CHAPTER 8

THE LARGE-SCALE STRUCTURE OF NETWORKS

A discussion of some of the recurring patterns and structures revealed when we apply the concepts developed in previous chapters to the study of real-world networks

IN PREVIOUS chapters of this book we have looked at different types of natural and man-made networks and techniques for determining their structure (Chapters 2 to 5), the mathematics used to represent networks formally (Chapter 6), and the measures and metrics used to quantify network structure (Chapter 7). In this chapter we combine what we have learned so far, applying our theoretical ideas and measures to empirical network data to get a picture of what networks look like in the real world.

As we will see, there are a number of common recurring patterns seen in network structures, patterns that can have a profound effect on the way networked systems work. Among other things, we discuss in this chapter component sizes, path lengths and the small-world effect, degree distributions and power laws, and clustering coefficients.

8.1 COMPONENTS

We begin our discussion of the structure of real-world networks with a look at component sizes. In an undirected network, we typically find that there is a large component that fills most of the network—usually more than half and not infrequently over 90%—while the rest of the network is divided into a large number of small components disconnected from the rest. This situation is sketched in Fig. 8.1. (The large component is often referred to as the "giant component," although this is a slightly sloppy usage. As discussed in Section 12.5, the words "giant component" have a specific meaning in network theory and are not precisely synonymous with "largest component." In this book we will be careful to distinguish between "largest" and "giant.")

A typical example of this kind of behavior is the network of film actors discussed in Section 3.5. In this network the vertices represent actors in movies and there is an edge between two actors if they have ever appeared in the same movie. In a version of the network from May 2000 [253], it was found that 440 971 out of 449 913 actors were connected together in the largest component, or about 98%. Thus just 2% of actors were not part of the largest component.

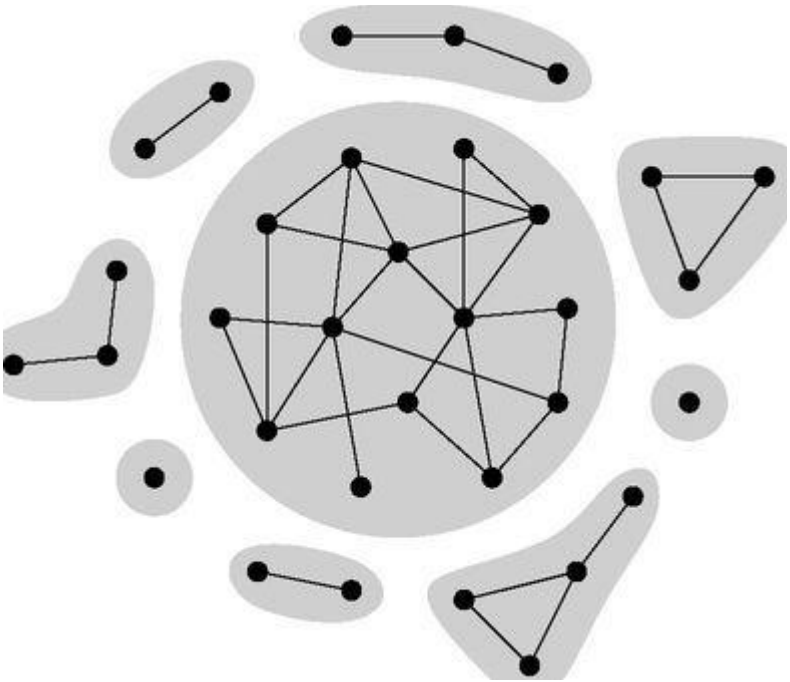


Figure 8.1: Components in an undirected network. In most undirected networks there is a single large component occupying a majority, or at least a significant fraction, of the network, along with a number of small components, typically consisting of only a handful of vertices each.

See Section 6.11.1 for the definition of a weakly connected component.

Table 8.1 summarizes the properties of many of the networks discussed in this chapter, and gives, among other things, the size S of the largest component in each case as a fraction of total

network size. (For the directed networks in the table it is the size of the largest weakly connected component that is quoted. Component sizes in directed networks are discussed further in the following section.) As we can see from the table our figure for the actor network is quite typical for the networks listed and not unusually large.

As the table also shows, there are quite a few networks for which the largest component fills the entire network so that $S = 1$, i.e., the network has only a single component and no smaller components. In the cases where this happens there is usually a good reason. For instance, the Internet is a communication network—its reason for existence is to provide connections between its nodes. There must be at least one path from your vertex to your friend's vertex if the network is to serve its purpose of allowing your and your friend to communicate. To put it another way, there would be no point in being a part of the Internet if you are not part of its largest component, since that would mean that you are disconnected from and unable to communicate with almost everyone else. Thus there is a strong pressure on every vertex of the Internet to be part of the largest component and thus for the largest component to fill the entire network. In other cases the largest component fills the network because of the way the network is measured. The first Web network listed in the table, for instance, is derived from a single web crawl, as described in Section 4.1. Since a crawler can only find a web page if that page is linked to by another page, it follows automatically that all pages found by a single crawl will be connected into a single component. A Web network may, however, have more than one component if, like the ″Alta Vista″ network in the table, it is assembled using several web crawls starting from different locations.

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1			
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16	
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.
	Sexual contacts	Undirected	2 810					3.2			
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7			
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–			
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44	
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.

Table 8.1: Basic statistics for a number of networks. The properties measured are: type of network, directed or undirected; total number of vertices n; total number of edges m; mean degree c; fraction of vertices in the largest component S (or the largest weakly connected component in the case of a directed network); mean geodesic distance between connected vertex pairs ℓ; exponent α of the degree distribution if the distribution follows a power law (or "-" if not; in/out-degree exponents are given for directed graphs); clustering coefficient C from Eq. (7.41); clustering coefficient C_{WS} from the alternative definition of Eq. (7.44); and the degree correlation

coefficient r from Eq. (7.82). The last column gives the citation(s) for each network in the bibliography. Blank entries indicate unavailable data.

Can a network have two or more large components that fill a sizable fraction of the entire graph? Usually the answer to this question is no. We will study this point in more detail in Section 12.6, but the basic argument is this. If we had a network of n vertices that was divided into two large components of about $\frac{1}{2}n$ vertices each, then there would be $\frac{1}{4}n^2$ possible pairs of vertices such that one vertex was in one large component and the other vertex in the other large component. If there is an edge between *any* of these pairs of vertices, then the two components are joined together and are in fact just one component. For example, in our network of movie actors, with half a million vertices, there would about 50 billion pairs, only one of which would have to be joined by an edge to join the two large components into one. Except in very special cases, it is highly unlikely that not one such pair would be connected, and hence also highly unlikely that we will have two large components.

And what about networks with no large component? It is certainly possible for networks to consist only of small components, small groups of vertices connected among themselves but not connected to the rest of the world. An example would be the network of immediate family ties, in which two people are considered connected if they are family members living under the same roof. Such a network is clearly broken into many small components consisting of individual families, with no large component at all. In practice, however, situations like this arise rather infrequently in the study of networks for the anthropocentric reason that people don't usually bother to represent such situations by networks at all. Network representations of systems are normally only useful if most of the network is connected together. If a network is so sparse as to be made only of small components, then there is normally little to be gained by applying techniques like those described in this book. Thus, essentially all of the networks we will be looking at do contain a large component (and certainly all those in Table 8.1, although for some of them the size of that component has not been measured and the relevant entry in the table is blank).

So the basic picture we have of the structure of most networks is that of Fig. 8.1, of a large component filling most of the network, sometimes all of it, and perhaps some other small components that are not connected to the bulk of the network.

8.1.1 COMPONENTS IN DIRECTED NETWORKS

As discussed in Section 6.11, the component structure of directed networks is more complicated than for undirected ones. Directed graphs have weakly and strongly connected components. The weakly connected components correspond closely to the concept of a component in an undirected graph, and the typical situation for weakly connected components is similar to that for undirected graphs: there is usually one large weakly connected component plus, optionally, other small ones. Figures for the sizes of the largest weakly connected components in several directed network are given in Table 8.1.

A strongly connected component, as described in Section 6.11, is a maximal subset of vertices in a network such that each can reach and is reachable from all of the others along a directed path. As with weakly connected components, there is typically one large strongly connected component in a directed network and a selection of small ones. The largest strongly connected component of the World Wide Web, for instance, fills about a quarter of network [56].

Associated with each strongly connected component is an out-component (the set of all vertices that can be reached from any starting point in the strongly connected component along a directed path) and an in-component (the set of vertices from which the strongly connected component can be reached). By their definition, in–and out-components are supersets of the strongly connected component to which they belong and if there is a large strongly connected component then the corresponding in–and out-components will often contain many vertices that lie outside the strongly connected component. In the Web, for example, the portion of the in–and out-components that lie outside the largest strongly connected component each also occupy about a quarter of the network [56].

Each of the small strongly connected components will have its own in–and out-components also. Often these will themselves be small, but they need not be. It can happen that a small strongly connected component C is connected by a directed path to the large strongly connected component, in which case the out-component of the large strongly connected component belongs to (and probably forms the bulk of) C's out-component. Notice that the large out-component can be reachable from many small components in this way—the out-components of different strongly connected components can overlap in directed networks and any vertex can and usually does belong to many out-components. Similar arguments apply, of course, for in-components as well.

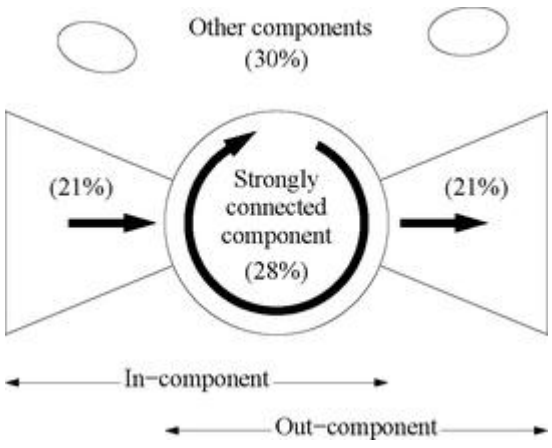


Figure 8.2: The "bow tie" diagram of components in a directed network. The typical directed

network consists of one large strongly connected component and many small ones, each with an in-component and an out-component. Note that by definition each in-component includes the corresponding strongly connected component as a subset, as does each out-component. The largest strongly connected component and its inand out-components typically occupy a significant fraction of the whole network. The percentages shown here indicate how much of the network is taken up by each part of the bow tie in the case of the World Wide Web. After Broder *et al.* [56].

The overall picture for a directed network can be represented using the "bow tie" diagram introduced by Broder and co-workers [56]. In Fig. 8.2 we show the bow tie for the case of the World Wide Web, including percentages (from Ref. [56]) for the fraction of the network occupied by its different parts.

Not all directed networks have a large strongly connected component. In particular, any acyclic directed network has no strongly connected components of size greater than one since if two vertices belong to the same strongly connected component then by definition there exists a directed path through the network in both directions between them, and hence there is a cycle from one vertex to the other and back. Thus if there are no cycles in a network there can be no strongly connected components with two or more vertices. Real-life networks are not usually perfectly acyclic, but some, such as citation networks (Section 4.2) are approximately so. Such networks typically have a few small strongly connected components of two or perhaps three vertices each, but no large ones.

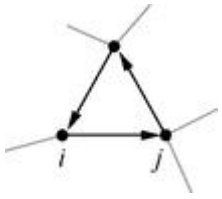
8.2 SHORTEST PATHS AND THE SMALL-WORLD EFFECT

One of the most remarkable and widely discussed of network phenomena is the *small-world effect*, the finding that in many—perhaps most—networks the typical network distances between vertices are surprisingly small. In Section 3.6 we discussed Stanley Milgram's letter-passing experiment in the 1960s, in which people were asked to get a letter from an initial holder to a distant target person by passing it from acquaintance to acquaintance through the social network. The letters that made it to the target did so in a remarkably small number of steps, around six on average. Milgram's experiment is a beautiful and powerful demonstration of the small-world effect, although also a rather poorly controlled one. But with the very complete network data we have for many networks these days it is now possible to measure directly the path lengths between vertices and verify the small-world effect explicitly.

In Section 7.6 we defined the mean distance ℓ between vertices in a network (see Eqs. (7.31) and (7.32)). In mathematical terms, the small-world effect is the hypothesis that this mean distance is small, in a sense that will be defined shortly. In Table 8.1 we list the value of ℓ for each of the networks in the table, and we see that indeed it takes quite small values, always less than 20 and usually less than 10, even though some of the networks have millions of vertices.

One can well imagine that the small-world effect could have substantial implications for networked systems. Suppose a rumor is spread over a social network for instance (or a disease for that matter). Clearly it will reach people much faster if it is only about six steps from any person to any other than if it is a hundred, or a million. Similarly, the speed with which one can get a response from another computer on the Internet depends on how many steps or "hops" data packets have to make as they traverse the network. Clearly a network in which the typical number of hops is only ten or twenty will perform much better than one in which it is ten times as much. (While this point was not articulated by the original designers of the Internet in the 1960s, they must have had some idea of its truth, even if only vaguely, to believe that a network like the Internet could be built and made to work.)

In fact, once one looks more deeply into the mathematics of networks, which we will do in later chapters, one discovers that the small-world effect is not so surprising after all. As we will see in Section 12.7, mathematical models of networks suggest that path lengths in networks should typically scale as $\log n$ with the number n of network vertices, and should therefore tend to remain small even for large networks because the logarithm is a slowly growing function of its argument.



The shortest path from i to j in this network has length 1, but the shortest path from j to i has length 2.

One can ask about path lengths on directed networks as well, although the situation is more complicated there. Since in general the path from vertex i to vertex j is different in a directed

network from the path from j to i, the two paths can have different lengths. Our average distance ℓ should therefore include terms for both distances separately. It's also possible for there to be no path in one direction between two vertices, which we would conventionally denote by setting $d_{ij} = \infty$. As before we could get around the problems caused by the infinite values by defining ℓ as an average over only the finite ones, as in Eq. (7.32). Values calculated in this way are given for the directed networks in Table 8.1. One could also (and perhaps more elegantly) use a harmonic mean as in Eq. (7.34), although this is rarely done.

One can also examine the diameter of a network, which, as described in Section 6.10.1, is the length of the longest finite geodesic path anywhere in the network. The diameter is usually found to be relatively small as well and calculations using network models suggest that it should scale logarithmically with n just as the average distance does. The diameter is in general a less useful measure of real-world network behavior than mean distance, since it really only measures the distance between one specific pair of vertices at the extreme end of the distribution of distances. Moreover, the diameter of a network could be affected substantially by a small change to only a single vertex or a few vertices, which makes it a poor indicator of the behavior of the network as a whole. Nonetheless, there are cases where it is of interest. In Section 8.4 we discuss so-called "scale-free" networks, i.e., networks with power-law degree distributions. Such networks are believed to have an unusual structure consisting of a central "core" to the network that contains most of the vertices and has a mean geodesic distance between vertex pairs that scales only as $\log \log n$ with network size, and not as $\log n$, making the mean distance for the whole network scale as $\log \log n$ also. Outside of this core there are longer "streamers" or "tendrils" of vertices attached to the core like hair, which have length typically of order $\log n$, making the *diameter* of the network of order $\log n$ [67, 75]. This sort of behavior could be detected by measuring separately the mean geodesic distance and diameter of networks of various sizes to confirm that they vary differently with n. (It's worth noting, however, that behavior of the form $\log \log n$ is very difficult to confirm in real-world data because $\log \log n$ is a *very* slowly varying function of n.)

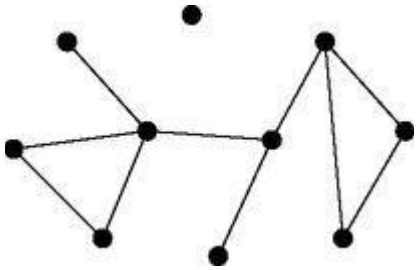
Another interesting twist on the small-world effect was discussed by Milgram in his original paper on the problem. He noticed, in the course of his letter-passing experiments, that most of the letters destined for a given target person passed through just one or two acquaintances of the target. Thus, it appeared, most people who knew the target person knew him through these one or two people. This idea, that one or two of your acquaintances are especially well connected and responsible for most of the connection between you and the rest of the world has been dubbed *funneling*, and it too is something we can test against complete networks with the copious data available to us today. If, for instance, we focus on geodesic paths between vertices, as we have been doing in this section, then we could measure what fraction of the shortest paths between a vertex i and every other reachable vertex go through each of i's neighbors in the network. For many networks, this measurement does reveal a funneling effect. For instance, in the coauthorship network of physicists from Table 8.1 it is found that, for physicists having five or more collaborators, 48% of geodesic paths go through one neighbor of the average vertex, the remaining 52% being distributed over the other four or more neighbors. A similar result is seen in the Internet. Among nodes having degree five or greater in a May 2005 snapshot of Internet structure at the autonomous system level, an average of 49% of geodesic paths go through one neighbor of the average vertex. It is tempting to draw conclusions about the routing of Internet packets from this latter result—perhaps that the network will tend to overload a small number of well-connected nodes rather than distributing load more evenly—but it is worth noticing that, although Internet packets tended to be routed along shortest paths during the early days of the Internet, much more sophisticated routing strategies are in place today, so statistics for shortest paths may not reflect actual packet flows very closely.

Milgram referred to these people as "sociometric superstars." We discussed them previously in Section 3.6.

8.3 DEGREE DISTRIBUTIONS

In this section, we look at one of the most fundamental of network properties, the frequency distribution of vertex degrees. This distribution will come up time and again throughout this book as a defining characteristic of network structure.

As described in Section 6.9, the degree of a vertex is the number of edges attached to it. Let us first consider undirected networks. We define p_k to be the fraction of vertices in such a network that have degree k. For example, consider this network:



It has $n = 10$ vertices, of which 1 has degree 0, 2 have degree 1, 4 have degree 2, 2 have degree 3, and 1 has degree 4. Thus the values of p_k for $k = 0,..., 4$ are

$$p_0 = \frac{1}{10}, \quad p_1 = \frac{2}{10}, \quad p_2 = \frac{4}{10}, \quad p_3 = \frac{2}{10}, \quad p_4 = \frac{1}{10},$$

(8.1)

and $p_k = 0$ for all $k > 4$. The quantities p_k represent the *degree distribution* of the network.

The value p_k can also be thought of as a probability: it is the probability that a randomly chosen vertex in the network has degree k. This will be a useful viewpoint when we study theoretical models of networks in Chapters 12 to 15.

Sometimes, rather than the fraction of vertices with a given degree, we will want the total number of such vertices. This is easily calculated from the degree distribution, being given simply by np_k, where n is as usual the total number of vertices.

Another construct containing essentially the same information as the degree distribution is the *degree sequence*, which is the set $\{k_1, k_2, k_3,...\}$ of degrees for all the vertices. For instance, the degree sequence of the small graph above is $\{0, 1, 1, 2, 2, 2, 2, 3, 3, 4\}$. (The degree sequence need not necessarily be given in ascending order of degrees as here. For instance, in many cases the vertices are given numeric labels and their degrees are then listed in the order of the labels.)

It is probably obvious, but bears saying anyway, that a knowledge of the degree distribution (or degree sequence) does not, in most cases, tell us the complete structure of a network. For most choices of vertex degrees there is more than one network with those degrees. These two networks, for instance, are different but have the same degrees:

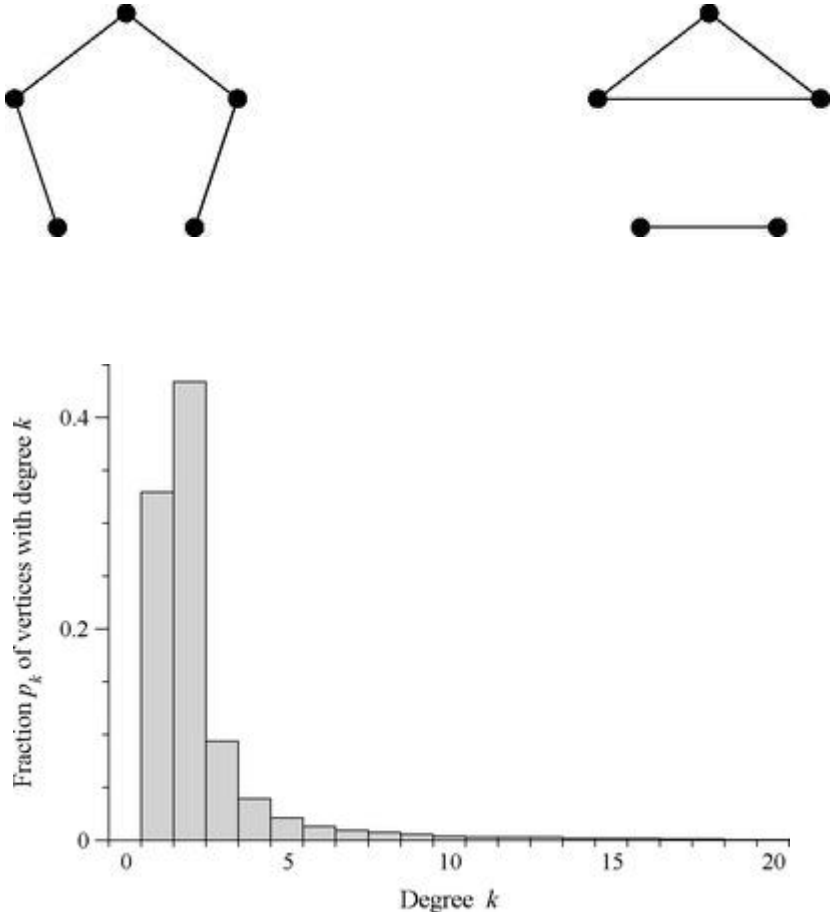


Figure 8.3: The degree distribution of the Internet. A histogram of the degree distribution of the vertices of the Internet graph at the level of autonomous systems.

Thus we cannot tell the complete structure of a network from its degrees alone. The degree sequence certainly gives us very important information about a network, but it doesn't give us complete information.

It is often illuminating to make a plot of the degree distribution of a large network as a function of k. Figure 8.3 shows an example of such a plot for the Internet at the level of autonomous systems. The figure reveals something interesting: most of the vertices in the network have low degree—one or two or three—but there is a significant "tail" to the distribution, corresponding to vertices with substantially higher degree.[118] The plot cuts off at degree 20, but in fact the tail goes much further than this. The highest degree vertex in the network has degree 2407. Since there are, for this particular data set, a total of 19 956 vertices in the network, that means that the most highly connected vertex is connected to about 12% of all other vertices in the network. We call such a well-connected vertex a *hub*[119]. Hubs will play an important role in the developments of the following chapters.

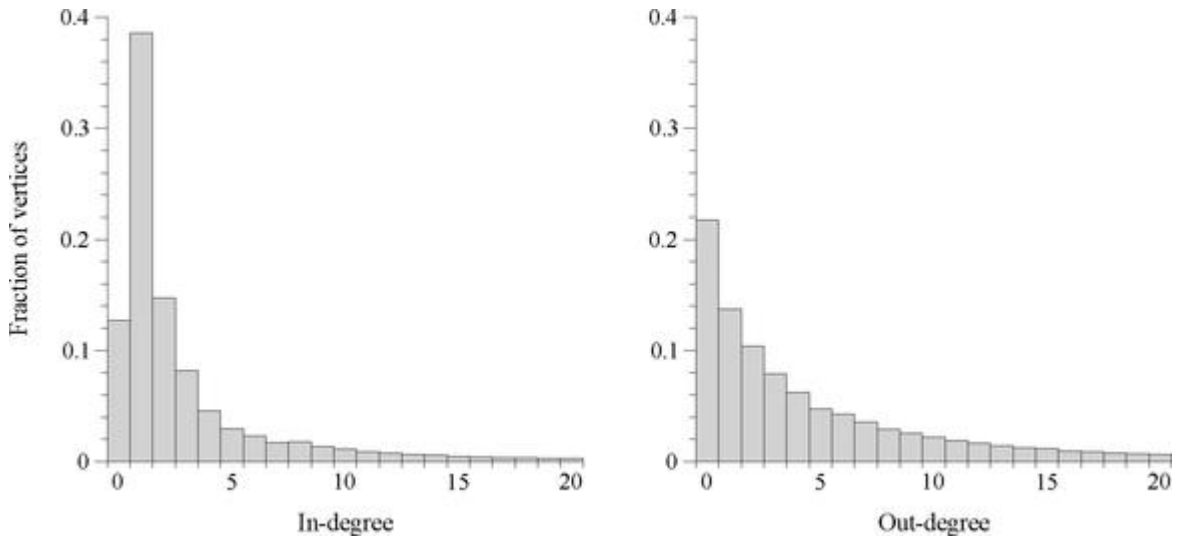


Figure 8.4: The degree distributions of the World Wide Web. Histograms of the distributions of in–and out-degrees of pages on the World Wide Web. Data are from the study by Broder *et al.* [56].

In fact, it turns out that almost all real-world networks have degree distributions with a tail of high-degree hubs like this. In the language of statistics we say that the degree distribution is *right-skewed*. Right-skewed degree distributions are discussed further in Section 8.4, and will reappear repeatedly throughout this book.

One can also calculate degree distributions for directed networks. As discussed in Section 6.9, directed networks have two different degrees for each vertex, the in-degree and the out-degree, which are, respectively, the number of edges ingoing and outgoing at the vertex of interest. There are, correspondingly, two different degree distributions in a directed network, the in-degree and out-degree distributions, and one can make a plot of either, or both. Figure 8.4, for example, shows the degree distributions for the World Wide Web.

If we wish to be more sophisticated, we might observe that the true degree distribution of a directed network is really a joint distribution of in–and out- degrees. We can define p_{jk} to be the fraction of vertices having simultaneously an in-degree j and an out-degree k. This is a two-dimensional distribution that cannot be plotted as a simple histogram, although it could be plotted as a two-dimensional density plot or as a surface plot. By using a joint distribution in this way we can allow for the possibility that the in–and out-degrees of vertices might be correlated. For instance, if vertices with high in-degree also tended to have high out-degree, then we would see this reflected in large values of p_{jk} when both j and k were large. If we only have the separate distributions of inand out-degree individually, but not the joint distribution, then there is no way of telling whether the network contains such correlations.

In practice, the joint in/out degree distribution of directed networks has rarely been measured or studied, so there is relatively little data on it. This is, in some ways, a pity, since many of our theories of directed networks depend on a knowledge of the joint distribution to give accurate answers (see Section 13.11), while others make predictions about the joint distribution that we would like to test against empirical data. For the moment, however, this is an area awaiting more thorough exploration.

8.4 POWER LAWS AND SCALE-FREE NETWORKS

Returning to the Internet, another interesting feature of its degree distribution is shown in Fig. 8.5, where we have replotted the histogram of Fig. 8.3 using logarithmic scales. (That is, both axes are logarithmic. We have also made the range of the bins bigger in the histogram to make the effect clearer—they are of width five in Fig. 8.5 where they were only of width one before.) As the figure shows, when viewed in this way, the degree distribution follows, roughly speaking, a straight line. In mathematical terms, the logarithm of the degree distribution p_k is a linear function of degree k thus:

$$\ln p_k = -\alpha \ln k + c,$$

(8.2)

where α and c are constants. The minus sign here is optional—we could have omitted it—but it is convenient, since the slope of the line in Fig. 8.5 is clearly negative, making α a positive constant equal to minus the slope in the figure. In this case, the slope gives us a value for α of about 2.1.

Taking the exponential of both sizes of Eq. (8.2), we can also write this logarithmic relation as

$$p_k = Ck^{-\alpha},$$

(8.3)

where $C = e^c$ is another constant. Distributions of this form, varying as a power of k, are called *power laws*. Based on the evidence of Fig. 8.5 we can say that, roughly speaking, the degree distribution of the Internet follows a power law.

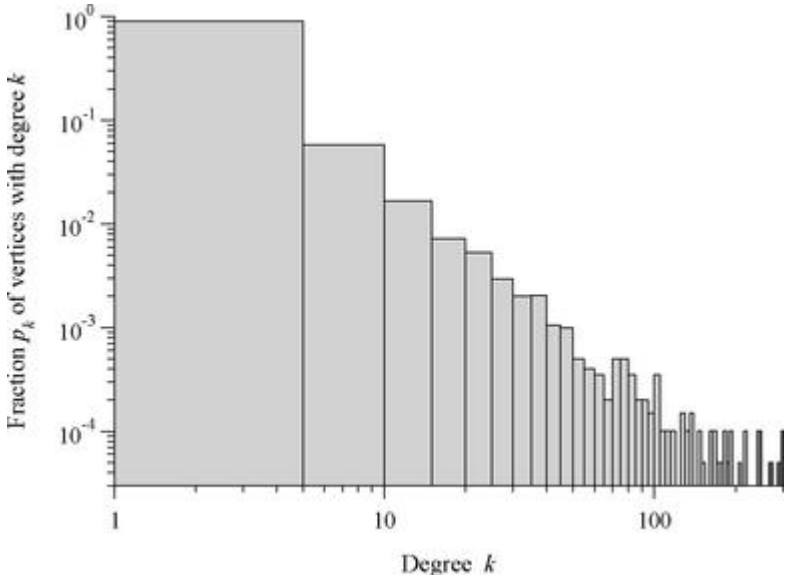


Figure 8.5: The power-law degree distribution of the Internet. Another histogram of the degree distribution of the Internet graph, plotted this time on logarithmic scales. The approximate straight-line form of the histogram indicates that the degree distribution roughly follows a power law of the form (8.3).

This is, in fact, a common pattern seen in quite a few different networks. For instance, as shown in Fig. 8.8 on page 253, both the in–and out-degrees of the World Wide Web roughly follow power-law distributions, as do the indegrees in many citation networks (but not the out-degrees).

The constant α is known as the *exponent* of the power law. Values in the range $2 \leq \alpha \leq 3$ are typical, although values slightly outside this range are possible and are observed occasionally. Table 8.1 gives the measured values of the exponents for a number of networks that have power-law or approximately power-law degree distributions, and we see that most of them fall in this range. The constant C in Eq. (8.3) is mostly uninteresting, being fixed by the requirement of normalization, as described in Section 8.4.2.

Degree distributions do not usually follow Eq. (8.3) over their entire range. Looking at Fig. 8.3, for example, we can see that the degree distribution is not monotonic for small k, even allowing for statistical fluctuations in the histogram. A true power-law distribution is monotonically decreasing over its entire range and hence the degree distribution must in this case deviate from the true power law in the small-k regime. This is typical. A common situation is that the power law is obeyed in the tail of the distribution, for large values of k, but not in the small-k regime. When one says that a particular network has a power-law degree distribution one normally means only that the tail of the distribution has this form. In some cases, the distribution may also deviate from the power-law form for high k as well. For instance, there is often a cut-off of some type that limits the maximum degree of vertices in the tail.

Networks with power-law degree distributions are sometimes called *scale-free networks*, and we will use this terminology occasionally. Of course, there are also many networks that are not scale-free, that have degree distributions with non-power-law forms, but the scale-free ones will be of particular interest to us because they have a number of intriguing properties. Telling the scale-free ones from the non-scale-free is not always easy however. The simplest strategy is to look at a histogram of the degree distribution on a log-log plot, as we did in Fig. 8.5, to see if we have a straight line. There are, however, a number of problems with this approach and where possible we recommend you use other methods, as we now explain.

8.4.1 DETECTING AND VISUALIZING POWER LAWS

As a tool for visualizing or detecting power-law behavior, a simple histogram like Fig. 8.5 presents some problems. One problem obvious from the figure is that the statistics of the histogram are poor in the tail of the distribution, the large-k region, which is precisely the region in which the power law is normally followed most closely. Each bin of the histogram in this region contains only a few samples, which means that statistical fluctuations in the number of samples from bin to bin are large. This is visible as a "noisy signal" at the righthand end of Fig. 8.5 that makes it difficult to determine whether the histogram really follows a straight line or not, and what the slope of that line is.

There are a number of solutions to this problem. The simplest is to use a histogram with larger bins, so that more samples fall into each bin. In fact, we already did this in going from Fig. 8.3 to Fig. 8.5—we increased the bin width from one to five between the two figures. Larger bins contain more samples and hence give less noise in the tail of the histogram, but at the expense of less detail overall, since the number of bins is correspondingly reduced. Bin width in this situation is always something of a compromise: we would like to use very wide bins in the tail of the distribution where noise is a problem, but narrower ones at the left-hand end of the histogram where there are many samples and we would prefer to have more bins if possible.

Alternatively, we could try to get the best of both worlds by using bins of different sizes in different parts of the histogram. For example, we could use bins of width one for low degrees and switch to width five for higher degrees. In doing this we must be careful to normalize the bins correctly: a bin of width five will on average accrue five times as many samples as a similarly placed bin of width one, so if we wish to compare counts in the two we should divide the number of samples in the larger bin by five. More generally, we should divide sample counts by the width of their bins to make counts in bins of different widths comparable.

We need not restrict ourselves to only two different sizes of bin. We could use larger and larger bins as we go further out in the tail. We can even make every bin a different size, each one a little larger than the one before it. One commonly used version of this idea is called *logarithmic binning*. In this scheme, each bin is made wider than its predecessor by a constant factor a. For instance, if the first bin in a histogram covers the interval $1 \le k < 2$ (meaning that all vertices of degree 1 fall in this bin) and $a = 2$, then the second would cover the interval $2 \le k < 4$ (vertices of degrees 2 and 3), the third the interval $4 \le k < 8$, and so forth. In general the nth bin would cover the interval $a^{n-1} \le k < a^n$ and have width $a^n - a^{n-1} = (a-1)\,a^{n-1}$. The most common choice for a is $a = 2$, since larger values tend to give bins that are too coarse while smaller ones give bins with non-integer limits.

Figure 8.6 shows the degree distribution of the Internet binned logarithmically in this way. We have been careful to normalize each bin by dividing by its width, as described above. As we can see, the histogram is now much less noisy in the tail and it is considerably easier to see the straight-line behavior of the degree distribution. The figure also reveals a nice property of logarithmically binned histograms, namely that when plotted on logarithmic scales as here, the bins in such a histogram appear to have equal width. This is, in fact, the principal reason for this particular choice of bins and also the origin of the name "logarithmic binning."

Note that on a logarithmically binned histogram there is never any bin that contains vertices of degree zero. Since there is no zero on logarithmic scales like those of Fig. 8.6, this doesn't usually make much difference, but if we do want to know how many vertices there are of degree zero we will have to measure this number separately.

A different solution to the problem of visualizing a power-law distribution is to construct the *cumulative distribution function*, which is defined by

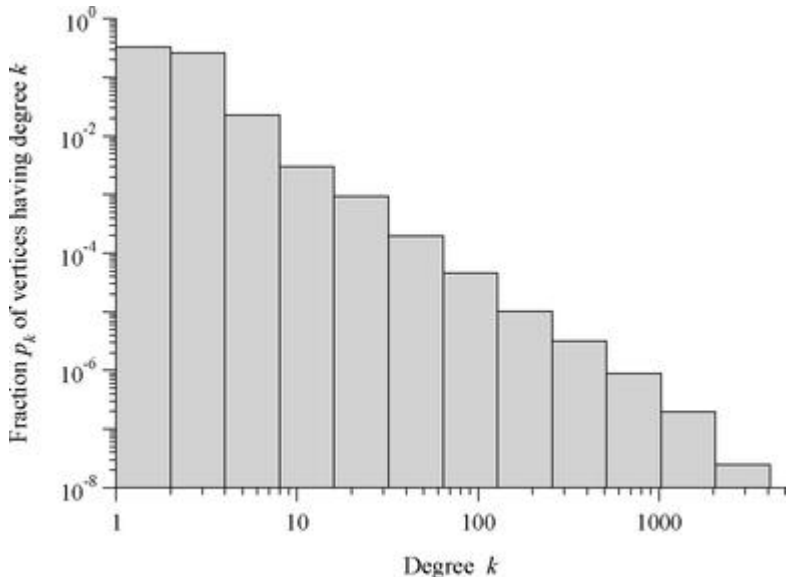


Figure 8.6: Histogram of the degree distribution if the Internet, created using logarithmic binning. In this histogram the widths of the bins are constant on a logarithmic scale, meaning that on a linear scale each bin is wider by a constant factor than the one to its left. The counts in the bins are normalized by dividing by bin width to make counts in different bins comparable.

$$P_k = \sum_{k'=k}^{\infty} p_{k'}.$$

(8.4)

In other words, P_k is the fraction of vertices that have degree k or greater. (Alternatively, it is the probability at a randomly chosen vertex has degree k or greater.)

Suppose the degree distribution p_k follows a power law in its tail. To be precise, let us say that $p_k = Ck^{-\alpha}$ for $k \geq k_{\min}$ for some $k_{\min}$. Then for $k \geq k_{\min}$ we have

$$\begin{aligned} P_k &= C \sum_{k'=k}^{\infty} k'^{-\alpha} \simeq C \int_k^{\infty} k'^{-\alpha} \, \mathrm{d}k' \\ &= \frac{C}{\alpha - 1} k^{-(\alpha-1)}, \end{aligned}$$

(8.5)

where we have approximated the sum by an integral, which is reasonable since the power law is a slowly varying function for large k. (We are also assuming that $\alpha > 1$ so that the integral

converges.) Thus we see that if the distribution p_k follows a power law, then so does the cumulative distribution function P_k, but with an exponent $\alpha - 1$ that is 1 less than the original exponent.

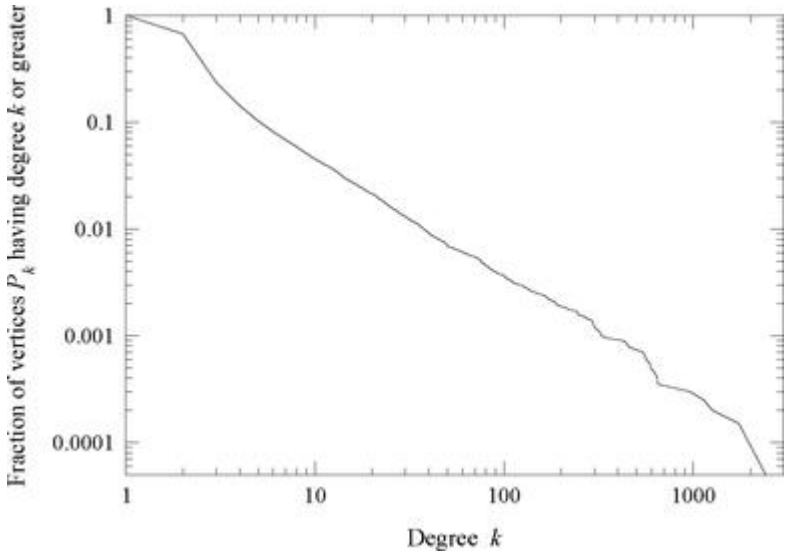
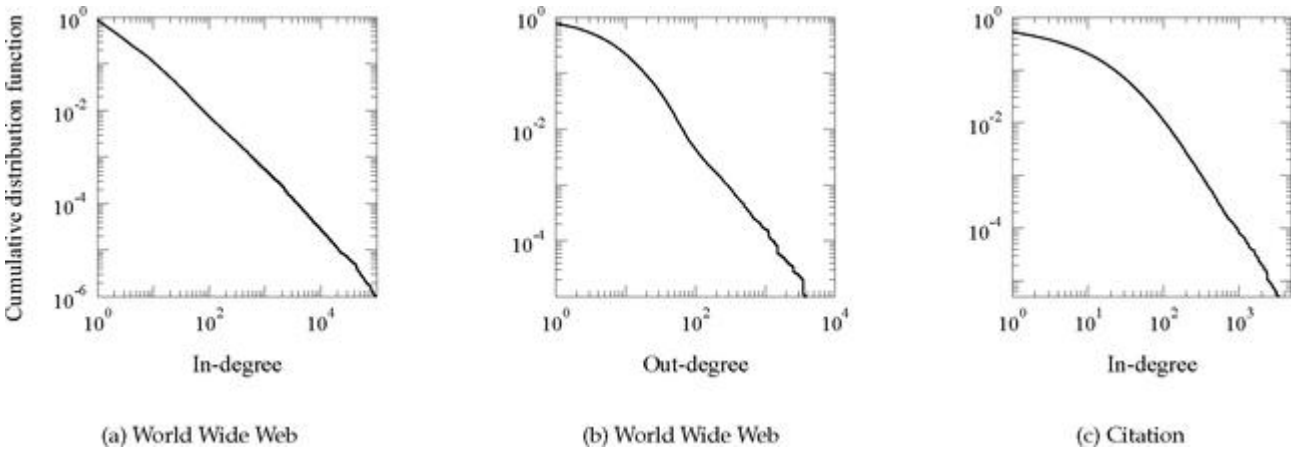


Figure 8.7: Cumulative distribution function for the degrees of vertices on the Internet. For a distribution with a power-law tail, as is approximately the case for the degree distribution of the Internet, the cumulative distribution function, Eq. (8.4), also follows a power law, but with a slope 1 less than that of the original distribution.

This gives us another − way of visualizing a power-law distribution: we plot the cumulative distribution function on log-log scales, as we did for the original histogram, and again look for straight-line behavior. We have done this in Fig. 8.7 for the case of the Internet, and the (approximate) straight-line form is clearly visible. Three more examples are shown in Fig. 8.8, for the in–and out-degree distributions of the World Wide Web and for the in-degree distribution of a citation network.

This approach has some advantages. In particular, the calculation of P_k does not require us to bin the values of k as we do with a normal histogram. P_k is perfectly well defined for any value of k and can be plotted just as a normal function. When bins in a histogram contain more than one value of k—i.e., when their width is greater than 1—the binning of data necessarily throws away quite a lot of the information contained in the data, eliminating, as it does, the distinction between any two values that fall into the same bin. The cumulative distribution function on the other hand preserves all of the information contained in the data, because no bins are involved. The most obvious manifestation of this difference is that the number of points in a plot like Fig. 8.5 or Fig. 8.6 is relatively small, whereas in a cumulative distribution plot like Fig. 8.7 there are as many points along the k (horizontal) axis as there are distinct values of k.



(a) World Wide Web (b) World Wide Web (c) Citation

Figure 8.8: Cumulative distribution functions for in–and out-degrees in three directed networks. (a) The in-degree distribution of the World Wide Web, from the data of Broder *et al.* [56]. (b) The out-degree distribution for the same Web data set. (c) The in-degree distribution of a citation network, from the data of Redner [280]. The distributions follow approximate power-law forms in each case.

The cumulative distribution function is also easy to calculate. The number of vertices with degree greater than or equal to that of the *r*th-highest-degree vertex in a network is, by definition, *r*. Thus the *fraction* with degree greater than or equal to that of the *r*th-highest-degree vertex in a network is $P_k = r/n$. So a simple way of finding P_k is to sort the degrees of the vertices in descending order and then number them from 1 to n in that order. These numbers are the so-called *ranks* r_i of the vertices. A plot of r_i/n as a function of degree k_i, with the vertices in rank order, then gives us our cumulative distribution plot.[120]

For instance, consider again the small example network we looked at at the beginning of Section 8.3, on page 244. The degrees of the vertices in that case were {0, 1, 1, 2, 2, 2, 2, 3, 3, 4}. Listing these in decreasing order and numbering them, we can easily calculate P_k as follows:

Degree k	Rank r	$P_k = r/n$
4	1	0.1
3	2	0.2
3	3	0.3
2	4	0.4
2	5	0.5
2	6	0.6
2	7	0.7
1	8	0.8
1	9	0.9
0	10	1.0

Then a plot of the last column as a function of the first gives us our cumulative distribution function.

Cumulative distributions do have some disadvantages. One is that they are less easy to interpret than ordinary histograms, since they are only indirectly related to the actual distribution of vertex

degrees. A more serious disadvantage is that the successive points on a cumulative plot are correlated—the cumulative distribution function in general only changes a little from one point to the next, so adjacent values are not at all independent. This means that it is not valid for instance to extract the exponent of a power-law distribution by fitting the slope of the straight-line portion of a plot like Fig. 8.7 and equating the result with $\alpha - 1$, at least if the fitting is done using standard methods such as least squares that assume independence between the data points.

In fact, it is in general not good practice to evaluate exponents by performing straight-line fits to either cumulative distribution functions or ordinary histograms. Both are known to give biased answers, although for different reasons [72, 141]. Instead, it is usually better to calculate α directly from the data, using the formula

$$\alpha = 1 + N \left[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}} \right]^{-1}.$$

(8.6)

Here, $k_{\min}$ is the minimum degree for which the power law holds, as before, and N is the number of vertices with degree greater than or equal to $k_{\min}$. The sum is performed over only those vertices with $k \geq k_{\min}$, and not over all vertices.

We can also calculate the statistical error on α from the formula:

$$\sigma = \sqrt{N} \left[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}} \right]^{-1} = \frac{\alpha - 1}{\sqrt{N}}.$$

(8.7)

For example, applying Eqs. (8.6) and (8.7) to the degree sequence of the Internet from Fig. 8.3 gives an exponent value of $\alpha = 2.11 \pm 0.01$.

The derivation of these formulas, which makes use of maximum likelihood techniques, would take us some way from our primary topic of networks, so we will not go into it here. The interested reader can find a discussion in Ref. [72], along with many other details such as methods for determining the value of $k_{\min}$ and methods for telling whether a particular distribution follows a power law at all.

8.4.2 PROPERTIES OF POWER-LAW DISTRIBUTIONS

Quantities with power-law distributions behave in some surprising ways. We take a few pages here to look at some of the properties of power-law distributions, since the results will be of use to us later on.

Power laws turn up in a wide variety of places, not just in networks. They are found in the sizes of city populations [24, 336], earthquakes [153], moon craters [230], solar flares [203], computer files [84], and wars [283]; in the frequency of use of words in human languages [109, 336], the frequency of occurrence of personal names in most cultures [335], the numbers of papers scientists write [201], and the number of hits on web pages [5]; in the sales of books, music recordings, and almost every other branded commodity [83, 185]; and in the numbers of species in biological taxa [58, 330]. A review of the data and some mathematical properties of power laws can be found in Ref. [244]. Here we highlight just a few issues that will be relevant for our study of networks.

Normalization: The constant C appearing in Eq. (8.3) is fixed by the requirement that the degree distribution be normalized. That is, when we add up the total fraction of vertices having all possible degrees $k = 0 \ldots \infty$, we must get 1:

$$\sum_{k=0}^{\infty} p_k = 1.$$

(8.8)

If our degree distribution truly follows a pure power law, obeying Eq. (8.3) for all k, then no vertices of degree zero are allowed, because p_0 would then be infinite, which is impossible since it is a probability and must lie between 0 and 1. Let us suppose therefore that the distribution starts at $k = 1$. Substituting from Eq. (8.3) we then find that $C\sum_k k^{-\alpha} = 1$, or

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)},$$

(8.9)

where $\zeta(\alpha)$ is the Riemann zeta function. Thus the correctly normalized power-law distribution is

$$p_k = \frac{k^{-\alpha}}{\zeta(\alpha)},$$

(8.10)

for $k > 0$ with $p_0 = 0$.

This is a reasonable starting point for mathematical models of scale-free networks—we will use it in Chapter 13—but it's not a very good representation of most real-world networks, which deviate from pure power-law behavior for small k as described above and seen in Fig. 8.3. In that case, the normalization constant will take some other value dependent on the particular shape of the distribution, but nonetheless it is still fixed by the requirement of normalization and we must make sure we get it right in our calculations.

For some of our calculations we will be interested only in the tail of the distribution where the power-law behavior holds and can discard the rest of the data. In such cases, we normalize over only the tail, starting from the minimum value $k_{\min}$ for which the power law holds, as above. This gives

$$p_k = \frac{k^{-\alpha}}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})},$$

(8.11)

where $\zeta(\alpha, k_{\min})$ is the so-called generalized or incomplete zeta function.

Alternatively, we could observe, as we did for Eq. (8.5), that in the tail of the distribution the sum over k is well approximated by an integral, so that the normalization constant can written

$$C \simeq \frac{1}{\int_{k_{\min}}^{\infty} k^{-\alpha}\, \mathrm{d}k} = (\alpha - 1) k_{\min}^{\alpha - 1},$$

(8.12)

or

$$p_k \simeq \frac{\alpha - 1}{k_{\min}} \left(\frac{k}{k_{\min}} \right)^{-\alpha}.$$

(8.13)

In the same approximation the cumulative distribution function, Eq. (8.5), is given by

$$P_k = \left(\frac{k}{k_{\min}} \right)^{-(\alpha - 1)}.$$

(8.14)

Moments: Of great interest to us will be the moments of the degree distribution. The first moment of a distribution is its mean:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k.$$

(8.15)

The second moment is the mean square:

$$\langle k^2 \rangle = \sum_{k=0}^{\infty} k^2 p_k.$$

(8.16)

And the *m*th moment is

$$\langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k.$$

(8.17)

Suppose we have a degree distribution p_k that has a power-law tail for $k \geq k_{\min}$, in the manner of the Internet or the World Wide Web. Then

$$\langle k^m \rangle = \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^{m-\alpha}.$$

(8.18)

Since the power law is a slowly varying function of k for large k, we can again approximate the second sum by an integral thus:

$$\begin{aligned}\langle k^m \rangle &\simeq \sum_{k=0}^{k_{\min}-1} k^m p_k + C \int_{k_{\min}}^{\infty} k^{m-\alpha}\, \mathrm{d}k \\ &= \sum_{k=0}^{k_{\min}-1} k^m p_k + \frac{C}{m-\alpha+1}\Big[k^{m-\alpha+1}\Big]_{k_{\min}}^{\infty},\end{aligned}$$

(8.19)

The first term here is some finite number whose value depends on the particular (non-power-law) form of the degree distribution for small k. The second term however depends on the values of m and α. If $m - \alpha + 1 < 0$, then the bracket has a finite value, and k^m is well-defined. But if $m - \alpha + 1 \geq 0$ then the bracket diverges and with it the value of k^m. Thus, the mth moment of the degree distribution is finite if and only if $\alpha > m + 1$. Put another way, for a given value of α all moments will diverge for which $m \geq \alpha - 1$.

Of particular interest to us will be the second moment k^2, which arises in many calculations to do with networks (such as mean degree of neighbors, Section 13.3, robustness calculations, Section 16.2.1, epidemiological processes, Section 17.8.1, and many others). The second moment is finite if and only if $\alpha > 3$. As discussed above, however, most real-world networks with power-law degree distributions have values of α in the range $2 \leq \alpha \leq 3$, which means that the second moment should diverge, an observation that has a number of remarkable implications for the properties of scale-free networks, some of which we will explore in coming chapters. Notice that this applies even for networks where the power law only holds in the tail of the distribution—the distribution does not have to follow a power law everywhere for the second moment to diverge.

These conclusions, however, are slightly misleading. In any real network all the moments of the degree distribution will actually be finite. We can always calculate the mth moment directly from the degree sequence thus:

$$\langle k^m \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i^m,$$

(8.20)

and since all the k_i are finite, so must the sum be. When we say that the mth moment is infinite, what we mean is that if we were to calculate it for an arbitrarily large network with the same power-law degree distribution the value would be infinite. But for any finite network Eq. (8.20) applies and all moments are finite.

There is however another factor that limits the values of the higher moments of the degree distribution, namely that most real-world networks are simple graphs. That is, they have no multiedges and no self-loops, which means that a vertex can have, at most, one edge to every other vertex in the network, giving it a maximum degree of $n - 1$, where n is the total number of vertices. In practice, the power-law behavior of the degree distribution may be cut off for other reasons before we reach this limit, but in the worst case, an integral such as that of Eq. (8.19) will be cut off in a simple graph at $k = n$ so that

$$\langle k^m \rangle \sim \left[k^{m-\alpha+1} \right]_{k_{\min}}^{n} \sim n^{m-\alpha+1},$$

(8.21)

as $n \to \infty$ for $m > \alpha - 1$. This again gives moments that are finite on finite networks but become infinite as the size of the network becomes infinite. For instance, the second moment goes as

$$\langle k^2 \rangle \sim n^{3-\alpha}.$$

(8.22)

In a network with $\alpha = \frac{5}{2}$, this diverges as $n^{1/2}$ as the network becomes large.

We will throughout this book derive results that depend on moments of the degree distributions of networks. Some of those results will show unusual behavior in power-law networks because of the divergence of the moments. On practical, finite networks that divergence is replaced by large finite values of the moments. In many cases, however, this produces similar results to a true divergence. On the Internet, for instance, with its power-law degree distribution and a total of about $n \simeq 20\,000$ autonomous systems as vertices, we can expect the second (and all higher moments) to take not infinite but very large values. For the Internet data we used in Figs. 8.3 and 8.5 the second moment has the value $k^2 = 1159$, which can in practice be treated as infinite for many purposes.

Top-heavy distributions: Another interesting quantity is the fraction of edges in a network that connect to the vertices with the highest degrees. For a pure power-law degree distribution, it can be shown [244] that a fraction W of ends of edges attach to a fraction P of the highest-degree vertices in the network, where

$$W = P^{(\alpha-2)/(\alpha-1)},$$

(8.23)

A set of curves of W against P is shown in Fig. 8.9 for various values of α. Curves of this kind are called *Lorenz curves*, after Max Lorenz, who first studied them around the turn of the twentieth century [200]. As the figure shows, the curves are concave downward for all values of α, and for values only a little above 2 they have a very fast initial increase, meaning that a large fraction of the edges are connected to a small fraction of the highest degree nodes.

Thus, for example, the in-degree distribution of the World Wide Web follows a power law above about $k_{\min} = 20$ with exponent around $\alpha = 2.2$. Equation (8.23) with $P = \frac{1}{2}$ then tells us that we would expect that about $W = 0.89$ or 89% of all hyperlinks link to pages in the top half of the degree distribution, while the bottom half gets a mere 11%. Conversely, if we set $W = \frac{1}{2}$ in Eq.

(8.23) we get $P = 0.015$, implying that 50% of all the links go to less than 2% of the "richest" vertices. Thus the degree distribution is in a sense "top-heavy," a large fraction of the "wealth"—meaning incoming hyperlinks in this case—falling to a small fraction of the vertices.

This calculation assumes a degree distribution that follows a perfect power law, whereas in reality, as we have seen, degree distributions usually only follow a power law in their high-degree tail. The basic principle still holds, however, and even if we cannot write an exact formula like Eq. (8.23) for a particular network we can easily evaluate W as a function of P directly from degree data. For the real degree distribution of the Web[122] we find that 50% of the incoming hyperlinks point to just 1.1% of the richest vertices (so Eq. (8.23) was not too bad in this case).

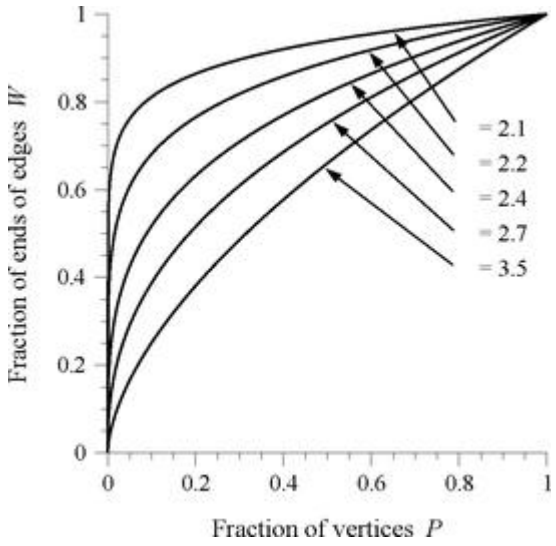


Figure 8.9: Lorenz curves for scale-free networks. The curves show the fraction W of the total number of ends of edges in a scale-free network that are attached to the fraction P of vertices with the highest degrees, for various values of the power-law exponent α.

Similarly, for paper citations 8.3% of the highest cited papers get 50% of all the citations[123] and on the Internet just 3.3% of the most highly connected nodes have 50% of the connections.[124]

In the remaining chapters of this book we will see many examples of networks with power-law degree distributions, and we will make use of the results of this section to develop an understanding of their behavior.

8.5 DISTRIBUTIONS OF OTHER CENTRALITY MEASURES

Vertex degree is just one of a variety of centrality measures for vertices in networks, as discussed in Chapter 7. Other centrality measures include eigenvector centrality and its variations (Sections 7.2 to 7.5), closeness centrality (Section 7.6), and betweenness centrality (Section 7.7). The distributions of these other measures, while of lesser importance in the study of networks than the degree distribution, are nonetheless of some interest.

Eigenvector centrality can be thought of as an extended form of degree centrality, in which we take into account not only how many neighbors a vertex has but also how central those neighbors themselves are (Section 7.2). Given its similarity to degree centrality, it is perhaps not surprising to learn that eigenvector centrality often has a highly right-skewed distribution. The left panel of Fig. 8.10 shows the cumulative distribution of eigenvector centralities for the vertices of the Internet, using again the autonomous-system-level data that we used in Section 8.3. As the figure shows, the tail of the distribution approximately follows a power law but the distribution rolls off for vertices with low centrality. Similar roughly power-law behavior is also seen in eigenvector centralities for other scale-free networks, such as the World Wide Web and citation networks, while other networks show right-skewed but non-power-law distributions.

Betweenness centrality (Section 7.7) also tends to have right-skewed distributions on most networks. The right panel of Fig. 8.10 shows the cumulative distribution of betweenness for the vertices of the Internet and, as we can see, this distribution is again roughly power-law in form. Again there are some other networks that also have power-law betweenness distributions and others still that have skewed but non-power-law distributions.

An exception to this pattern is the closeness centrality (Section 7.6), which is the mean geodesic distance from a vertex to all other reachable vertices. As discussed in Section 7.6 the values of the closeness centrality are typically limited to a rather small range from a lower bound of 1 to an upper bound of order log n, and this means that their distribution cannot have a long tail. In Fig. 8.11, for instance, we show the distributions of closeness centralities for our snapshot of the Internet, and the distribution spans well under an order of magnitude from a minimum of 2.30 to a maximum of 7.32. There is no long tail to the distribution, and the distribution is not even roughly monotonically decreasing (as our others have been) but shows clear peaks and dips.

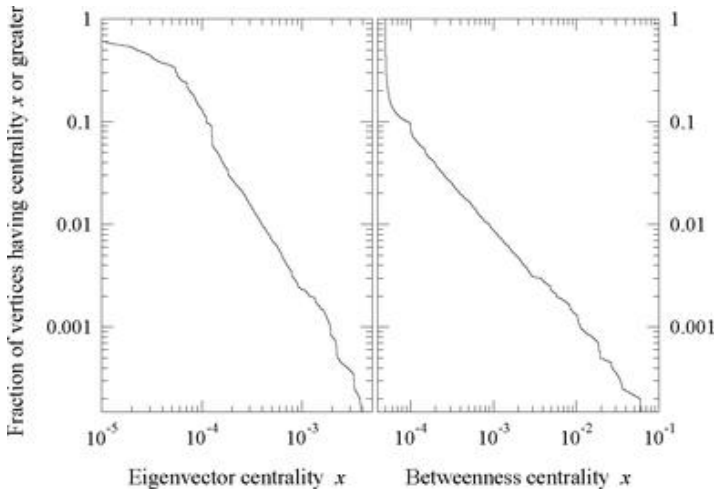


Figure 8.10: Cumulative distribution functions for centralities of vertices on the Internet. Left panel: eigenvector centrality. Right panel: betweenness centrality.

8.6 CLUSTERING COEFFICIENTS

See Section 7.9 for a discussion of clustering coefficients.

The clustering coefficient measures the average probability that two neighbors of a vertex are themselves neighbors. In effect it measures the density of triangles in the networks and it is of interest because in many cases it is found to have values sharply different from what one would expect on the basis of chance. To see what we mean by this, look again at Table 8.1 on page 237, which gives measured values of the clustering coefficient for a variety of networks. (Look at the column denoted C, which gives values for the coefficient defined by Eq. (7.41).) Most of the values are of the order of tens of percent—there is typically a probability between about 10% and maybe 60% that two neighbors of a vertex will be neighbors themselves. However, as we will see in Section 13.4, if we consider a network with a given degree distribution in which connections between vertices are made at random, the clustering coefficient takes the value

$$C = \frac{1}{n}\frac{[\langle k^2\rangle - \langle k\rangle]^2}{\langle k\rangle^3}, \tag{8.24}$$

In networks where k^2 and k have fixed finite values, this quantity becomes small as $n \to \infty$ and hence we expect the clustering coefficient to be very small on large networks. This makes the values in Table 8.1, which are of order 1, quite surprising, and indeed many of them turn out to be much larger than the estimate given by Eq. (8.24). For instance, the collaboration network of physicists is measured to have a clustering coefficient of 0.45. Plugging the appropriate values for n, k, and k^2 into Eq. (8.24) on the other hand gives $C = 0.0023$. Thus the measured value is more than a hundred times greater than the value we would expect if physicists chose their collaborators at random.

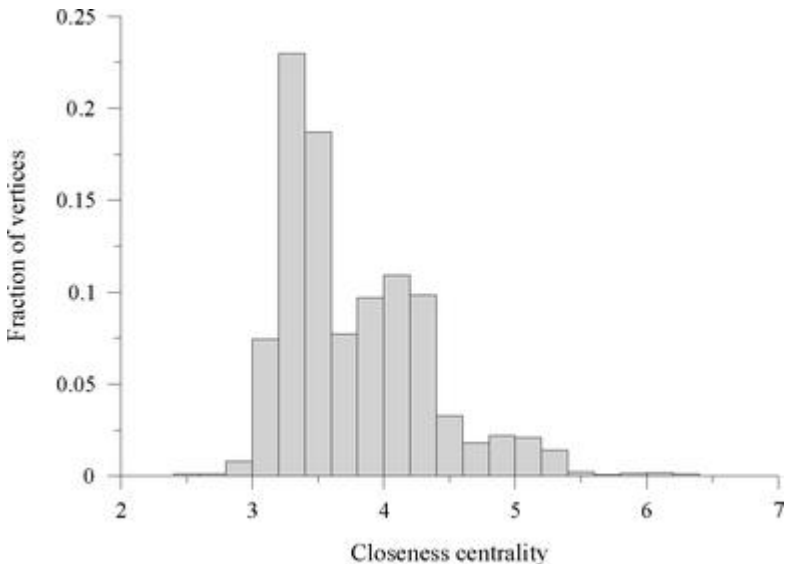


Figure 8.11: Histogram of closeness centralities of vertices on the Internet. Unlike Fig. 8.10 this is a normal non-cumulative histogram showing the actual distribution of closeness centralities. This distribution does not follow a power law.

Presumably this large difference is indicative of real social effects at work. There are a number of reasons why a real collaboration network might contain more triangles than one would expect by chance, but for example it might be that people introduce pairs of their collaborators to one another and those pairs then go on to collaborate themselves. This is an example of the process that social network analysts call *triadic closure*: an "open" triad of vertices (i.e., a triad in which one vertex is linked to the other two, but the third possible edge is absent) is "closed" by the addition of the last edge, forming a triangle.

One can study triadic closure processes directly if one has time-resolved data on the formation of a network. The network of physics collaborators discussed here was studied in this way in Ref. [233], where it was shown that pairs of individuals who have not previously collaborated, but who have another mutual collaborator, are enormously more likely to collaborate in future than pairs who do not—a factor of 45 times as likely in that particular study. Furthermore, the probability of future collaboration also goes up sharply as the number of mutual collaborators increases, with pairs having two mutual collaborators being more than twice as likely to collaborate in future as those having just one.

However, it is not always the case that the measured clustering coefficient greatly exceeds the expected value given by Eq. (8.24). Take the example of the Internet again. For the data set we examined earlier the measured clustering coefficient is just 0.012. The expected value, if connections were made at random, is 0.84. (The large value arises because, as discussed in Section 8.4, the Internet has a highly right-skewed degree distribution, which makes k^2 large.) Clearly in this case the clustering is far *less* than one would expect on the basis of chance, suggesting that in the Internet there are forces at work that shy away from the creation of triangles.[125]

In some other networks, such as food webs or the World Wide Web, clustering is neither higher nor lower than expected, taking values roughly comparable with those given by Eq. (8.24). It is not yet well understood why clustering coefficients take such different values in different types of network, although one theory is that it may be connected with the formation of groups or communities in networks [252].

The clustering coefficient measures the density of triangles in a network. There is no reason, however, for us to limit ourselves to studying only triangles. We can also look at the densities of

other small groups of vertices, or *motifs*, as they are often called. One can define coefficients similar to the clustering coefficient to measure the densities of different motifs, although more often one simply counts the numbers of the motifs of interest in a network. And, as with triangles, one can compare the results with the values one would expect to find if connections in the network are made at random. In general, one can find counts that are higher, lower, or about the same as the expected values, all of which can have implications for the understanding of the networks in question. For example, Milo *et al.* [221] looked at motif counts in genetic regulatory networks and neural networks and found certain small motifs that occurred far more often than was expected on the basis of chance. They conjectured that these motifs were playing the role of functional "circuit elements," such as filters or pulse generators, and that their frequent occurrence in these networks might be an evolutionary result of their usefulness to the organisms involved.

8.6.1 LOCAL CLUSTERING COEFFICIENT

In Section 7.9.1 we introduced the local clustering coefficient for a vertex:

$$C_i = \frac{\text{(number of pairs of neighbors of } i \text{ that are connected)}}{\text{(number of pairs of neighbors of } i\text{)}},$$

(8.25)

which is the fraction of pairs of neighbors of vertex i that are themselves neighbors. If we calculate the local clustering coefficient for all vertices in a network, an interesting pattern emerges in many cases: we find that on average vertices of higher degree tend to have lower local clustering [278, 318]. Figure 8.12, for example, shows the average value of the local clustering coefficient for vertices of degree k on the Internet as a function of k. The decrease of the average C_i with k is clear. It has been conjectured that plots of this type take either the form $C_i \sim k^{-0.75}$ [318] or the form $C_i \sim k^{-1}$ [278]. In this particular case neither of these conjectures matches the data very well, but for some other networks they appear reasonable.

Community structure in networks is discussed at some length in Chapter 11.

On possible explanation for the decrease in C_i with increasing degree is that vertices group together into tightly knit groups or communities, with vertices being connected mostly to others within their own group. In a network showing this kind of behavior vertices that belong to small groups are constrained to have low degree, because they have relatively few fellow group members to connect to, while those in larger groups can have higher degree. (They don't have to have higher degree, but they can.) At the same time, the local clustering coefficient of vertices in small groups will tend to be larger. This occurs because each group, being mostly detached from the rest of the network, functions roughly as its own small network and, as discussed in Section 8.6, smaller networks are expected to have higher clustering. When averaged over many groups of different sizes, therefore, we would expect vertices of lower degree to have higher clustering on average, as in Fig. 8.12.[126]

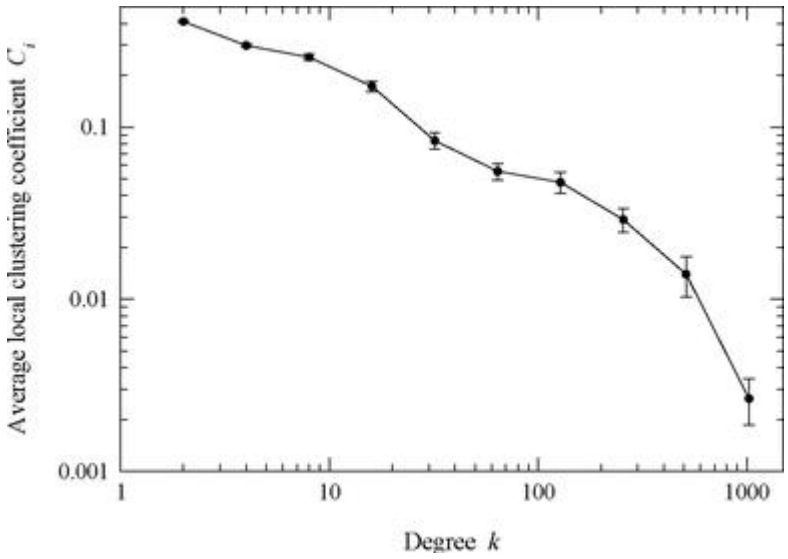


Figure 8.12: Local clustering as a function of degree on the Internet. A plot of the measured mean local clustering coefficient of vertices on the Internet (at the level of autonomous systems) averaged over all vertices with the given degree.

8.7 ASSORTATIVE MIXING

Assortative mixing or homophily is the tendency of vertices to connect to others that are like them in some way. We discussed assortative mixing in Section 7.13, where we gave some examples from social networks, such as the high school friendships depicted in Figs. 7.10 and 7.11 in which school students tend to associate more with others of the same ethnicity or age as themselves.

Of particular interest is assortative mixing by degree, the tendency of vertices to connect others with degrees that are similar to their own. We can also have disassortative mixing by degree, in which vertices connect to others with very different degrees. As we saw in Section 7.13.3, assortative mixing can have substantial effects on the structure of a network (see particularly Fig. 7.12).

Assortative mixing by degree can be quantified in a number of different ways. One of them is to use the correlation coefficient defined in Eq. (7.82):

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)k_i k_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2m)k_i k_j}.$$

(8.26)

If we were going to calculate the value of this coefficient, however, we should not do it directly from this equation, because the double sum over vertices i and j has a lot of terms (n^2 of them) and is slow to evaluate on a computer. Instead we write

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2},$$

(8.27)

with

$$S_e = \sum_{ij} A_{ij} k_i k_j = 2 \sum_{\text{edges } (i,j)} k_i k_j,$$

(8.28)

where the second sum is over all distinct (unordered) pairs of vertices (i, j) connected by an edge, and

$$S_1 = \sum_i k_i, \qquad S_2 = \sum_i k_i^2, \qquad S_3 = \sum_i k_i^3.$$

(8.29)

The computer time needed to calculate network quantities is an important topic in its own right. We discuss the main issues in Chapter 9.

The sum in (8.28) has m terms, where m is the number of edges in the network and the sums in (8.29) have n terms each, so Eq. (8.27) is usually a lot faster to evaluate than Eq. (8.26).

In Table 8.1 we show the values of r for a range of networks and the results reveal an interesting pattern. While none of the values are of very large magnitude—the correlations between degrees are not especially strong—there is a clear tendency for the social networks to have positive r, indicating assortative mixing by degree, while the rest of the networks—technological, information, biological—have negative r, indicating disassortative mixing.

The reasons for this pattern are not known for certain, but it appears that many networks have a tendency to negative values of r because they are simple graphs. As shown by Maslov *et al.* [211], graphs that have only single edges between vertices tend in the absence of other biases to show disassortative mixing by degree because the number of edges that can fall between high-degree vertex pairs is limited. Since most networks are represented as simple graphs this implies that most should be disassortative, as indeed Table 8.1 indicates they are.

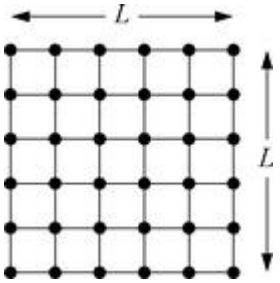
And what about the social networks? One suggestion is that social networks are assortatively mixed because they tend to be divided into groups, as discussed in Section 8.6.1. If a network is divided up into tightly knit groups of vertices that are mostly disconnected from the rest of the network, then, as we have said, vertices in small groups tend to have lower degree than vertices in larger groups. But since the members of small groups are in groups with other members of the same small groups, it follows that the low-degree vertices will tend to be connected to other low-degree vertices, and similarly for high-degree ones. This simple idea can be turned into a quantitative calculation [252] and indeed it appears that, at least under some circumstances, this mechanism does produce positive values of r.

Thus a possible explanation of the pattern of r-values seen in Table 8.1 is that most networks are naturally disassortative by degree because they are simple graphs while social networks (and perhaps a few others) override this natural bias and become assortative by virtue of their group structure.

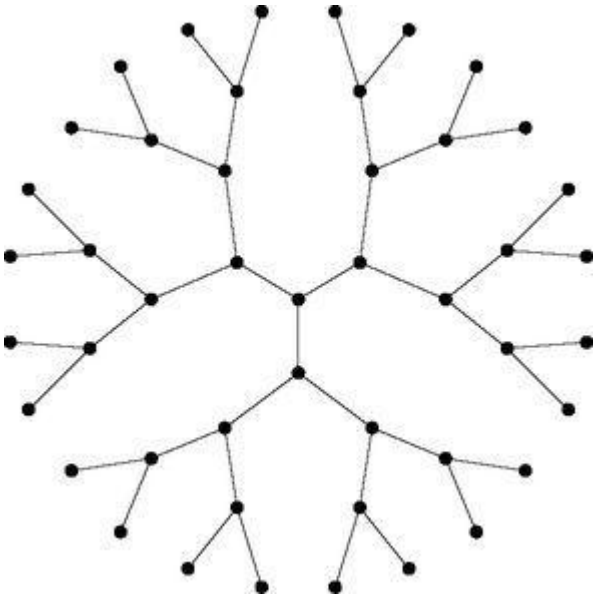
PROBLEMS

8.1 One can calculate the diameter of certain types of network exactly.

a. What is the diameter of a clique?
b. What is the diameter of a square portion of square lattice, with L edges (or equivalently $L + 1$ vertices) along each side, like this:



What is the diameter of the corresponding hypercubic lattice in d dimensions with L edges along each side? Hence what is the diameter of such a lattice as a function of the number n of vertices?
c. A Cayley tree is a symmetric regular tree in which each vertex is connected to the same number k of others, until we get out to the leaves, like this:



(We have $k = 3$ in this picture.)

Show that the number of vertices reachable in d steps from the central vertex is $k(k-1)^{d-1}$ for $d \geq 1$. Hence find an expression for the diameter of the network in terms of k and the

number of vertices n.

d. Which of the networks in parts (i), (ii), and (iii) displays the small-world effect, defined as having a diameter that increases as log n or slower?

8.2 Suppose that a network has a degree distribution that follows the exponential form $p_k = Ce^{-\lambda k}$, where C and λ are constants.

a. Find C as a function of λ.

b. Calculate the fraction P of vertices that have degree k or greater.

c. Calculate the fraction W of ends of edges that are attached to vertices of degree k or greater.

d. Hence show that for the exponential degree distribution with exponential parameter λ, the Lorenz curve—the equivalent of Eq. (8.23)—is given by

$$W = P - \frac{1-e^{\lambda}}{\lambda} P \ln P.$$

e. Show that the value of W is greater than one for some values of P in the range $0 \le P \le 1$. What is the meaning of these "unphysical" values?

8.3 A particular network is believed to have a degree distribution that follows a power law. Among a random sample of vertices in the network, the degrees of the first 20 vertices with degree 10 or greater are:

16	17	10	26	13
14	28	45	10	12
12	10	136	16	25
36	12	14	22	10

Estimate the exponent α of the power law and the error on that estimate using Eqs. (8.6) and (8.7).

8.4 Consider the following simple and rather unrealistic mathematical model of a network. Each of n vertices belongs to one of several groups. The mth group has n_m vertices and each vertex in that group is connected to others in the group with independent probability $p_m = A(n_m - 1)^{-\beta}$, where A and β are constants, but not to any vertices in other groups. Thus this network takes the form of a set of disjoint clusters or communities.

a. Calculate the expected degree k of a vertex in group m.

b. Calculate the expected value $\overline{C}_m$ of the local clustering coefficient for vertices in group m.

c. Hence show that $\overline{C}_m \propto \langle k \rangle^{-\beta/(1-\beta)}$.

d. What value would β have to have for the expected value of the local clustering to fall off with increasing degree as $k^{-3/4}$?