broder *et al.* [44].

[8] It is impossible to put on a page, say, one million links to other web documents.

4.3 Cellular networks

The Internet and WWW networks are sufficiently large to allow reliable statistical analysis of their architectures. Even on the AS level, the Internet network has many thousand nodes. In this section we will touch upon a few really small cellular and genetic networks, whose specific structures could be revealed despite their tiny size.[9] In 2000, Jeong, Tombor, Albert, Oltvai, and Barabási made a thorough study of the networks of the metabolic reactions of 43 simple organisms belonging to different domains of life [106]. All these networks were very small, from 200 to 800 nodes. The network of metabolic reactions, in essence, is a typical chemical reaction graph. Its nodes are molecular compounds participating in various metabolic reactions as educts or products. The same compound can be educt in one reaction and product in another. If two compounds participate in some reaction as an educt and a product, connect them by a directed link going from educt to product. Thus Jeong and coauthors had 43 sparse directed graphs to analyse. The remarkable conclusion was that the in- and out-degree distributions of all 43 networks are approximately scale-free with exponent equal to 2.2. The definite conclusion for such small networks was possible due to the low value, 2.2, of the degree distribution exponent. The lower this exponent, the wider the range of degrees to observe a power law. This scale-free architecture has a direct consequence on the distributions of metabolic reaction fluxes in these networks. These distributions were also found to be skewed [86, 8].

[9] For detailed discussion of complex network concepts in cellular biology, see the review of Barabási and Oltvai [15].

The second example of a cellular network, which we touch upon, is the network of physical protein interactions of the yeast *Saccharomyces cerevisiae* [105]. This network of 1870 nodes and 2240 undirected links is bigger than the metabolic networks that we have discussed in this section, nonetheless, the statistical data are less conclusive. The nodes of this network are different proteins and the undirected links are physical interactions (direct contacts) between them. It was impossible to fit the empirical degree distribution by a power-law function. Instead, the distribution was fitted by a power law with an exponential cut-off. The power-law exponent can be roughly estimated as 2.5, which, we believe, is already too large to reliably observe scale-free distributions in networks of this size. Still, the general architecture of this network is rather similar to that of, say, the AS-level Internet network. In particular, the empirical dependencies $\overline{C}(q)$ and $\overline{q}_{\rm nn}(q)$ look similar to those in Figs. 4.3 and 4.4 [122, 123].

In both these examples, the networks are clearly defined, and a graph for empirical study can be obtained in a quite strict way. For many other systems, an underlying network structure is not that obvious. Using genetic networks as an illustration [156], let us demonstrate how the network structure can be unveiled.

Without going into detail, *genome* is a large set of interacting genes which encode the genetic information of an organism. A given living organism is characterized by a set of features, the so-called gene expres-

sions. The expression of each gene is quantitatively described by its level—*an expression level*. The term 'interaction between genes' is used here in the sense that genes function not independently, but in cooperation: the expression of a gene depends on other genes. Consequently, expression levels are not independent. These correlations between expression levels of genes are treated as the result of the cooperative function of genes, that is of their 'interaction'.

The procedure for obtaining correlations between gene expression levels is routine in genetic research. Suppose that the expression levels of N genes in a genome are e_i, $i = 1, \ldots, N$. The original ('wild-type') cell culture is 'perturbed' M times. For example, the culture is exposed to radiation and mutant strains with distorted genomes are produced. The full set of gene expression levels, $\{e_i^{(s)}\}$, $s = 1, \ldots, M$, of these mutant strains is measured (see Fig. 4.8). The correlation c_{ij} between the expression levels of two genes, i and j, may be easily obtained by averaging over all the mutant strains:[10]

$$c_{ij} = \langle e_i e_j \rangle - \langle e_i \rangle \langle e_j \rangle \,, \tag{4.1}$$

where the average $\langle\ \rangle = M^{-1} \sum_s$. Correlations between genes i and j are absent when $c_{ij} = 0$. The resulting numbers c_{ij} are the elements of a large $N \times N$ matrix which shows how different genes interact with each other. However, most of the matrix elements are small and inessential, and this matrix, in its original form, provides superfluous information. For unveiling the basic structure of gene interactions, one must take into account only the important matrix elements. So, genes i and j are believed to be interacting to (be connected) if only the element c_{ij} is greater than some threshold value. This value is chosen such that the resulting network is sufficiently sparse, and the structure of essential pairwise interactions is clearly visible.

The same approach may be applied to many other situations. In general, network constructions of such a kind use two basic restrictions:

(i) Take into account only 'important' pairwise interactions or correlations between 'nodes'.

(ii) Ignore the difference between the magnitudes of these interactions.

In principle, one can present the complete information in the form of a *weighted network*, where each link has its real number value c_{ik}, taken, for example, from eqn (4.1). Usually, however, these weighted networks are too informative for a direct analysis.

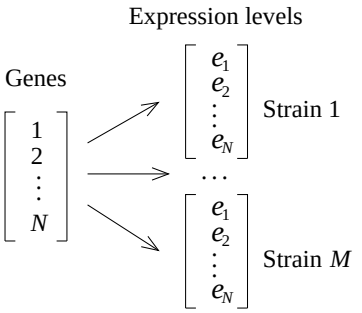

Fig. 4.8 Genes in a genome (N genes) are expressed in terms of characteristic features—traits—of the organism (N gene expression levels). These features are not independent. Exposure to radiation (M times) leads to mutant strains with M sets of gene expression levels. The statistical analysis of these sets allows one to find correlations between different gene expressions and to unveil the cooperative function of the corresponding genes.

[10] This is a simplified form of the expression. Moreover, instead of the expression levels e_i, the logarithms of the ratios of expression levels of mutant strains and those of the wild-type cell culture are actually used.

4.4 Co-occurrence networks

In many systems allowing network representation, groups of elements co-occur in various associations. That is, a given set of elements each have something in common, or they participate together in some action, or, they present in the same list, and so on. If these associations are pairwise, then we have a simple network (nodes connected by ordinary

single links). On the other hand, if the associations are more complicated, then we arrive at multi-partite graphs and hypergraphs. Let us touch upon these more interesting constructions.

The standard example of these networks is the bipartite network of directors sitting on many boards. We will consider this in the next lecture. Here we mention another typical example, namely the human disease network constructed by Goh and coauthors in 2007 [96]. The network was based on a long list of human genetic disorders and all known disease genes in the human genome. The disorders and disease genes are nodes of two types. If mutations in a gene are involved in some disorder, connect these two nodes by an undirected link. Since the same gene may be implicated in various disorders, the network has a rich structure of connections. The researchers hope that analysing the global organization and statistics of connections in this bipartite network will allow them to find as yet unknown relations between genetic disorders and disease genes. The point is that the complete system of disorders and disease genes is too large to uncover all the relations of this kind without using network representation.

In principle, the human disease network is a quite typical bipartite graph. Zlatić, Ghoshal, and Caldarelli (2009) studied more exotic networks [186] based on file-sharing databases Flickr and CiteULike. In these databases, users upload photos (Flickr) or put links to scientific papers (CiteULike) with short text descriptions. All registered users can supply these photos and papers by keywords—descriptive tags. The resulting network has three sorts of nodes: (i) users, (ii) photos (papers), and (iii) tags. When a user uploads a photo with a tag or assigns a tag to a photo uploaded by another user, he creates a new hyperedge interconnecting three nodes: this user, the photo (paper), and the tag; see Fig. 4.9. So these networks are tripartite hypergraphs, in which every hyperedge interconnects three nodes of different types. In these tagged social networks, there are three 'hyperedge distributions', for each kind of node. The researchers found that these distributions differ significantly from each other. The degree distribution for photos (papers) decays most rapidly of the three, the distribution for users decays the most slowly (apparently, there exist 'crazy taggers'), and the degree distribution for tags demonstrates an intermediate rate of decay. Thus, interestingly, there are no really 'superpopular' photos and papers either in Flickr or in CiteULike. Instead, there are plenty of people interested in tagging!

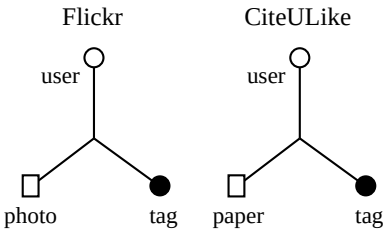

Fig. 4.9 Hyperedges in Flickr and CiteULike.

Uncorrelated networks

5

Most real-world networks are very far from the classical random graph models. In this lecture we show that these models can be greatly improved.

5.1 The configuration model

By the late 1970s, the theory of classical random graphs was well developed, and mathematicians started to search for more general network constructions. In 1978, Edward A. Bender and E. Rodney Canfield published a paper entitled 'The asymptotic number of labelled graphs with given degree sequences' [22], where they described random networks with essentially richer architectures than the Erdős–Rényi graph. Béla Bollobás strictly formulated this generalization of the Erdős–Rényi model in his paper 'A probabilistic proof of an asymptotic formula for the number of labelled random graphs' (1980) and named it the *configuration model* [38]. This generalization turned out to be a major step toward real networks in the post-Erdős epoch.

Before introducing the configuration model, we have to recall the idea of classical random graphs. A classical random graph is the maximally random network that is possible for a given mean degree of a node, $\langle q \rangle$. The Erdős–Rényi model is one of the two versions of classical random graphs: a maximally random network under two restrictions: (i) the total number of nodes N is fixed and (ii) the total number of links is fixed. These constraints result in the Poisson form of the degree distribution which differs from the degree distributions of real-world networks. To make a step towards real networks, one should be able to construct a network with, at least, a real degree distribution $P(q)$, and not only a real mean degree. The idea was to build the maximally random network for a given degree distribution. The configuration model provides a way (more precisely, one of the ways) to achieve this goal by directly generalizing the Erdős–Rényi construction. In graph theory, the term 'sequence of degrees' usually means the set of numbers $N(q)$ of nodes of degree q in a graph, $\sum_q N(q) = N$ [128]. Let this sequence be given. The configuration model is a statistical ensemble whose members are all possible labelled graphs, each with the same given sequence of degrees. All these members are realized with equal probability. We have explained that this corresponds to the the maximum possible randomness—uniform randomness.

The same can be done in the following way. Consider a set of N

[1] Note that the term 'at random' without further specification usually means 'uniformly at random'. To build a particular realization of this network, make the following: (i) create the full set of nodes with a given sequence of stub bunches; (ii) from all these stubs, choose at random a pair of stubs and join them together; (iii) from the rest of the stubs, choose at random a pair of stubs and join them together, and so on until no stubs remain.

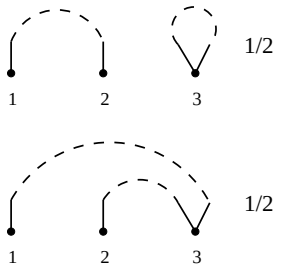

Fig. 5.1 A random network ensemble in the configuration model. In this example, $N(1)=2$ and $N(2)=1$. The two members of the ensemble have equal probability of realization, 1/2.

nodes: $N(1)$ nodes of degree 1, $N(2)$ nodes of degree 2, and so on; supply each of the nodes by q stubs of links; choose stubs in pairs at random and join each pair together into a link (see Fig. 5.1).[1] As a result we have a network with a given sequence of degrees, but otherwise random. Clearly, if every node has the same number of connections, then this network is reduced to a random regular graph. On the other hand, if the sequence of degrees is drawn from a Poisson distribution, then (in the infinite network limit) we get a classical random graph.

The heterogeneity of these networks is completely determined by degree distributions; correlations are absent in contrast to real-world networks. Fortunately, the majority of phenomena in complex networks can be explained qualitatively based only on the form of degree distribution, without accounting for correlations.

5.2 Hidden variables

The second way to build an uncorrelated network with an arbitrary degree distribution was found quite recently, in 2001–2002. This construction [97, 55, 48] directly generalizes the Gilbert (that is the $G_{N,p}$) model and is essentially based on the notion of 'hidden variables'. The general idea of the algorithm used is very simple. (i) To each of the nodes, $i = 1, 2, \ldots, N$, ascribe a number—*a hidden variable*, d_i. (ii) Then connect nodes in pairs with probabilities depending on the hidden variables of these nodes, so that the probability that there is a link between nodes i and j is $p_{ij} = f(d_i, d_j)$. The architecture of the resulting random network is determined by the statistics of the hidden variables and the form of a given function $f(d, d')$. In particular, if $p_{ij} = p$ is a constant, we arrive at the Gilbert model.

Suppose we aim to get an uncorrelated network with a degree distribution $P(q)$. Then use the 'desired degrees' as the hidden variables. Namely, (i) ascribe 'desired degrees' d_i—random numbers drawn from the distribution $P(d)$—to N nodes, and (ii) connect nodes in pairs (i, j) with probabilities p_{ij} proportional to the products $d_i d_j$. Normalization gives $p_{ij} = d_i d_j / N\langle d\rangle$ if we assume that the network is sparse. One can prove that the degree distribution of the resulting network approaches the desired form $P(q)$ at large degrees. One can also show that this network is uncorrelated. Chung and Lu (2002) called this construction 'the random graphs with a given desired sequence of degrees' [55]. This random network and the configuration model approach each other when the networks are infinite. However, networks built by using hidden variables are more convenient for practical purposes than the configuration model. In modern numerical experiments (simulations), researchers usually use this easy algorithm or its variations to build uncorrelated networks, and not the configuration model. Furthermore, these networks are easily treatable analytically.

Maybe, the reader has already noticed a serious problem with this construction. If the product $d_i d_j$ exceeds $N\langle d\rangle$, the probability p_{ij},

defined as above, becomes greater than 1, which is pure nonsense. This problem certainly arises for slowly decaying distributions $P(d)$. A simple patch to remedy this flaw was proposed in the very first work introducing a construction with hidden variables (2001) [97]. Its authors—Goh, Kahng, and Kim—proposed to restrict the probability p_{ij} from above by choosing this probability in the form: $p_{ij} \propto 1 - \exp(-d_i d_j / N\langle d\rangle)$.[2] The resulting construction is called the *static model*. The flaw has been fixed but, in return, another problem has emerged. For slowly decaying distributions $P(d)$, a network built with this p_{ij} appears correlated. Without going into detail, these networks are uncorrelated only if their degree distributions decay sufficiently rapidly. For more detail, see Lecture 8.

[2] This is, of course, only one of possible forms.

5.3 Neighbour degree distribution

In Section 2.3, we have shown that if the degree distribution of an arbitrary network is $P(q)$, then the degree distribution of any of the end nodes of a randomly chosen link is equal to $qP(q)/\langle q\rangle$. We stress that this is the case for any network. Let us introduce a joint distribution $P(q, q')$ of the degrees q and q' of the end nodes of a randomly chosen link. In uncorrelated networks, these degrees, q and q', are independent. So for these networks, a joint degree–degree distribution takes the following factorised form:

$$P(q, q') = \frac{qP(q)}{\langle q\rangle} \frac{q'P(q')}{\langle q\rangle} . \tag{5.1}$$

This also means that the branching of a link does not depend on the degree of its second end. The mean degree of an end of a randomly chosen link is $\langle q^2\rangle/\langle q\rangle$. This is also the mean degree of a randomly chosen node. The mean branching is smaller by 1, and so, as we already know, $\bar{b} = \langle q^2\rangle/\langle q\rangle - 1$. Note that this is also the ratio of z_2 (the average number of the second nearest neighbours of a node) and z_1 (the average number of the nearest neighbours of a node, $\langle q\rangle$).

It is important that almost always $\langle q^2\rangle/\langle q\rangle$ is greater than the mean degree $\langle q\rangle$ in the network.[3] If a degree distribution decays slowly, the difference may be great. For various properties of networks, it is the organization of connections of the nearest neighbours of a node that matters. In respect of these properties, a network seems more dense than it really is. This basic observation explains a great number of phenomena in complex networks. We will exploit this extensively.

[3] Check that the equality $\langle q^2\rangle/\langle q\rangle = \langle q\rangle$ takes place only if a network has no nodes other than the bare ones or the dead ends.

5.4 Loops in uncorrelated networks

The first thing we should do is to find whether these networks are loopy or not. Actually, we could expect that they have very few loops by analogy with classical random graphs, but we must check. Let us, to be concrete, find the clustering coefficient C in the configuration model. Let a network be large, $N \gg 1$. We should find the probability that

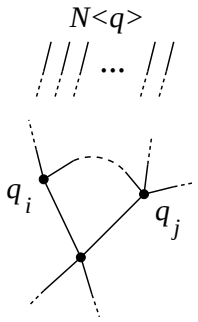

Fig. 5.2 Calculating the clustering coefficient in the configuration model. All the stubs in a network are connected in pairs at random, with equal probability.

two nearest neighbours of a randomly chosen node have at least one link between them (multiple connections are allowed). Suppose that these two nodes, i and j, have q_i and q_j connections, respectively, see Fig. 5.2. Then we have $q_i - 1$ stubs at the first node and $q_j - 1$ stubs at the second, and also nearly $N\langle q\rangle$ stubs at the other nodes. We use the fact that in the configuration model these stubs are connected in pairs at random. Then a stub at node i is connected to one of the stubs at node j with probability $(q_j - 1)/(N\langle q\rangle)$. So for all $q_i - 1$ stubs together we get the total probability $(q_i - 1)(q_j - 1)/(N\langle q\rangle)$. Now average this expression over degrees q_i and q_j, taking into account that these degrees are distributed as $qP(q)/\langle q\rangle$. This readily gives the clustering coefficient of an uncorrelated network [133]:

$$C = \overline{C} = \frac{1}{N\langle q\rangle}\left(\frac{\langle q^2\rangle - \langle q\rangle}{\langle q\rangle}\right)^2 = \frac{\overline{b}^2}{N\langle q\rangle}\,. \tag{5.2}$$

For a Poisson degree distribution, this result is reduced to the expression $C = \langle q\rangle/N$ for classical random graphs.

At first sight, formula (5.2) does not look radically different from that for classical random graphs. The clustering vanishes in both cases as N approaches infinity. In other words, the clustering is only a finite-size effect in these models. Nonetheless, formally substituting empirical data ($\langle q^2\rangle$, $\langle q\rangle$, and N) for real networks into eqn (5.2) usually provides far more reasonable values than the classical random graph formula. Heavy tails in degree distributions of real networks lead to large $\langle q^2\rangle$, and this in turn results in a sufficiently high value of the calculated clustering coefficient at finite N. Typically, the classical formula underestimates C by several orders of magnitude—three, four, five orders, while with eqn (5.2) we often underestimate it by 'only' several times. Roughly speaking, the configuration model provides the smallest clustering that is possible in a random network of a given size with a given degree distribution, and ignores other details. Thus, clustering of many real-world networks turns out to be not that far from these 'minimum possible values'. To explain real values and size-independent contribution to the clustering coefficient, we should go beyond the configuration model and its variations.

Similarly, one can find the number $\mathcal{N}_L$ of loops of length L in an uncorrelated network. For sufficiently short (e.g., finite) loops, $\mathcal{N}_L \sim \overline{b}^L/(2L)$ [30, 32]. On the other hand, the number of loops longer than the diameter of a network is extremely large, $\ln \mathcal{N}_L \propto N$. In this respect, the situation is very similar to that for classical random graphs—few short loops and many long loops, which corresponds to a locally tree-like architecture.[4] This sea of long loops is a necessary feature of uncorrelated networks and, generally, of constructions of this kind. Thanks to the long loops, these networks have no boundaries, no centres. Their nodes can be distinguished only by their degrees, otherwise they are 'statistically equivalent', as physicists often say. We will show that this absence of boundaries is critically important for cooperative effects in complex networks.

[4] More rigorously, if the second moment of a degree distribution diverges, then the tree-like character disappears. This takes place in infinite networks with scale-free degree distributions if exponent γ is less than or equal to 3. Surprisingly, even in this difficult situation, the tree ansatz sometimes works.

5.5 Statistics of shortest paths

Using the tree ansatz we readily arrive at the asymptotic formula (large N) for the mean internode distance (and diameter) of an uncorrelated network:

$$\overline{\ell} \cong \frac{\ln N}{\ln \overline{b}} \,. \tag{5.3}$$

We have obtained this relation discussing classical random graphs. This result, typical for small worlds, is valid only if the second moment of a degree distribution is finite in the infinite network limit.[5] Otherwise, $\overline{\ell}$ can grow even slower than $\ln N$—*ultra-small worlds*. The resulting form $\overline{\ell}(N)$ depends on how $\langle q^2 \rangle$ approaches infinity with growing N. It is impossible to show a general formula, since this approach differs between different versions of uncorrelated network models. In particular, it may be $\overline{\ell} \sim \ln N / \ln \ln N$ for networks with exponent $\gamma = 3$ and $\overline{\ell} \sim \ln \ln N$ if exponent γ is smaller than 3 [60]. Furthermore, in some situations, $\overline{\ell}(N)$ even approaches a constant as $N \to \infty$. This is the case, for example, if a single node in a network attracts a finite number of all connections, as in Fig. 2.7. Unfortunately, it is hardly possible to distinguish between various slowly varying dependences on N in finite real-world networks.

[5] Recall that $\overline{b} = (\langle q^2 \rangle - \langle q \rangle)/\langle q \rangle$.

Not everything can be done in the framework of the convenient tree ansatz. Here we have indicated only one problem which cannot be solved within this approximation. Let us introduce an important notion which helps to characterize the distribution of shortest paths over a network. *The betweenness centrality* (physicists also call it *load*) shows how often shortest paths in a network pass through a given node.[6] The betweenness centrality of a given node is proportional to the relative number of shortest paths between other nodes which run through this node.[7] More rigorously, the betweenness centrality of a node is defined as follows. Let $s(i,j) > 0$ be the number of shortest paths between nodes i and j, while $s(i,v,j)$ is the number of these paths passing through node v. Then the betweenness centrality $B(v)$ of node v is

[6] The notion of betweenness centrality was first proposed in sociology.

[7] Similarly, one can introduce the betweenness centrality of a link.

$$B(v) \equiv \sum_{i,j \neq v} \frac{s(i,v,j)}{s(i,j)} \,, \tag{5.4}$$

where the sum is over all nodes other than node v. Similarly to the degree distribution, for a random network, one should introduce the betweenness centrality distribution $\mathcal{P}(B)$.

Figure 5.3 shows schematically all possible configurations of shortest paths between three nodes in the same connected component. Note configuration (c) which implies that the presence of loops in a network certainly changes the value of betweenness centrality. Therefore, to calculate the betweenness centrality distribution, one has take into account loops. It turns out that this is a difficult task, and so the problem for loopy networks is still open. As far as we know (2008), even for classical random graphs, an exact betweenness centrality distribution has not been found. Nonetheless, this distribution has been investigated in numerous empirical works and numerical simulations. In a wide range

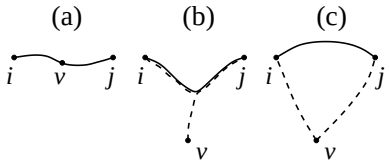

Fig. 5.3 Possible configurations of the shortest paths connecting nodes i, j, and v. The dotted lines show the shortest paths between nodes i and v, j and v.

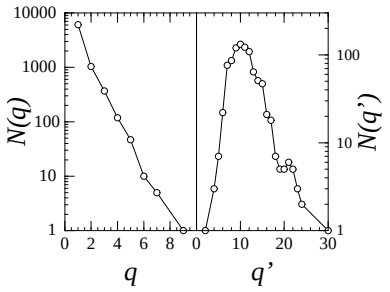

Fig. 5.4 Empirical statistics of the bipartite Fortune 1000 graph. Frequency with which a director sits on q board is on the left panel; frequency with which q' directors sit on a board is on the right panel. Adapted from Newman, Strogatz and Watts [140].

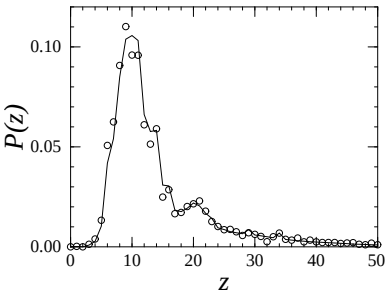

Fig. 5.5 Statistics of the one mode projection (to directors) of the bipartite Fortune 1000 graph. Dots are the empirical data: probability that a director has in total z co-directors on all his boards. The solid line was calculated for the one-mode projection of an uncorrelated bipartite graph with the same statistics as in Fig. 5.4. Adapted from [140].

of networks, including scale-free nets, the observed distribution is rather close to $\mathcal{P}(B) \sim B^{-2}$ [97].

In uncorrelated networks, the width of the distribution of internode distances was found to be independent of N. If a network is sufficiently small, then a noticeable fraction of nodes may be separated by distances essentially greater than $\overline{\ell}$. The statistics for this remote part of a network is of particular interest. The relevant quantity is the distribution of the number of the ℓ-th nearest neighbours of a randomly chosen node if $\ell \gg \overline{\ell}$. It is even more convenient to use the very close distribution of the number of nodes $z_{>\ell}$ at a distance greater than ℓ from a node. Interestingly, a form of these distributions is essentially determined by the presence (or absence) of dead ends in a network. If there are no nodes with a single link in a network, then these distributions rapidly decay. If a finite fraction of all nodes are dead ends, then these distributions decay quite slowly, as $z_{>\ell}^{-2}$.

5.6 Uncorrelated bipartite networks

We have demonstrated how to build an uncorrelated undirected one-partite network. Similarly, one can define uncorrelated models of other networks: directed networks, multi-partite networks, etc. A number of examples were described in a seminal paper by Newman, Strogatz and Watts (2001) [140]. For demonstration purposes, let us focus on undirected bipartite networks, see Fig. 1.4 (a). If the reader needs a real-world example, it can be a bipartite collaboration network of movie actors or one of the networks of the members of boards of directors. In the movie actor network, one type of node is actors and the other is movies where they played. In the network of directors, one kind of node is directors and the other is boards on which they sit. An uncorrelated bipartite network is a network with (i) given numbers of nodes of each sort, N_1 and N_2 and (ii) given degree distributions $P_1(q)$ and $P_2(q')$ for each kind of node, otherwise they are uniformly random. As usual, we can also say that this is the maximally random bipartite network with given N_1, N_2, $P_1(q)$, and $P_2(q')$. Similarly to the one-partite uncorrelated networks, this network has few short loops. Remarkably, the one-mode projection of an uncorrelated bipartite network, Fig. 1.4 (b), is already correlated. Furthermore, this projection is already not tree-like—it has many short loops. In particular, the clustering coefficient C of the one-mode projection does not vanish as the size of this network approaches infinity. This shows how one can easily get large clustering starting from an uncorrelated network.

How far is this model from real bipartite networks? It turns out that it sometimes describes a real situation surprisingly well. In their work, Newman, Strogatz and Watts inspected the Fortune 1000 graph, namely a bipartite network of the members of boards of directors of the US companies with the highest revenues. They obtained three empirical degree distributions: two degree distributions for each of two kinds of

nodes, Fig. 5.4, and the degree distribution to the one-mode projection of this graph to the set of nodes-directors, the dots in Fig. 5.5. Then they built an uncorrelated bipartite network with the same distributions $P_1(q)$ and $P_2(q')$ as in the real network. It was easy to compute the degree distribution of the one-mode projection of this model network, the solid line in Fig. 5.5. The reader can see that the result is very close to the real-world data. The calculated clustering coefficient of the one-mode projection, $C = 0.59$, practically coincided with the measured one. Note the large value of the clustering coefficient. One has to admit that such close agreement is the exception rather than the rule.

Percolation and epidemics

6

The uncorrelated models provide us with a discovery tool to explore complex networks. With this tool, in this lecture we approach a set of related problems for complex networks. What is the global organization of these networks? How can the networks be destroyed? How do epidemics spread through complex networks?

6.1 Connected components in uncorrelated networks

We have described the system of connected components in classical random graphs, see Fig. 2.4. In essence, this qualitative picture with a single giant and numerous finite connected components is quite general. The problem is that in different complex networks, the giant component can emerge in very different ways, with differently distributed finite components. Between 1995 and 1998, graph theory mathematicians Michael Molloy and Bruce A. Reed found a way to explicitly obtain the statistics of connected components in the configuration model [128,129]. Here we explain the idea of this basic approach which physicists mostly know from the previously mentioned paper of Newman, Strogatz, and Watts (2001). Many key results in the science of complex networks have been obtained using this approach.

Let us, for example, find a giant connected component in the configuration model of an infinite undirected network with a degree distribution $P(q)$. Recall that in this model, (i) all nodes and links are statistically independent, (ii) the degrees of the nearest neighbours are independent of each other, (iii) the degree distribution of an end node of a link, coinciding with the degree distribution of any nearest neighbour of a node, is $qP(q)/\langle q\rangle$, and (iv) the structure of a network is locally tree-like.

These properties allow us to introduce the following basic probability x in this problem. Choose at random a link. Select with equal probability one of its ends. Then x is the probability that strictly following this link in the direction of the chosen end we arrive at a 'finite connected component'. In this 'finite component', the nodes approachable following the link in the opposite direction are ignored. So this 'finite component', in principle, may belong to the giant connected component of the network. Figure 6.1 explains this definition. One can graphically represent probability x as shown in Fig. 6.2. This is simply a useful notation.

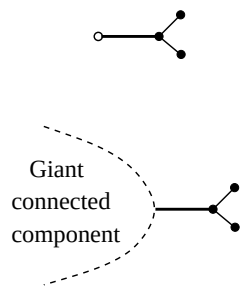

Fig. 6.1 Explanation of the probability x. The right end of the horizontal link is selected. Both of these two, at first sight very different configurations, contribute to x.

Fig. 6.2 Graphical notation for probability x.

Fig. 6.3 The probability that a link is in one of the finite connected components.

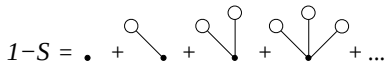

Fig. 6.4 The probability $1 - S$ that a randomly selected node belongs to one of the finite connected components. Here S is the relative number of nodes in the giant connected component.

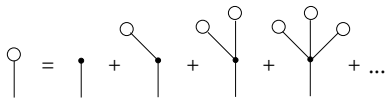

Fig. 6.5 Graphical representation of the equation for probability x.

Using this notation, we can conveniently represent various characteristics of a network. For example, the probability that a randomly chosen link belongs to one of the finite connected components has a graphical representation shown in Fig. 6.3. Indeed, in this situation, following the link in any direction we must approach a finite component. Clearly, this probability equals x^2. Then the fraction of links belonging to the giant connected component equals $1 - x^2$. This provides us with a way to get x if we succeed in measuring the number of links in the giant component of an uncorrelated network. Our aim, however, is not to measure but to calculate characteristics of connected components for a given $P(q)$.

If we find a way to get x, we will be able to obtain the relative size S of the giant connected component and many other quantities. Figure 6.4 shows how to find the probability $1 - S$ that a randomly chosen node belongs to one of the finite components, if x is known. This is a sum of the probabilities: the probability that a node has no connections plus the probability that it has a single link and that this link leads to a finite component, and so on. Note that this equality necessarily implies a tree-like structure of a network. Figure 6.4 leads to the following simple formula:

$$1 - S = \sum_q P(q) x^q \,. \tag{6.1}$$

Now we should find probability x. To derive an equation for x we again use graphical notations. Figure 6.5 explains the form of this equation. To write it down we use the degree distribution of an end node of a randomly chosen link, $qP(q)/\langle q \rangle$. As a result we have

$$x = \sum_q \frac{qP(q)}{\langle q \rangle} x^{q-1} \,. \tag{6.2}$$

Thus the problem is essentially reduced to the analysis of this equation for a given distribution $P(q)$. When it has a non-trivial solution $x < 1$, the network has a giant connected component. One can easily show that this takes place when the mean number of second nearest neighbours of a node exceeds the mean number of nearest neighbours: $z_2 > z_1$. This is the celebrated *Molloy–Reed criterion* (1995),

$$\langle q^2 \rangle - \langle q \rangle > \langle q \rangle \tag{6.3}$$

which is the mean branching $\bar{b} > 1$. At the birth point of a giant connected component, the mean number of second nearest neighbours coincides with the mean number of nearest neigbours, $z_2 = z_1$. In particular, this leads to the proper critical point $\langle q \rangle = 1$ for classical random graphs.

Surprisingly, these equations and the criterion may still be valid even when the second moment of a degree distribution diverges and the tree ansatz becomes dubious. Look at criterion (6.3): if $\langle q^2 \rangle$ diverges, a network necessarily has a giant connected component. On the other hand, one can check using these equations that, if an uncorrelated network has

no dead ends, then almost all nodes in the network are in the giant connected component. In the next section we will show how strongly the size of the giant component depends on the form of the degree distribution.

Using these equations, one can find the organization of connections inside the giant component. Importantly, the distribution of connections in a giant connected component may essentially differ from that of the entire network. The strongest difference is observed close to the birth point, where the degree distribution in a giant component is proportional to $qP(q)$.

The same ideas were successfully used to find the statistics of finite connected components in uncorrelated networks [140]. The derivation, however, was more cumbersome, so here we only outline the main results. Similarly to classical random graphs, in any uncorrelated network the size distribution of finite components, $\mathcal{P}(s)$, is a power law at the birth point of a giant connected component. If the degree distribution of a network decays sufficiently rapidly, then the distribution $\mathcal{P}(s) \sim s^{-5/2}$ at this critical point, as in classical random graphs. For slowly decaying degree distributions, the exponent of $\mathcal{P}(s)$ differs from 5/2.[1]

That was a 'critical' distribution. What about the size distribution away from the critical point? When a network with any degree distribution contains a giant connected component, its finite components have exponentially rapidly decaying size distribution.[2] For 'non-critical' situations, this is not surprising but typical. What is surprising is that in the other phase, where a giant connected component is absent, scale-free and non-scale-free networks have very different distributions of finite components. In scale-free uncorrelated networks in this situation the size distribution decays slowly, $\mathcal{P}(s) \sim s^{-\gamma}$ [136]. In contrast, in non-scale-free networks ($\gamma = \infty$), $\mathcal{P}(s)$ decays exponentially rapidly.

[1] In particular, if an uncorrelated network is scale-free, $P(q) \sim q^{-\gamma}$, then (i) for $\gamma > 4$, the distribution $\mathcal{P}(s) \sim s^{-5/2}$, and (ii) as γ decreases from 4 to 3, the exponent of $\mathcal{P}(s)$ increases from 5/2 to 3 [61].

[2] This is true for an arbitrary network, and not only for uncorrelated.

6.2 Ultra-resilience phenomenon

One of main questions to ask about a network concerns its robustness. The standard approach to this problem is to study how the removal of nodes influences the global structure and function of a network. Remove a fraction of nodes from a network and investigate how the size of a giant connected component diminishes—this is the essence of hundreds of papers on this topic. In terms of condensed matter theory, this is a site percolation problem.

The nodes for deletion may be chosen in various ways. The simplest way is to choose them uniformly at random. Let us remove a fraction $1 - p$ of uniformly selected nodes from an uncorrelated network with a given degree distribution, and so with given $z_1 = \langle q \rangle$, $z_2 = \langle q^2 \rangle - \langle q \rangle$, and $\overline{b} = z_2/z_1$. Then, when will a giant connected component disappear? The answer was found using the same approach as in the preceding section. The giant connected component is present in a network if

$$pz_2 > z_1 \,, \tag{6.4}$$

so the percolation threshold (the birth point of a giant connected component) is at

$$p_c = \frac{z_2}{z_1} = \frac{\langle q \rangle}{\langle q^2 \rangle - \langle q \rangle} = \frac{1}{\overline{b}}\,. \tag{6.5}$$

[3] For derivations and details see papers [50, 58].

Here z_1, z_2, $\overline{b}$, $\langle q \rangle$, and $\langle q^2 \rangle$ are for an undamaged network.[3] This relation shows that if the degree distribution of an uncorrelated network has so heavy a tail that $\langle q^2 \rangle$ diverges, then the threshold p_c is zero. In other words, we have to remove practically all nodes from this network to eliminate its giant connected component. In scale-free networks, $\langle q^2 \rangle$ diverges when exponent $\gamma \leq 3$, which is the case for many real-world networks. In 2000, Albert, Jeong, and Barabási investigated a number of real networks, including the WWW and the Internet, and for all of them observed this ultra-resilience against random failures [6].

Often the ultra-resilience is considered as one of the most impressive phenomena induced by complex network architectures. However, this effect of the heavy tails of degree distributions (the abundance of hubs) is almost trivial. Let us discuss the extreme situation—a finite fraction c of nodes are connected to a single node. For simplicity, we assume that without this super-hub, all of these cN nodes are not connected. Suppose that we select nodes for removal uniformly randomly, with probability $1-p$. Then the average size of a giant connected component in this network is $(1-p)\cdot 0 + p \cdot pcN = p^2 cN$. So for any non-zero p, the giant connected component contains a finite fraction $p^2 c$ of nodes in the network, which means ultra-resilience. Now suppose that we want to cause the maximum damage. Then, as is natural, remove the hub and there will be a chance to destroy a giant component. Thus networks with heavy degree distributions are extremely robust against uniformly random damage but, simultaneously, quite weak against intentional damage. It is usually enough to remove a few per cent of highly connected nodes to completely split a network with a heavy-tailed degree distribution [50, 59]. This is in sharp contrast to the networks with rapidly decreasing distributions of connections, where both kinds of damage produce similar effects. Irrespective of its simplicity, the ultra-resilience phenomenon have numerous consequences with regard to the functioning of complex networks, which we will discuss in the following lectures.

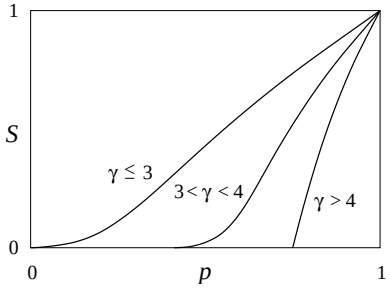

Fig. 6.6 Relative size of a giant connected component versus the fraction of nodes retained after random damage. The three curves show the typical dependences at various values of the exponent of a scale-free degree distribution.

Let us return to the uniformly random removal of nodes and discuss the role of the form of degree distribution regarding the critical properties. Suppose our network is scale-free. Consider the variation of the relative size $S(p)$ of a giant component, varying with exponent γ near the percolation threshold $p_c(\gamma)$. A schematic plot, Fig. 6.6, gives a qualitative picture. When $\gamma > 4$, the critical behaviour is very similar to that for the classical random graphs. A sharp difference occurs below $\gamma = 4$, that is when the third moment of the degree distribution diverges. Quantitatively, the critical features of $S(p)$ are as follows [57]:

(i) if $\gamma > 4$, then $S \propto p - p_c$,
(ii) if $3 < \gamma < 4$, then $S \propto (p - p_c)^{1/(\gamma-3)}$,
(iii) if $\gamma < 3$, then $S \propto p^{1+1/(3-\gamma)}$ and $p_c = 0$.

These results were obtained by using the approach described in the preceding section. They assume that any finite neighbourhood of a node has no loops. Items (ii) and (iii) in this list must surprise physicists. Let us discuss this point in more detail.

As is well known from condensed matter physics, critical phenomena in models defined on the top of infinite-dimensional objects are exactly described by mean-field (molecular-field) theories.[4] There are three standard infinite-dimensional objects in condensed matter: (i) an infinite-dimensional lattice, (ii) a fully connected graph, and (iii) a Bethe lattice. For percolation on each of these objects, $S(p) \propto p - p_c$, and this is also the result of mean-field theory for sufficiently homogeneous systems. This theory, however, is not applicable to low-dimensional systems. It fails, for example, in two- and three-dimensional lattices. One of the challenges of theoretical physics in the 20th century was to advance beyond mean-field theories. During several of the last decades of the 20th century physicists arrived at a clear understanding of critical phenomena in low-dimensional ordered and disordered lattices in situations where the mean-field theories fail. In these situations, critical features of observables are power-law singularities, e.g., $S(p) \propto (p-p_c)^\beta$, with non-mean-field exponents. The birth of a giant component in classical random graphs and networks with rapidly decaying degree distributions is well described by standard mean-field theory, as it should be for small worlds. Our list, however, shows that when $\gamma < 4$, the standard mean-field theory fails. For γ between 3 and 4, the critical feature of $S(p)$ differs from the standard mean-field one as in two- and three-dimensional lattices, though our networks are infinite-dimensional. On the other hand, it turns out that for any value of exponent γ, the mean size of a finite component to which a node belongs has the same critical singularity: $\langle s\rangle' \propto 1/|p - p_c|$. So for this quantity, in contrast to $S(p)$, the standard mean-field works at any degree distribution.

At first sight, this combination of singularities of different kinds looks very strange. The actual reason for this 'strange' critical behaviour is the strong heterogeneity of these complex networks, namely a wide range of node degrees. This heterogeneity does not allow one to use the standard mean-field theory developed for homogeneous systems. Nonetheless, if we modify a mean-field theory, accounting for the heterogeneity of a network, then we will get correct critical singularities.[5]

[4] In physics, a mean-field theory is the main approach to systems of interacting agents, for example, spin $\mathbf{S}$. In this approximate approach, instead of solving a full set of equations for all spins, a single spin in a mean field from its neighbours is considered. This effective field is taken as $\mathbf{H}_{\text{eff}} \approx \langle\mathbf{S}\rangle z_1$, i.e. the mean value of spin times the coordination number of a lattice. Within this approximation, one can obtain $\langle\mathbf{S}\rangle = f(\mathbf{H}_{\text{eff}}) = f(\langle\mathbf{S}\rangle z_1)$ and then find $\langle\mathbf{S}\rangle$ from this self-consistent equation.

[5] We can even treat eqn (6.2) as a self-consistent equation for order parameter in a specific mean-field theory.

6.3 Finite-size effects

The reader has certainly noticed that the strongest effect takes place when the second moment of the degree distribution diverges. The problem is that this divergence is possible only if a network is infinite, while all real networks are, of course, finite. In finite networks any heavy-tailed degree distribution necessarily has a cut-off—an end part of the distribution with a rapid decay, see Fig. 6.7. The position of this cut-off, $q_{\text{cut}}(N)$, determines the values of higher moments of a degree

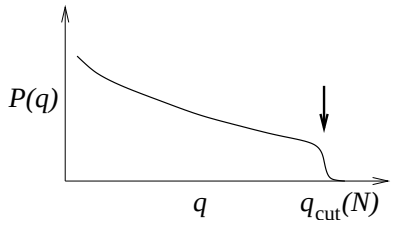

Fig. 6.7 Size-dependent cut-off of a degree distribution.

distribution and so strongly influences various properties of a network. This is why $q_{\rm cut}(N)$ is one of the basic network characteristics. For example, in a scale-free network with exponent γ less than 3, the second moment of a degree distribution takes the form $\langle q^2\rangle \sim q_{\rm cut}^{3-\gamma}(N)$. Size dependent $q_{\rm cut}(N)$ differs from model to model.[6]

For a single scale-free graph, that is for a single member of an ensemble, the highest possible cut-off $q_{\rm cut}(N)$ is estimated to be of the order of $N^{1/(\gamma-1)}$. If the cut-off is of this order, then the graph has a finite number of nodes with degrees exceeding $q_{\rm cut}$, or even a single node in this range. This cut-off is called 'natural'. In some network models, however, $q_{\rm cut}$ is far below that 'natural' value. Here we indicate only one network model with a cut-off of this kind. Consider a scale-free uncorrelated network with exponent $\gamma < 3$, where multiple connections and one-loops are forbidden. One can show that with these restrictions, the cut-off $q_{\rm cut}(N)$ cannot be higher than of the order of $N^{1/2}$. Nodes of higher degrees would make this network correlated. Remarkably, if nodes with this (of the order of $N^{1/2}$) or higher degrees are present, then they form a densely interconnected club within a sparse network. Thus nodes with the highest numbers of links are interconnected. This was named a *rich-club phenomenon* and observed in many real-world networks, including the AS and router level Internet networks [185]. Note, however, that this club contains only a tiny fraction of nodes in a network.[7]

To estimate the size effect, one can formally substitute the resulting second moment of the degree distribution in a finite network into eqn (6.5) for the percolation threshold and readily obtain a finite p_c. Any physicist will understand that this is a very naive approach. Indeed, continuous phase transitions are, in principle, impossible in finite systems, and $p_c(N)$ and a giant connected component are well defined only in the limit $N \to \infty$. Nonetheless, this dubious way is possible if we only need a simple estimate. Thus, in real-world networks, the heavy tails of degree distributions cannot cause absolute resilience. We can claim, however, that even finite networks of this architecture are far more robust against random failures than the classical random graphs.

[6] See the list of cut-offs for various network models in review [77].

[7] The reader can easily check that this elite club contains only about $N^{(3-\gamma)/2}$ nodes [63]; which is much less than N.

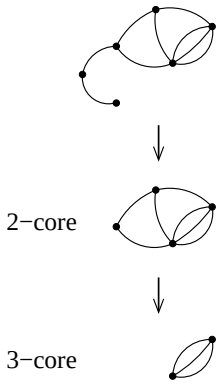

Fig. 6.8 The k-cores of a small graph. To find a k-core, one can use the 'pruning algorithm'. Remove from a graph all nodes of degree less than k. After that, some of the remaining nodes may occur with less than k links. Prune these nodes, and so on. The final result, if it exists, is the k-core.

6.4 k-cores

The k-core of a network is its largest subgraph whose nodes have at least k connections (within this subgraph, of course) [54]. Figure 6.8 shows the k-cores in a small graph. In general, there is a set of successively enclosed k-cores, similarly to a Russian nesting doll—'matrioshka'. The full set of the k-cores of a network, $k = 1, 2, 3, \ldots$, presents essentially a more detailed description of the network than the one based on connected components. Knowledge of k-cores enables us to indicate the best interconnected parts in a network. The k-cores have been used as a network visualization tool. The reader will find the image gallery of picturesque k-core decompositions of various networks in [9] and on a special site http://xavier.informatics.indiana.edu/lanet-vi/.

In essence, the 2-cores are rather close to connected components—only the dangling chains are eliminated, see Fig. 6.8. The 3- and higher-k-cores, however, show a sharp contrast to connected components. Let us discuss these k-cores.

First note that the trees have no ($k\geq 2$)-cores. Use pruning, similarly to Fig. 6.8, to understand this claim. Then if a network is locally tree-like, it, in principle, cannot have finite ($k\geq 2$)-cores. In loopy networks (more precisely, networks having short loops), a single giant and numerous finite k-cores can coexist, but in locally tree-like networks there can only be a single giant k-core. In 1979, Chalupa, Leath, and Reich found that the birth of this giant k-core, for $k \geq 3$, is a quite unusual phase transition [54], different both from continuous and first-order transitions, see Fig. 6.9. Formally speaking, these authors focused not on networks, but rather on lattices. Namely, they studied the process of the elimination of a giant k-core by removing random sites from a lattice. Their findings, however, turned out to be valid for a wide range of lattice and network architectures.[8]

[8] In their calculations, these researchers used a Bethe lattice, which we know is infinite-dimensional.

Normally, physicists divide phase transitions into continuous and first order, Fig. 6.9 (a) and (b). The birth of a giant connected component is a continuous transition. There is no jump of S (relative size of a giant component) at the birth point, Fig. 6.9 (a). In contrast, in a first-order phase transition, an order parameter X emerges abruptly, Fig. 6.9 (b). Surprisingly, the transition associated with the birth of a k-core combines the features of continuous and first-order phase transitions, see Fig. 6.9 (b). (i) At the birth point, q_c, there is a jump of S_k, the relative size of a k-core, from zero to a finite value. (ii) In addition, there is a square root singularity at the same point:

$$S_k(\langle q\rangle) - S_k(q_c) \cong \text{const}\sqrt{\langle q\rangle - q_c}\,. \tag{6.6}$$

This phase transition, which is called *a hybrid transition* or, sometimes, a mixed transition, is not so common as first- and second- order phase transitions. Note, however, that k-cores have many applications, for example, in various jamming phenomena and in systems for limiting metastable states.

Here we will not discuss the nature of the hybrid phase transition, see discussion in review [77]. For us, it is more interesting how k-cores are organized in various networks. It was found that this specific transition takes place in networks with sufficiently rapidly decaying degree distributions, namely if $\langle q^2\rangle < \infty$, see schematic Fig. 6.10 (a). If the degree distribution of an infinite network decays more slowly (in scale-free networks, this corresponds to the region $\gamma \leq 3$), then this network has an infinite sequence of k-cores for any non-zero value of mean degree, Fig. 6.10 (b). In this range of γ, the k-cores behave similarly to a giant connected component. Instead of mean degree, we can use another control parameter, p, which is the probability that a node is retained in a uniformly damaged network. On a qualitative level, we simply substitute $\langle q\rangle$ in Fig. 6.10 for p. Then compare Figs. 6.6 and 6.10. Clearly, for $\gamma \leq 3$ in an infinite network, all k-cores are ultra-resilient against

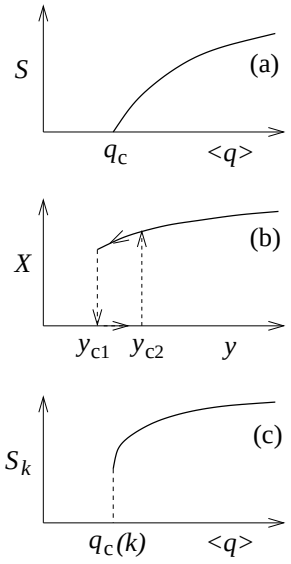

Fig. 6.9 Three phase transitions. (a) Continuous transition, the birth of a giant connected component. S is the relative size of a giant component. (b) First-order transition. X stays for an order parameter, y denotes a control parameter. (c) Hybrid phase transition, the birth of a k-core. Here S_k is the relative size of a giant k-core. Near the jump at $q_c(k)$, the deviation of S_k varies as $\sqrt{\langle q\rangle - q_c(k)}$.

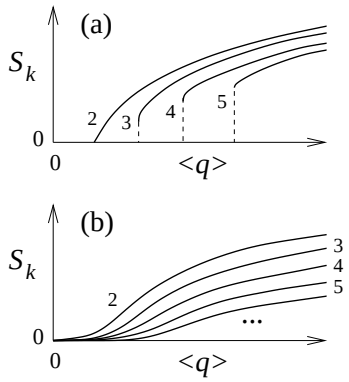

Fig. 6.10 Schematic view of the birth process of the k-core set in uncorrelated networks when: (a) the second moment of a degree distribution is finite, (b) the γ exponent of a power-law degree distribution does not exceed 3 [76]. Instead of mean degree $\langle q \rangle$, one can use a concentration of retained nodes in a uniformly damaged network.

random damage. If a network is finite, the k-cores are already not absolutely robust—Fig. 6.10 (b) transforms into Fig. 6.10 (a). Notice that as random damage increases (p diminishes), first the highest k-core disappears, then the second highest, and so on. That is, the damage primarily spoils the best interlinked areas in a network.

6.5 Epidemics in networks

How does an infectious disease spread? Generations of epidemiologists and population biologists have studied this challenging problem. One can safely say that the spread of disease in homogeneous systems is now well understood. Traditional theories describe epidemics within fully connected graphs, classical random graphs, and lattices. Usually in epidemic models, each of the individuals can be in two or three states: S—susceptible, I—infective, and R—removed or recovered but not susceptible. So the dynamics of epidemics should be described in terms of three probabilities for each of the individuals in a population: the probabilities $i(t)$, $s(t)$, and $r(t)$ that this individual is infective, susceptible, and removed, respectively. Specific epidemic models use different sets of states and set rules for transitions between them.

To be more concrete, in *the SIS model*, an individual can be susceptible or infective. The rate of recovery of an infected individual ($I \to S$) is μ. A susceptible individual becomes infected ($S \to I$) with a rate β if at least one of its nearest neighbours is infective. The main parameter of the problem is the so-called reproductive number $\lambda = \beta/\mu$.

In the second basic model—*the SIR model*—an individual can be susceptible, infective, or recovered. The 'recovery' rate for an infected individual ($I \to R$) is μ.[9] As in the SIS model, a susceptible individual becomes infected ($S \to I$) with a rate β if at least one of its nearest neighbours is infective. The main parameter of the problem is the reproductive number for this infection, $\lambda = \beta/\mu$.

[9] This 'recovery' may occur as removal.

These models are defined on various substrates—networks or lattices—where each individual permanently occupies its own node and never changes it. Let us first touch upon classical homogeneous situations, where the nodes in a network or a lattice have a narrow distribution of connections.[10] Suppose that initially a single or a few nodes are infected. It turns out that the disease will quickly die out if the reproductive number is below some value, *an epidemic threshold*, λ_c. This is one of the fundamental notions in epidemiology. Another one is *the prevalence*, which is the fraction of infected individuals. In homogeneous situations, the epidemic threshold is determined by the mean degree of a node in a network: $\lambda_c \sim 1/\langle q \rangle$. Above the epidemic threshold, an epidemic spreads throughout a network.

[10] Here we discuss only the spread of disease within a giant connected component.

This general picture is valid for the SIS and SIR, and for many other epidemic models. There are differences, but they are not that important. Above the epidemic threshold in the SIS model, the prevalence $\langle i \rangle(t)$ in a population grows monotonously. It is finite in the final state. In

contrast to this, in the SIR model, $\langle i \rangle(t)$ shows an epidemic outbreak but approaches zero at $t \to \infty$: $\langle i \rangle = 0$. This outbreak, however, results in a finite fraction of removed (or recovered) individuals in the final state, $\langle r \rangle > 0$. Figure 6.11 shows the evolution of an epidemic in these two models.

In 2001, Pastor-Satorras and Vespignani [149] moved beyond this traditional focus on epidemics in homogeneous media. They considered the speed of infectious disease within an uncorrelated network with an arbitrary degree distribution. The results that they obtained naturally generalize the picture of epidemics described above. The most important result was the epidemic threshold. In the SIR model,

$$\lambda_c = \frac{\langle q \rangle}{\langle q^2 \rangle - \langle q \rangle} = \frac{1}{\overline{b}}\,, \tag{6.7}$$

which coincides with the percolation threshold, eqn (6.5). For physicists, this coincidence is not surprising. They know that the SIR model is equivalent to the percolation problem in many respects.[11] Consequently, in networks with heavy-tailed degree distribution, the epidemic threshold is low. In real-world networks, it may be dramatically lower than the $1/\langle q \rangle$ value of classical random graphs. What about the dynamics of epidemics above the epidemic threshold? In the qualitative Fig. 6.11 for the evolution of prevalence, the characteristic time scale is $\tau = \lambda_c/\lambda$. This is a typical time for an epidemic outbreak.

Thus the presence of hubs in networks leads to low epidemic threshold and τ. How can we protect these networks against epidemics, that is, increase the epidemic threshold? The obvious answer is primarily to immunize the strongly connected nodes. Unfortunately, for this it is necessary to know these nodes, which is often not the case. This is why a number of immunization strategies were proposed on how to pick up the 'best' individuals for immunization with restricted information about a network [62]. The weak point of these algorithms is that usually they assume some type of the network architecture. If, for example, we suppose that a network is uncorrelated, than a relevant strategy appears as follows. Choose a node at random and immunize one of its nearest neighbours. Repeat this as many times as you like. If the real network is indeed uncorrelated, then this strategy certainly works—you will immunize the hubs with higher probability. However, the same strategy, applied to networks of another type, may be a complete failure.[12]

In many respects, the basic infection models, which we have discussed, are not realistic. In these models individuals stay permanently at their nodes, one individual per node. In the real world, the spread of infectious disease, like pandemic influenza, is rather due to high population mobility. Millions of passengers travel across transportation networks, from city to city, increasing the potential for rapid spread of disease. In models for this process, nodes—urban areas—are occupied by many individuals. With some mobility rate κ, these individuals travel from node to node with a chance to transmit infectious disease from one population to another. (Within populations, the spread of disease is de-

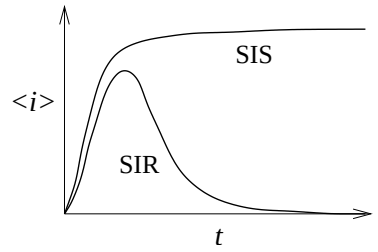

Fig. 6.11 Fraction of infected individuals versus time above an epidemic threshold in the SIS and SIR models. In the SIR model, the fraction of removed individuals $\langle r \rangle$ grows monotonously with time and approaches a constant value at $t \to \infty$.

[11] For the SIS model, a close expression was found, $\lambda_c = \langle q \rangle / \langle q^2 \rangle$.

[12] If you want to choose a set of nodes with degree distribution $qP(q)/\langle q \rangle$ from a network with an unknown architecture, use the following approach. Choose a node uniformly from the set of all nodes. Then select each of its nearest neighbours with some probability p, where $0 < p \leq 1$. Repeat this for as many randomly chosen nodes as you want.

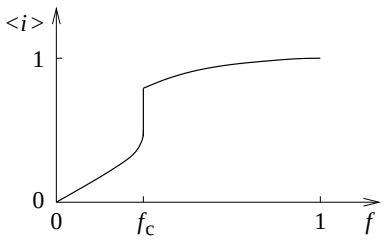

Fig. 6.12 Schematic plot of a final prevalence as a function of the fraction f of initially infected nodes in the bootstrap percolation problem. Note a square-root singularity on the low-f side of the transition: $\langle i\rangle(f_c{-}0){-}\langle i\rangle(f) \propto \sqrt{f_c{-}f}$.

scribed by standard epidemic models.) In other words, these metapopulation models describe the process of the invasion of disease in a network of well-interlinked populations. Remarkably, the main conclusions for epidemics obtained using the SIS and SIR models remain valid in metapopulation models. If these models are defined on top of uncorrelated networks, an invasion threshold for an epidemic, κ_c, is proportional to $\langle q\rangle/(\langle q^2\rangle{-}\langle q\rangle) = 1/\overline{b}$, similarly to the percolation and epidemic thresholds (6.5) and (6.7) [64].

In standard epidemic models, a final stationary prevalence $\langle i\rangle$ emerges at the epidemic threshold in a similar way as a giant connected component at the percolation threshold, see Fig. 6.6. Similarly to a giant connected component, $\langle i\rangle$ rises continuously from zero as is should be at a continuous phase transition. There is a class of models, however, where prevalence emerges explosively, after a sudden jump from zero. In the best known of these models, *the bootstrap percolation problem*, disease spreads in the following way [3].

- At the initial moment, a fraction f of nodes in a network are infected for ever. Then, at each time step, all nodes with k or more infected nearest neighbours become infected. Here $k \geq 3$.

Figure 6.12 demonstrates how a final prevalence, resulting from this process, depends on f. This picture is valid for classical random graphs. There is a discontinuous jump at the critical point f_c accompanied by a square-root singularity on the left-hand side of the transition. So this transition is hybrid similarly to the k-core transition. This is not strange, because the bootstrap percolation and k-core problems are generically related. Recall the definition of a k-core and compare somewhat complementary Figs. 6.12 and 6.9(c). Notice that the square-root singularities in these plots are on opposite sides of the transitions. Finally, it is worthwhile mentioning that the main applications of bootstrap percolation are not to epidemiology, but rather to the failure of units in large arrays and related problems.

Self-organization of networks

7

The formal network constructions (the configuration model and others) which we have discussed do not explain why and how networks get their complex structures. In this lecture we show how the evolution of a network shapes its architecture.

7.1 Random recursive trees

Let us start with the simplest growing random networks. We have already introduced random recursive trees in Lecture 3. Recall the growth process illustrated by Fig. 3.6 (b): at each time step, a new node is attached to a uniformly randomly chosen existing node by a single link—uniform random attachment. We explained that these networks differ strongly from equilibrium ones. In particular, their degree distribution, $P(q) = 2^{-q}$ decays more slowly than the Poisson degree distributions of the equilibrium random trees and the classical graphs. The reason for this exponential decay is clear. The number of connections of a node depends on its age. The 'old' nodes certainly get more links than the 'young' ones. For example, the mean degree of the oldest node—the root—equals $1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{t} \cong \ln t$ in the network of $t + 1$ nodes, while the youngest node has a single link.[1] This results in a wider degree distribution than a Poisson one.

There is another important distribution function—the distribution $p(q, u, t)$ of the number of connections q of a node born at a given time $u < t$. To measure this distribution we must grow many members of the ensemble by repeating the growth process many times. One can show that the resulting distribution is Poissonian. It is known that if the first moment of a Poisson distribution is large, then this distribution is relatively narrow, its width is about the square root of the first moment. For example, the mean degree of the root is $\ln t$, and possible fluctuations from this value are only about $\sqrt{\ln t} \ll \ln t$. To be clear, if we grow a single random recursive tree and measure the degree of its root, then we obtain the value $\ln t \pm \sqrt{\ln t}$.

One should note that these results for degree distributions are qualitatively valid even if the random recursive network is not a tree, see Fig. 3.6 (a). It is only important that the attachment is uniformly random. There are, however, interesting characteristics, relevant only for

[1] More generally, the mean degree of a node born at time u varies as $\ln(t/u)+1$.

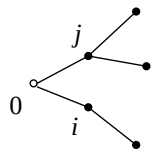

Fig. 7.1 In this recursive tree, node i has one descendant, node j has two, and the root 0 has five. The root has two branches of sizes 2 and 3.

recursive trees. We mention here only two of these. The first is the distribution $\mathcal{P}_d(s)$ of the number of descendants of a node in this tree, see Fig. 7.1. In 2001 Krapivsky and Redner found that this distribution decays very slowly, as s^{-2} [115]. Moreover, it was shown that this law for $\mathcal{P}_d(s)$ is practically general for recursive trees. The uniform attachment was not necessary. For branches of the root, even more slowly decaying distribution of branch sizes was found [92]. The mean number of branches with s nodes is exactly $1/s$. That is, in the random recursive tree, on average, the root has one leaf ($s=1$). Thus, for some distributions, power laws can be found even in the random recursive trees. Researchers, however, were more interested in degree distributions.

7.2 The Barabási–Albert model

In 1999 Barabási and Albert published a truly timely paper addressing the problem of heavy-tailed degree distributions [14]. In this celebrated work they presented a very simple growth model—actually, a concept—pretending to explain the scale-free architectures of real-world networks. The idea was to grow a network, preferentially adding new connections to already highly connected nodes, that is, to exploit the 'rich get richer' effect. Barabási and Albert named this process *preferential attachment*—preferential attachment of new nodes to existing nodes of higher degrees [4]. It is obvious that preferential attachment should increase the number of hubs. The question is: how can it produce scale-free networks?

The Barabási–Albert model is an undirected recursive network, Fig. 3.6 (b), defined as follows. The growth starts from some initial configuration of nodes and links which is actually not important for the structure of the network when it is already large. At each time step,

(i) add a new node to the network and

(ii) attach it to $m \geq 1$ preferentially chosen nodes. Each of these nodes is chosen with a probability proportional to its degree—proportional preference.

Since each new node has m links, the mean degree of a node in this network is $2m$. In particular, $m=1$ corresponds to a tree. The tail of the degree distribution, however, turns out to be independent of m. It is important that the probability of selecting a node of degree q in this process is $q/(2mt)$. When this network is already large, of $t \to \infty$ nodes, its degree distribution can easily be found.[2] The result is a stationary power-law distribution, $P(q) \propto q^{-\gamma}$ with exponent $\gamma = 3$. Of course, this is a very particular power law, but the Barabási–Albert model only demonstrates the principal possibility of generating a scale-free distribution by preferential attachment. Note that the model assumes proportional preferential attachment only for simplicity. In the next sections we will show how to obtain power laws with arbitrary exponents and more general heavy-tailed distributions.

[2] Consider the mean number $\langle N(q,t)\rangle$ of nodes of degree q in a network of size t. For brevity, set $m=1$. Then, for two consecutive moments,

$$\langle N(q,t{+}1)\rangle = \langle N(q,t)\rangle + \frac{(q{-}1)\langle N(q{-}1,t) - q\langle N(q,t)\rangle}{2t} + \delta_{q,1}.$$

The second term on the right-hand side is due to attachments to nodes of degree $q{-}1$ with probability $(q{-}1)/(2t)$ and to nodes of degree q with probability $q/(2t)$. The Kronecker symbol $\delta_{q,1}$ describes the increase in the number of nodes of degree 1 due to a new node. The degree distribution is $P(q,t) = \langle N(q,t)\rangle/t$. Insert this into the equation, assume that the resulting distribution is stationary at $t{\to}\infty$, and finally approximate the discrete difference in degree by a derivative d/dq in the range $q{\gg}1$. This gives the equation:

$$P(q) = -\frac{1}{2}\frac{d[qP(q)]}{dq}\,.$$

Its solution is $P(q) \propto q^{-3}$.

One can easily see that for $m > 1$, the Barabási–Albert model and other recursive networks of this kind have a locally tree-like structure if a network is large.[3] So its clustering coefficient approaches zero in the large network limit ($t = N \to \infty$) [40]:

$$C \approx \overline{C} \sim (m-1)\frac{(\ln N)^2}{N}\,. \tag{7.1}$$

[3] There is a very small chance that a new node will be attached to closely separated nodes.

The low clustering, however, is not related to the preferential attachment mechanism. There are many network models with preferential attachment and, simultaneously, with high clustering and many loops. As one could expect, this network is a small world, but, interestingly, there is a marked difference in the small-world effect at $m = 1$ (a tree) and $m > 1$ (a locally tree-like network). The mean internode distances grow with N differently [41]:

$$\overline{\ell}(m=1) \sim \ln N \quad \text{but} \quad \overline{\ell}(m>1) \sim \frac{\ln N}{\ln\ln N}\,. \tag{7.2}$$

Thus long loops play an important role. They make this network more compact.

7.3 General preferential attachment

One should note that the general idea of using preferential choice to obtain heavy-tailed distributions emerged long before the boom in complex networks.[4] In 1925 G. Udny Yule published 'A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis' which at first sight was very far from networks [183]. Yule proposed a possible explanation for the earlier observed power-law distribution of the number of species in genera. This distribution is a power law with exponent close to 2. The idea of Yule was ultimately simple. It is known that new species emerge due to mutations. Yule supposed that the evolution of a set of genera is a growth process determined by (i) the emergence of new genera with a single species and (ii) the emergence of new species in already existing genera. Two possibilities were taken into account. (1) There is a small chance that mutations create a new genus. (2) Mutations in genera with more species are more frequent and so the probability of the emergence of a new species is greater. The latter dramatically increases the number of large genera. Similarly to the Barabási–Albert model, if the frequency of the emergence of a new species is proportional to the number of species in a genus (a very natural assumption), then this growth leads to a power-law distribution. It is very easy to check that its exponent is indeed 2. Processes of this kind are often called *Yule processes*.

[4] For a brief history of concepts of power-law distributions and relations between various models see [127, 42, 162].

Power laws have been observed in many other distributions. Vilfredo Pareto found this in an income distribution (1897). J. B. Estoup (1916) and George Kingsley Zipf (1932) observed a power-law distribution for the frequency of occurrence of distinct words in a text (exponent close to 2). The distribution of city size (F. Auerbach, 1913) and the distribution

of the number of papers by a scientist (Alfred James Lotka, 1926) are also power laws. And these are only a few examples. In 1955 Herbert A. Simon (later Nobel Prize winner, 1978) mathematically developed and presented the ideas of Yule in what is usually called the *Simon model* [163]. Numerous scale-free distributions were interpreted using this model. It is just common sense that the rich attract more wealth, a popular girl attracts more looks, a failure attracts more misfortunes. One may say: 'money attracts money', 'popularity is attractive', and so on. According to Yule and Simon, processes underlying these truisms may result in power-law distributions. In general terms, the mechanism of the emergence of power-laws in these processes is *self-organization*. While growing, systems of this type self-organize into states with scale-free statistics. Sometimes physicists treat these states as 'critical'.

A few 'premature' works of the 1970s and 1980s considering these ideas in the context of networks had no immediate impact.[5] It was the paper of Barabási and Albert that initiated a wave of studies exploiting this mechanism. The Barabási–Albert model only demonstrates the essence of the preferential attachment mechanism. To approach real-world networks, further steps are needed. The first natural step is to find a way to generate scale-free networks with an arbitrary degree distribution exponent and not only 3. For this it is sufficient to change slightly a rule of preferential attachment. Let the probability of attaching to a node of degree q be proportional to a function $f(q)$ which is called a *preference function*. In the Barabási–Albert model, the preference function is q, that is the proportional preference. Let us change in this model just the preference function. Let it now be not proportional but linear, $q + A$. Here $A \equiv am$ plays the role of an 'additional attractiveness', a is a constant number, m is the number of links of each new node. We assume $a > -1$. It turns out that this linear preference leads to exponent

$$\gamma = 3 + a\,, \tag{7.3}$$

which can take values in the range from 2 to infinity [81].[6] Note that the limiting case $a \to \infty$ corresponds to a random recursive graph with uniform attachment.

One can obtain the same result as with linear preference, mixing proportional and uniform preferential attachment. With some probability, make connections as in the Barabási–Albert model and with the complementary probability attach to uniformly randomly selected nodes. These straightforward generalizations provide far more realistic networks than the Barabási–Albert model, but they are not more difficult to work with. For numerical simulations, researchers usually need to generate very large networks, say of 10^7 nodes or more. This can be done easily if a graph is recursive and preference is linear.[7] Still, there is a natural question: what is the origin of the proportional or linear attachment? Indeed, preferential attachment explains power laws, but what can explain preferential attachment? Unfortunately, nobody has found a completely satisfactory, general answer up to now, and this is the weakest point of the self-organization (or self-organized criticality) concept for networks.

[5] In bibliometrics, the model of Derek de Solla Price (1976) based on preferential choice explained the statistics of growing directed networks of scientific citations [153]. In graph theory, Szymański defined recursive trees with proportional preferential attachment [174], which actually coincide with the particular case $m = 1$ of the Barabási–Albert model.

[6] Similarly, one can grow scale-free directed networks. Suppose, for example, that in the recursive network which we discussed, all links are directed from new to old nodes. Suppose that the attachment probability is a linear function of in-degrees, $q_{\rm i} + \tilde{a}m$, $\tilde{a} > 0$. Then the exponent of the in-degree distribution is $\gamma = 2 + \tilde{a}$.

[7] For generating these networks, it is inefficient to select nodes by using their degrees. Indeed, for each attachment, we would have to examine all the degrees. Instead, form a list where each node is repeated as many times as its degree (or, if you want, degree minus m). Selecting uniformly an entry from this list, we get proportional preference. To get linear preference, form in addition the relevantly weighted (with weight a) list of all nodes, and choose uniformly from the combination of these two lists. After every attachment, add new entries to these lists.

One of the possible explanations is a process based on uniform selection. In other words, we should arrive at the proportional (or linear) preference using only a uniform choice. Recall that in an arbitrary network the degree distribution of an end node of a uniformly randomly chosen link is $qP(q)/\langle q\rangle$. So, in principle, choosing a link at random and then its end, we get the proportional preference. The problem is that we have to choose from the set of all links, for which we need to know all of them. There is a better option. To explain, let us consider a network of scientific citations and look how references emerge in scientific papers. A researcher either learns about a paper, which he will cite, directly from journals and archives, or he finds a reference to it in another article. The second option results in preferential attachment. Note, however, that choosing one nearest neighbour of a uniformly selected node, we get the proportional preference only if a network is uncorrelated, which is usually not the case. For an arbitrary network, we have to do the following. Choose a node uniformly and select each of its nearest neighbours with some given nonzero probability.[8] It is easy to show that this gives the proportional preference. The reader may check this claim using any small graph. Here, application to the networks of scientific citations is not unique. Actually the same process drives the evolution of genetic networks [143]. In cellular biology, this process is called *duplication–divergence*. In protein–protein interaction networks,[9] new proteins are born as identical copies of the original ones (duplication), but afterwards new proteins lose some of their functions (divergence), see Fig. 7.2. (In addition, they acquire new functions but this can be treated separately.) In network terms, duplication–divergence means precisely the connection of a new node to a fraction of the nearest neighbours of a selected node. Thus we again arrive at a proportional preferential attachment.

The preferential attachment mechanism can also be applied to equilibrium networks, but in this section our focus is only on recursive graphs. Nodes in these networks are distinguished (labelled) according to their birth times. It is obvious that older nodes are better connected. Indeed, the mean degree of a node born at time u was found to be $\langle q(u,t)\rangle \sim (u/t)^{-\beta}$ [81]. Here exponent β is related to the exponent of the degree distribution: $\beta = 1/(\gamma - 1)$. Thus we have two related power laws. In comparison with uniform random recursive graphs, the linear preferential attachment sharply changes the statistics of connections of a node born at a given time. Recall that in the model with uniform attachment, fluctuations of $q(u,t)$ are very weak. In contrast to this, due to linear preferential attachment, the degree $q(u,t)$ of node u fluctuates strongly. The fluctuations are of the order of the mean value $\langle q(u,t)\rangle$.

Power laws in these recursive networks have been observed not only for degrees and degree distributions. We have already mentioned that in a wide range of recursive trees (with and without preferential attachment) the distribution of the number of ancestors of a node decays as a power law with exponent 2. Similarly, it was found that in all these recursive trees, the betweenness centrality distribution is a power law with the

[8] Note that you have a chance to select simultaneously several nearest neighbours, or zero. So this process effectively generates short loops.

[9] Nodes in these networks are proteins, while links indicate functional and other associations between protein pairs.

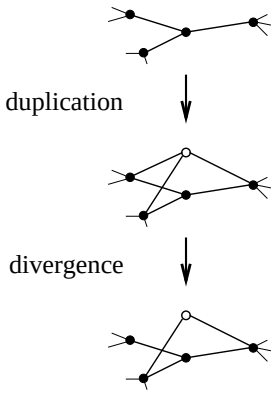

Fig. 7.2 A duplication–divergence event. The white node is a new protein.

same exponent 2 [97].[10] In these networks—recursive trees with linear preferential attachment—it is even possible to find distributions with exponent less than 2. One example is the size distribution for branches of the root. Let the exponent of the degree distribution of this recursive tree be γ. Then the mean number of branches of size s decays as $s^{-\gamma/(\gamma-1)}$. Exponent γ varies in range between 2 and infinity, so $\gamma/(\gamma-1)$ takes values in the range from 2 to 1.

The growth of a network starts from some initial configuration of nodes and links. What is its role? Can it influence the evolution of a network, in particular, the form of its degree distribution? In principle, the answer depends on the network model, but for recursive networks, the situation is clear. Figure 7.3 shows a typical degree distribution of a finite scale-free recursive network. As is natural, a power law describes the distribution only in the intermediate region, between low and high degrees. The cut-offs of the degree distributions in these networks increase with size N: $q_{\rm cut} \sim N^{1/(\gamma-1)}$. Note a bulge near the cut-off in the figure which is visible at any network size. This hump is just the trace of an initial condition—the first nodes will always remain best connected. By changing an initial configuration, we can increase, or diminish, or sometimes even wipe out the hump.

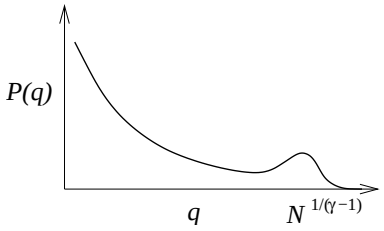

Fig. 7.3 Typical degree distribution of a finite scale-free recursive network.

We have provided some reasons why linear preference can emerge, but, in general, a preference function can be nonlinear. Suppose, for example, that a preference function is a power law, q^ϕ, with an arbitrary positive exponent ϕ. Can it produce scale-free distributions? Pavel Krapivsky and Sid Redner showed this is possible only if $\phi = 1$ [114,115]. When the function grows slower than q, that is $\phi < 1$, the resulting degree distribution decays slower than exponentially but faster than a power law. The distribution has a so-called stretched exponential form: $P(q) \propto \exp(-Cq^{1-\phi})$, where the value of C is determined by exponent ϕ. If, on the other hand, the preference function grows faster than q, that is $\phi > 1$, we again fail to get a scale-free distribution. In this case, all links in the growing network become seized by a tiny fraction of the oldest nodes or even by a single node. Thus, the preferential attachment mechanism generates power-law distributions only if preference is linear, which is a serious restriction.

In this lecture we have discussed only a few basic networks demonstrating the essence of the preferential attachment mechanism. Plenty of more detailed and realistic network models based on this mechanism have been considered. We stress that apart from preferential attachment, there are other ways, maybe not so extensively explored in networks, which allow one to generate power-law distributions. We will touch upon some of these alternative ways in the following lectures. Here we will mention only one of the alternative processes. Suppose that the evolution of a network incorporates the merging of nodes. Say, at each time step, a pair of some nodes merge together, and so the corresponding degrees are added together. It turns out that this merging can result in a variety of scale-free degree distributions [109].

[10] The coincidence of these exponents is not occasional. One can show that these two distributions are closely related.

7.4 Condensation phenomena

Discussing preferential attachment, we have focused on growing networks. Note, however, that growing systems are only a particular case of nonequilibrium ones. One can easily construct a nonequilibrium network with a constant number of nodes and links.[11] Main qualitative conclusions for the particular case of growing networks with preferential attachment are valid for nonequilibrium networks. On the other hand, it is more difficult to arrive at power-law degree distributions for equilibrium networks by using preferential attachment. It is time to discuss these networks in more detail.

[11] Consider, for example, the following process generating a non-equilibrium network of a fixed size. At each time step, reattach all the emanating links from a randomly chosen node to some other specially selected nodes. Clearly, the statistics of connections of a node in this network depends on the time elapsed after the reattachment.

Let us implement preferential attachment to an equilibrium net. For example, at each time step, choose a link uniformly from the set of all links and reattach one of its ends to a preferentially chosen node. As is natural, the resulting network architecture depends on a preference function. As in growing networks, we must use a linear preference to arrive at scale-free architecture. For equilibrium networks, this is, however, not sufficient. It turns out that the organization of connections is essentially determined by the mean degree of nodes in a network.

(i) If the mean degree is below some critical value, which is determined by the form of a preference function, then the degree distribution is rapidly decaying.

(ii) At the critical value, the degree distribution is power-law.

(iii) Above the critical value, one of the nodes attracts a finite fraction of all connections in the network. The connections of the rest of the nodes are described by a power-law degree distribution with the same exponent as at the critical point, see Fig. 7.4.

This condensation of links frequently occurs in the models of complex networks. As far as we know, condensation has never been observed in real-world networks. The difficulty is that for observation, a network must be very large. Otherwise the hub—a node attracting a finite fraction of connections—cannot be reliably distinguished from other strongly connected nodes.

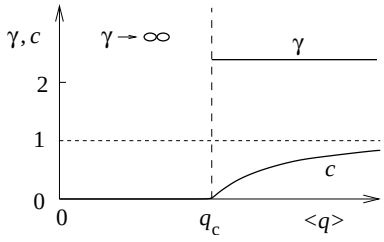

Fig. 7.4 Condensation of links in an equilibrium network. Fraction c of links attached to a single node and the exponent γ of the degree distribution of the rest of the nodes versus the mean degree $\langle q \rangle$. Below the critical value q_c, the degree distribution decays exponentially.

The same equilibrium network, with condensation of links and a power-law degree distribution, can be constructed by other means, without using the preferential attachment. Simply give higher weights to those statistical ensemble members that have many strongly connected nodes [45]. More generally, one can introduce the 'energy' of a graph—a quantity which is determined by some structural characteristics (for example, the total number of triangles—transitivity—in this graph) and then implement Boltzmann statistics to the ensemble of the graph. So here a given form of the 'energy' functional defines a network. After David Strauss [170], random networks constructed in this, natural for statistical mechanics way, are usually called *exponential random graphs*. Importantly, these networks easily develop condensation: of links, triangles, and other structural units, depending on the definition of 'energy'. Models of this kind are usually simulated based on the so-called *Metropolis*

algorithm.[12] Usually these simulations start from a homogeneously random configuration, and a final condensed state is approached only after a series of steps. In small networks, this transition is rapid, but, remarkably, it takes an astronomical time even if a network is moderately large [46].

Condensation also occurs in growing networks. Here we consider one of the possibilities. Let us return to recursive networks with a linear attachment, that is, the preference function is $f(q) = hq + a$. We assumed previously that the parameters h and a of the preference do not depend on a node.[13] In this respect, all nodes were assumed equal. In real-world networks, this is apparently not the case—nodes are very different and have different potentials for growth. Compare Google and your home page. To model this diversity, Ginestra Bianconi and László Barabási assumed that the parameters a and h vary from node to node [28]. A 'young' node with large a and/or h has a good chance to quickly attach many links and become more connected than older nodes with low parameters a and h. If fluctuations of the preference parameters are low, the effect of this heterogeneity is not very strong—the exponent of a degree distribution changes. If, however, the fluctuations of a and, especially, h, are sufficiently strong, the heterogeneity changes qualitatively the organization of a network. A finite fraction of connections become attracted by one of the nodes [29]. The degree distribution for the rest of the nodes is scale-free. For simplicity, assume that only one node in a network is distinct from the other, namely that this node has $h > 1$, while others have $h = 1$. Figure 7.5 for this particular case shows how the fraction c of condensed links and the exponent γ of the degree distribution depend on h (compare with Fig. 7.4 for condensation in equilibrium networks).

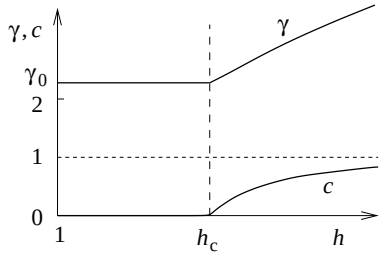

Fig. 7.5 Condensation of links in a growing network in which only one node has the preference function parameter $h > 1$ while others have $h = 1$, see book [80]. Fraction c of condensed links and the exponent γ of the degree distribution versus h. In the homogeneous network ($h = 1$ for all nodes) the value of the degree distribution exponent is assumed to be $\gamma_0 > 2$.

7.5 Accelerated growth

We have already mentioned that in many real-world networks the number of links grows faster the number of nodes, so that the mean degree of a node is a growing function of the network size [120]. This nonlinear, or, one can say, 'accelerated', growth has numerous consequences [78, 79].

We indicated that, for example, the diameter of a growing network of this kind may be constant or even decrease as the number of nodes increases. The models of networks that we have considered up to now do not show this accelerated growth. It is not very difficult, however, to incorporate this feature into network models with preferential attachment. Suppose, for example, that in a random recursive network, the number of connections of a new node $m(N)$ is a growing function of N. Then the mean degree of a node, $\langle q \rangle(N) = 2m(N)$, is also growing. With the acceleration, a network can still be scale-free, if the preferential attachment is linear. There are some differences compared to previously discussed degree distributions. Without the acceleration, a scale-free degree distribution is stationary in the limit $N \to \infty$, and its exponent

[12] This is one of the standard algorithms for physics and optimization problems [102]. In application to spin systems ($S_i = \pm 1$), the algorithm works as follows. At each time step, select at random a spin, say, in state S_i and compare the energies of two configurations—with S_i and with $-S_i$. If the latter has a lower energy than the former, then flip the spin: $S_i \to -S_i$. Otherwise pass to the next step, and so on until the system approaches an equilibrium state.

[13] In this situation, we could use $f(q) = q + a/h$.

γ cannot be smaller than 2, otherwise the first moment of a degree distribution would diverge. This is the case in most of the studied models with preferential attachment. If the growth is accelerated, both these restrictions are removed: degree distributions may be non-stationary, and γ may occur between 1 and 2. Low values of this exponent have indeed been observed in a number of real networks of this kind. For example, γ is below 2 in the degree distributions of large software maps [178].

7.6 The BKT transition

In each of the growing networks discussed above, all nodes were within a single connected component. More generally, growing networks, similarly to equilibrium ones, can consist of a rich set of connected components. How are they organized in a growing network and what is their difference from equilibrium networks?

For demonstration, consider the following network [49] with numerous connected components. The growth process consists of two channels: bare nodes are added at unit rate and links connecting uniformly chosen nodes are added at rate b. The rate for links, b, plays the role of a control parameter which determines the structure of the network. Clearly, the mean degree of a node in this network is $\langle q \rangle = 2b$. One can easily show that this growing random graph has an exponentially decaying degree distribution. If $b = 0$, the network is a set of bare nodes. With increasing b, the network contains bigger and bigger components. If the parameter b is greater than some critical value b_c, then the growing network has a giant connected component consisting of a finite fraction S of all nodes. The remarkable finding was that this giant component emerges in a drastically different way from that in equilibrium networks [49]. Namely, near the birth point, the relative size of a giant component is

$$S \propto \exp\Big(-\frac{D}{\sqrt{b-b_c}}\Big), \tag{7.4}$$

where D is a positive constant. So, any derivative $d^n S(q)/dq^n$ ($n = 0, 1, 2, \ldots$) is zero at the critical point. In standard classification, this is an infinite-order transition. Compare this singular function of b to the critical singularities of S in equilibrium networks, which we listed in the preceding lecture. In condensed matter physics, this so-called *Berezinskii–Kosterlitz–Thouless* (BKT) singularity is considered to be rather exotic. For theoretical physicists, it was surprising to observe this anomaly in small worlds. Similar critical features were later observed in many other growing networks, including scale-free [82].

There is another marked difference. In the absence of a giant component in these networks, the size distributions $\mathcal{P}(s)$ of finite components decay in a power-law fashion. In physics, a power-law decay often indicates a critical state. So in these networks, the entire phase without a giant component can be called 'critical', and not only its birth point.

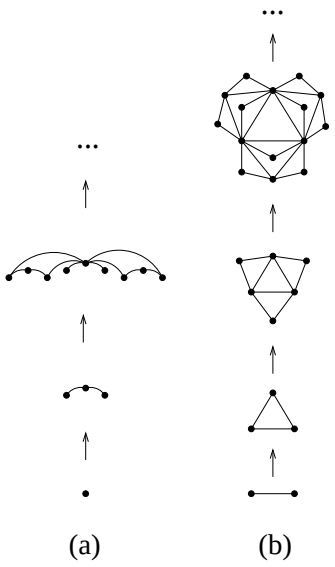

Fig. 7.6 Two of the constructions of scale-free deterministic graphs [16, 74].

7.7 Deterministic graphs

In this lecture we have mostly discussed growing networks distinguished by three features: (i) the small-world feature, (ii) a scale-free architecture, (iii) randomness. It is possible, however to build more simple non-random scale-free small worlds, in other words, deterministic scale-free graphs. Two examples, shown in Fig. 7.6, explain this construction. The diameters of these graphs grow logarithmically ($\ln N$) with their size, typically for small worlds but in sharp contrast with fractals (see Fig. 1.5). The degrees of deterministic graphs take a discrete set of values in contrast to random networks. On the other hand, the envelope of the occurrence frequency of these degrees is often power-law, in particular for the graphs in Fig. 7.6. In that sense, these graphs are scale-free. Surprisingly, many of their properties are very close to those of random scale-free networks. So these toy models are closer to real-world networks that one would expect.

Correlations in networks

8

In Lecture 5 we considered network models without structural correlations. However, real-world networks are correlated. In this lecture we discuss the simplest correlations in complex networks and explain their nature.

8.1 Degree–degree correlations

Among a wide spectrum of various structural correlations in networks, the simplest are correlations between degrees of nearest-neighbouring nodes. These correlations are described by the joint distribution of these degrees, $P(q, q')$. Usually, it is called a *degree–degree distribution*. This is the probability that a randomly chosen link connects nodes of degree q and q'.[1] Recall that in an uncorrelated network, this degree–degree distribution is determined by the degree distribution $P(q)$ of the network, namely, $P(q, q') = qP(q)q'P(q')/\langle q\rangle^2$. This is not true, if correlations are present. In an arbitrary network, the degree distribution of an end node for a uniformly chosen link is $qP(q)/\langle q\rangle$. So $P(q)$ is expressed in terms of a degree–degree distribution:

$$\sum_{q'} P(q, q') = \frac{qP(q)}{\langle q\rangle}\,. \tag{8.1}$$

Instead of the joint degree–degree distribution, equivalently one can use the conditional probability $P(q|q')$ that if one end node of a link is of degree q', then its second end node is of degree q.[2]

Let us take one step beyond uncorrelated networks defining an equilibrium network model with only correlations between degrees of nearest neighbours. In the spirit of the configuration model, this is the maximally random network with a given joint degree–degree distribution $P(q, q')$. Thus we have a hierarchy of equilibrium network models.

(i) The classical random graphs are the lowest level of the hierarchy. These are maximally random networks under a single constraint: $\langle q\rangle$ is fixed.

(ii) The uncorrelated networks are the second level. These are maximally random networks with a given degree distribution $P(q)$. As a consequence of this constraint, a mean degree $\langle q\rangle$ is also fixed.

(iii) Maximally random networks with a given $P(q, q')$ are the third level. This constraint ensures that $P(q)$ is also fixed, and then $\langle q\rangle$ is fixed as well.

[1] How can we obtain the degree–degree distribution from empirical data? Let the number of links in a graph be L. Find the number of links $L(q, q')$ connecting nodes of degree q and q'. If $q \neq q'$, then $P(q, q') = L(q, q')/2L$, and if $q = q'$, then $P(q, q) = L(q, q)/L$.

[2] One can see that

$$P(q|q') = \frac{P(q, q')}{\sum_q P(q, q')} = \bar{q}\frac{P(q, q')}{q'P(q')}\,.$$

The symmetry $P(q, q') = P(q', q)$ immediately leads to the following symmetry relation for the conditional probability [36]:

$$qP(q'|q)P(q) = q'P(q|q')P(q')\,.$$

One can continue this hierarchy by imposing more and more rigid constraints which leads to more complex multi-degree correlations. Unfortunately, already, the third level of hierarchy is sufficiently difficult for analysis.

8.2 How to measure correlations

The complete information about degree–degree correlations in a real-world network can be obtained only by measuring its joint degree–degree distribution $P(q,q')$ or $P(q|q')$. Alas, this is practically impossible in sparse networks with heavy-tailed degree distributions. Suppose the number of links in a network is $L \sim N$ and the cut-off of the degree distribution is at $q_{\rm cut}$. To arrive at a usable $P(q,q')$ with sufficiently small fluctuations, we must have a large number $L(q,q')$ of links connecting nodes of degrees q and q', for each pair of degrees. This is impossible if L is less or of the order of $q_{\rm cut}^2$, which is the case if a network is sparse and $q_{\rm cut}$ is large. To avoid strong fluctuations in empirical data, researchers have to use a less informative though convenient characteristic, namely the average degree of the nearest neighbour of a node as a function of the degree of this node [148]:

$$\overline{q}_{\rm nn}(q) = \sum_{q'} q' P(q'|q)\,. \tag{8.2}$$

We mentioned this quantity while discussing real-world networks. This dependence relates the degree of a node and the average degree of its nearest neighbours and allows one to find how nodes of different degrees are interconnected. If a network is uncorrelated, $\overline{q}_{\rm nn}(q) = \langle q^2\rangle/\langle q\rangle$, that is a constant. In correlated networks, dependences $\overline{q}_{\rm nn}(q)$ can be very diverse in shape. There is some restriction. It turns out that the second moment of the degree distribution and $\overline{q}_{\rm nn}(q)$ are related [37].[3] Namely,

$$\langle q^2\rangle = \sum_{q'} q' P(q') \overline{q}_{\rm nn}(q')\,. \tag{8.3}$$

If, in particular, the second moment of the degree distribution diverges, $\overline{q}_{\rm nn}(q)$ cannot be arbitrary. Its form must guarantee the divergence of the right-hand side of relation (8.3). If so, then what are the possible dependencies $\overline{q}_{\rm nn}(q)$?

[3] This relation directly follows from the definition of $\overline{q}_{\rm nn}(q)$.

8.3 Assortative and disassortative mixing

Sociologists usually characterize degree–degree correlations by an even more rough quantity than $\overline{q}_{\rm nn}(q)$. This is *the Pearson correlation coefficient* defined as

$$r = \frac{\langle qq'\rangle_l - \langle q\rangle_l \langle q'\rangle_l}{\langle q^2\rangle_l - \langle q\rangle_l^2}\,. \tag{8.4}$$

Here q and q' are the degrees of the end nodes of a link, and $\langle\ \rangle_l$ denotes the average over all links in a network.[4] The Pearson coefficient is a

[4] Equivalently, one can substitute $q-1$ and $q'-1$ in this definition for q and q'. $\langle qq'\rangle_l$ is the average product of the degrees of the end nodes of a link. $\langle q\rangle_l$ is the average degree of an end node of a link. As we explained, $\langle q\rangle_l = \langle q^2\rangle/\langle q\rangle$ for an arbitrary network. Similarly, the average square of the degree of an end node of a link, $\langle q^2\rangle_l$, equals $\langle q^3\rangle/\langle q\rangle$. Expression (8.4) is not very convenient for the computation of r from empirical data. See [134] for a more convenient form of this definition.

standard pair correlation function $\langle qq'\rangle_l - \langle q\rangle_l\langle q'\rangle_l$ normalized in such a way that r is in the range between -1 and 1. Clearly, if a network is uncorrelated, then the Pearson coefficient is zero. If, on average, strongly connected nodes have strongly connected neighbours, then $r > 0$. These correlations are called *assortative*, and this situation is referred to as *assortative mixing* [134]. In the opposite situation, on average, a weakly connected node has a strongly connected neighbour and vice versa. This is called *disassortative mixing*, and for these correlations, the Pearson coefficient is negative. Figure 8.1 shows examples of networks with ultimately strong assortative and disassortative mixing, $r = 1$ and -1, respectively.

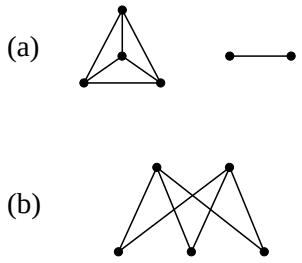

Fig. 8.1 Two small networks with ultimately assortative (a) and disassortative (b) mixing. In network (a), only nodes of equal degree are interlinked. In network (b), only nodes of different degrees are interlinked. Check that r equals 1 and -1 for (a) and (b), respectively.

We can distinguish between assortative and disassortative correlations in another way by inspecting the dependence $\overline{q}_{\rm nn}(q)$, see Fig. 8.2 (a). Clearly, if this curve is monotonously growing, then the degree–degree correlations are assortative, if it is monotonously decreasing, then the correlations are disassortative. The problem is that the curve $\overline{q}_{\rm nn}(q)$ is often not monotonous as in Figs. 8.2 (b) and 4.4, and then clear distinction between the two kinds of correlations is impossible. Furthermore, sometimes, the Pearson coefficient is itself confusing. To see this, we rewrite expression (8.5) using $\overline{q}_{\rm nn}(q)$. This gives

$$r = \frac{\langle q\rangle \sum_q q^2 \overline{q}_{\rm nn}(q) P(q) - \langle q^2\rangle^2}{\langle q\rangle\langle q^3\rangle - \langle q^2\rangle^2}\,. \qquad (8.5)$$

Note the third moment of the degree distribution, $\langle q^3\rangle$, in the denominator. In infinite scale-free networks, $\langle q^3\rangle$ diverges if exponent γ is less or equal to 4. So if the numerator is finite, we get zero Pearson's coefficient in an infinite network. It turns out that this is the case only if degree distribution exponent γ is between 3 and 4. In this situation for finite networks, r strongly depends on network size. So, it is practically impossible to compare the values of the Pearson coefficient for different networks and, for example, determine which one of them is 'more assortative' [73]. Only if exponent γ is outside of this $(3, 4)$ interval, is r finite in infinite networks and can it be used for the characterization of correlations.[5] Furthermore, zero Pearson's coefficient does not guarantee the absence of degree–degree correlations. In principle, some non-monotonous $\overline{q}_{\rm nn}(q)$ substituted in relation (8.5), may give $r = 0$.

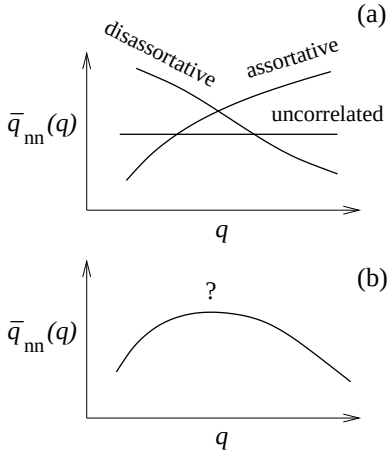

Fig. 8.2 (a) Typical curves $\overline{q}_{\rm nn}(q)$ for various kinds of degree–degree correlations. (b) It is impossible to conclude whether the mixing is assortative or disassortative by inspecting this non-monotonous dependence.

These drawbacks of r seriously hamper comparison of empirical data for different networks with heavy-tailed degree distributions. Nonetheless, the Pearson correlation coefficient is always listed among main network characteristics, providing information at least about the type of degree–degree correlations. For example, practically all empirical data on social and collaboration networks indicate strong assortative correlations (see a large survey of empirical data on networks [66]). That is, a sociable person typically has sociable friends, while (very few) friends of an introvert are also unsociable. On the other hand, many technological networks demonstrate disassortative mixing. In particular, the Internet on the Autonomous Systems level has these kinds of correlations [150]. However, the network of routers in the Internet does not have such clear

[5] If $\gamma < 3$, in infinite networks, both the numerator and denominator on the right-hand side of eqn (8.5) diverge, and these divergences compensate each other resulting in a finite r. The divergence of the numerator is explained by relation (8.5). It was found that $\overline{q}_{\rm nn}(q)$ is strongly size-dependent in this range of exponent γ [37].

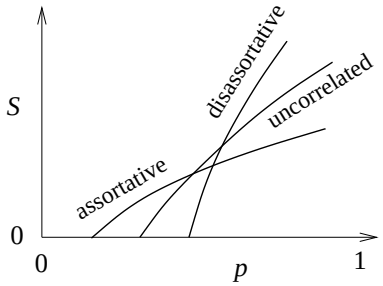

Fig. 8.3 The effect of degree–degree correlations on the emergence of a giant connected component. The size of a giant component versus the retained fraction p of nodes after damaging the network.

degree–degree correlations, see Fig. 4.4. The WWW and the network of protein interactions have disassortative degree–degree correlations.

What are the consequences of degree–degree correlations? To be specific, let us discuss how degree–degree correlations influence global organization of a network, namely a giant connected component. Clearly, we should compare correlated networks with uncorrelated ones having identical degree distributions. Consider, for example, the percolation problem. That is, remove a fraction $(1-p)$ of uniformly chosen nodes and inspect how this affects a giant connected component. We assume that the degree–degree correlations are sufficiently weak, so that they can only moderately change the organization of a network. In particular, these weak correlations cannot modify the ultraresilience criterion obtained for uncorrelated networks. If the second moment of a degree distribution is infinite in an infinite correlated network (which corresponds to exponent $\gamma \leq 3$), then the percolation threshold is still $p_c = 0$. The effect of correlations on the percolation threshold is visible if a degree distribution decays sufficiently rapidly, $\gamma > 3$. In this situation, assortative correlations diminish p_c from its value for an uncorrelated network, that is they make the network more resilient against random damage [134]. Disassortative mixing results in the opposite effect.

At first sight, enhancement of resilience due to assortative correlations is quite obvious. Indeed, thanks to these correlations, nodes of high degrees are better interconnected with each other than in uncorrelated networks. This strongly interconnected core part of a network belongs to the giant connected component (if it exists, of course). To eliminate the giant component, one has to split this strongly interconnected core, which is not an easy task. Nonetheless, the effect of correlations is not that simple. Consider the influence of correlations on a full curve—size of a giant component versus p. Figure 8.3 shows that assortative correlations indeed decrease p_c, making it harder to eliminate a giant component, but this is not all. Surprisingly, these correlations may also suppress the size of a giant component [134]. Thus, these two contrasting effects strangely coexist with each other.

8.4 Why are networks correlated?

Real-world networks are correlated. Even if empirical data indicate the absence of correlations in some real-world networks, this usually only means that correlations are weak or that an unsuitable or too crude quantity (e.g. r) was studied. Importantly, all growing network models are correlated. In the growth process, the presence of links between nodes depends on ages and degrees of these nodes, and there is an asymmetry between younger and older nodes. So, correlations are practically inevitable. The type of correlations depends on the model.

Degree–degree correlations are present even in uniform random recursive graphs. Consider more general, scale-free random recursive graphs which are characterized by a degree distribution exponent γ. In these

networks, if $\gamma < 3$, then, asymptotically,

$$\overline{q}_{\rm nn}(q) \sim q^{-(3-\gamma)}\,, \tag{8.6}$$

which is typical for disassortative mixing [115]. If $\gamma > 3$, the slope of $\overline{q}_{\rm nn}(q)$ is opposite, which indicates assortative mixing. The Barabási–Albert model is unique in the sense that at $\gamma = 3$, mixing changes from disassortative to assortative.[6] This is one of the reasons to use less special models with linear attachment and not the Barabási–Albert model.

In some equilibrium networks, degree–degree correlations are also inevitable. It turns out that additional structural constraints lead to correlations [123]. In particular, the banning of multiple links is a constraint of this kind. Let us discuss a uniformly random graph with a given degree distribution, but in contrast to the configuration model, now we forbid multiple links and loops of length 1. In this constrained model, the form of the degree distribution is particularly significant. In practice, this network may be built by using the following randomization algorithm [123]. (i) Create an arbitrary graph of a given size, with a given degree sequence, and without multiple connections. (ii) Rewire a pair of randomly chosen links in a way shown in Fig. 8.4. This retains the degrees of all nodes and does not create multiple links. (iii) Repeat (ii) until the network relaxes to an equilibrium state.[7]

A difference between the configuration and the constraint models exists only if a network has hubs of degrees of the order of $\sqrt{N}$ or greater.[8] In the configuration model, these hubs have a good chance of being multi-linked and of having 1-loops. Nodes with such high degrees exist only if the degree distribution slowly decays. Exponent γ must be smaller than 3. Furthermore, the cut-off of the degree distribution must grow more rapidly than $\sqrt{N}$. If we forbid multiple links, the excessive links between multiply connected nodes should be rewired. This (i) reduces the probability that high-degree nodes are interlinked, and (ii) increases the probability that the highly connected nodes are linked to weakly connected ones. Consequently, this 'constrained' network has disassortative degree–degree correlations. This also leads to a rough estimate $q_{\rm cut} \sim N^{1/2}$ for an uncorrelated network without multiple links, if exponent $\gamma \leq 3$. There cannot be nodes with higher degrees in networks of this kind.

By similar reasoning, strongly connected nodes (with degrees of the order of or larger than $N^{1/2}$) are strongly interlinked between each other. A network is sparse, but its subgraph based on these 'rich' nodes is dense. As was mentioned, this is called a rich-club phenomenon [185]. Remarkably, the rich-club phenomenon is possible in a wide range of networks: correlated and uncorrelated, with and without multiple links. This phenomenon was observed in many real-world networks with slowly decaying degree distributions [63]. For example, the tier-1 Autonomous Systems in the Internet form a rich club.

Although the rich club is relatively small, it plays an important role in a network. It belongs to a giant connected component (if it exists) and form its 'core'. This part of a network is vitally important for its

[6] For $\gamma \geq 3$, $\overline{q}_{\rm nn}(q)$ changes logarithmically slowly [20]. Interestingly, although these networks are correlated at any γ, their Pearson coefficient is zero when a network is infinite [73].

[7] Here we only outline the idea of the algorithm.

[8] For simplicity, estimate the mean number of 1-loops at a node of degree $q \gg 1$ in the configuration model. Use the construction shown in Fig. 5.1. Each of the q stubs joins one of the remaining stubs at this node with probability $(q-1)/\langle q\rangle N$. This gives, on average, about $q^2/\langle q\rangle N$ 1-loops at this node.

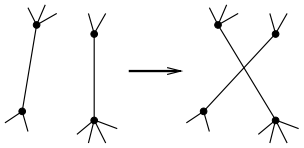

Fig. 8.4 Rewiring operation for the randomization algorithm [123]. This rewiring preserves degrees of nodes and does not create multiple links.

resilience against random damage. Indeed, the rich-club phenomenon is observed for the same slowly decaying degree distributions as the ultra-resilience. Therefore one can significantly improve the robustness of a network by enhancing (or introducing) the rich-club phenomenon. For this, add new links between nodes of high degrees.

Remarkably, correlations may be found even within uncorrelated networks, in particular, within the standard configuration model. Suppose that an uncorrelated network has a finite fraction of dead ends, and so, as we explained earlier, its giant connected component is smaller than the network. Then, counterintuitively, this giant component is correlated.[9]

[9] To see the reason for this, let us assume the opposite, i.e., that the component is itself an uncorrelated network. Since the giant component also apparently has a finite fraction of dead ends, we conclude that this component should contain finite connected components, which is false. Therefore, our assumption was wrong. There is another explanation. An uncorrelated network with dead ends necessarily contains 'dimers'—separated pairs of connected nodes. The giant component of this network has dead ends but no dimers, so it must be correlated.

8.5 Degree correlations and clustering

All networks belonging to the network constructions of Section 8.1 are locally tree-like if a degree distribution decays sufficiently rapidly. In these equilibrium network models, the clustering coefficient vanishes as $1/N$. High concentrations of triangles can be found only in a (small) rich club, if it exists in a network. Suppose that a network has only degree–degree correlations, $P(q, q')$. This distribution determines the clustering coefficient. One can directly generalize formula (5.2) obtained for uncorrelated networks.[10] The key difference is that in correlated networks, the clustering becomes degree dependent. As in real networks, the resulting mean clustering $\overline{C}(q)$ of a node depends on its degree, and so the clustering coefficient does not coincide with the mean clustering, $C \neq \overline{C}$.

[10] See review [77] and references therein.

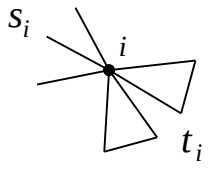

Fig. 8.5 Newman's generalization of the configuration model [137]. For each node i in this network, two numbers are given: the number t_i of triangles to which this node belongs and the number s_i of its rest connections. Note that the triangles have no joint sides.

Clustering in networks, constructed in this way, is only a finite size effect. Mark Newman, however, proposed a direct generalization of the configuration model, which has a finite clustering coefficient even if the network is infinite [137]. His network is made of two simple non-overlapping motifs (frequently repeated subgraphs): triangles (3-cliques) which have no joint sides and single links (2-cliques) which do not belong to any of the triangles. Instead of a given degree sequence, consider a sequence of pairs of numbers. For each node i two numbers are given—the number of triangles to which this node belongs, t_i, and the number of the remaining connections (single links), s_i, see Fig. 8.5. Importantly, all the triangles have no joint sides. So the degree of the node i is $q_i = s_i + 2t_i$. Newman's network is defined as a uniformly random graph with a given $\{s_i, t_i\}$ sequence. Therefore this network is a superposition of two 'configuration models': for single links and for triangles.[11] One can show that, remarkably, this network has degree–degree correlations.

[11] Check that local clustering in this model cannot be too strong:

$$C_i \leq \frac{q_i/2}{q_i(q_i-1)/2} = \frac{1}{q_i - 1}\,.$$

If we need a clustered network of this kind for numerical simulation purposes, it is more convenient to use a hidden variable construction. Recall our discussion of hidden variables in Lecture 5. In this approach, $\{s_i, t_i\}$, i=1,...,N, is a given sequence of desired numbers of single links and triangles. Using these numbers, connect pairs of nodes (i, j) with probability $p_{ij} \propto s_i s_j$, and interconnect triples of nodes (i, j, k) with probability $p_{ijk} \propto t_i t_j t_k$. Instead of triangles, we could use more complicated motifs—for example, long loops or large cliques, and so on.