CHAPTER **12**

RANDOM GRAPHS

An introduction to the most basic of network models, the random graph

SO FAR in this book we have looked at how we measure the structure of networks and at mathematical, statistical, and computational methods for making sense of the network data we get from our measurements. We have seen for instance how to measure the structure of the Internet, and once we have measured it how to determine its degree distribution, or the centrality of its vertices, or the best division of the network into groups or communities. An obvious next question to ask is, "If I know a network has some particular property, such as a particular degree distribution, what effect will that have on the wider behavior of the system?" It turns out that properties like degree distributions can in fact have huge effects on networked systems, which is one of the main reasons we are interested in them. And one of the best ways to understand and get a feel for these effects is to build mathematical models. The remainder of this book is devoted to the examination of some of the many network models in common use.

In Chapters 12 to 15 we consider models of the structure of networks, models that mimic the patterns of connections in real networks in an effort to understand the implications of those patterns. In Chapters 16 to 19 we consider models of processes taking place on networks, such as epidemics on social networks or search engines on the Web. In many cases these models of network processes are themselves built on top of our models of network structure, combining the two to shed light on the interplay between structure and dynamics in networked systems.

In Section 8.4, for instance, we noted that many networks have degree distributions that roughly follow a power law—the so-called scale-free networks.

A reasonable question would be to ask how the structure and behavior of such scale-free networks differs from that of their non-scale-free counterparts. A good way to address this question would be to create, on a computer for example, two artificial networks, one with a power-law degree distribution and one without, and explore their differences empirically. Better still, one could create a large number of networks in each of the two classes, to see what statistically significant features appear in one class and not in the other. This is precisely the rationale behind random graph models, which are the topic of this chapter and the following one. In random graph models, one creates networks that possess particular properties of interest, such as specified degree distributions, but which are otherwise random. Random graphs are interesting in their own right for the light they shed on the structural properties of networks, but have also been widely used as a substrate for models of dynamical processes *on* networks. In Chapter 17, for instance, we examine their use in epidemic modeling.

We also look at a number of other types of network model in succeeding chapters. In Chapter 14 we look at generative models of networks, models in which the network is "grown" according to a specified set of growth rules. Generative models are particularly useful for understanding how network structure arises in the first place. By growing networks according to a variety of different rules and comparing the results with real networks, we can get a feel for which growth processes are plausible and which can be ruled out. In Chapter 15 we look at "small-world models," which model the phenomenon of network transitivity or clustering (see Section 7.9), and at "exponential random graphs," which are particularly useful when we want to create model networks that match the properties of observed networks as closely as possible.

12.1 RANDOM GRAPHS

In general, a *random graph* is a model network in which some specific set of parameters take fixed values, but the network is random in other respects. One of the simplest examples of a random graph is the network in which we fix only the number of vertices n and the number of edges m. That is, we take n vertices and place m edges among them at random. More precisely, we choose m pairs of vertices uniformly at random from all possible pairs and connect them with an edge. Typically one stipulates that the network should be a simple graph, i.e., that it should have no multiedges or self-edges (see Section 6.1), in which case the position of each edge should be chosen among only those pairs that are distinct and not already connected.[1] This model is often referred to by its mathematical name $G(n,m)$.

Another entirely equivalent definition of the model is to say that the network is created by choosing uniformly at random among the set of all simple graphs with exactly n vertices and m edges.

Strictly, in fact, the random graph model is not defined in terms of a single randomly generated network, but as an *ensemble* of networks, i.e., a probability distribution over possible networks. Thus the model $G(n,m)$ is correctly defined as a probability distribution $P(G)$ over all graphs G in which $P(G) = 1/\Omega$ for simple graphs with n vertices and m edges and zero otherwise, where Ω is the total number of such simple graphs. We will see more complicated examples of random graph ensembles shortly.

When one talks about the properties of random graphs one typically means the average properties of the ensemble. For instance, the "diameter" of $G(n,m)$ would mean the diameter $\ell(G)$ of a graph G, averaged over the ensemble thus

$$\langle \ell \rangle = \sum_G P(G)\ell(G) = \frac{1}{\Omega}\sum_G \ell(G). \tag{12.1}$$

This is a useful definition for a several of reasons. First, it turns out to lend itself well to analytic calculations; many such average properties of random graphs can be calculated exactly, at least in the limit of large graph size. Second, it often reflects exactly the thing we want to get at in making our model network in the first place. Very often we are interested in the typical properties of networks. We might want to know, for instance, what the typical diameter is of a network with a given number of edges. Certainly there are special cases of such networks that have particularly large or small diameters, but these don't reflect the typical behavior. If it's typical behavior we are after, then the ensemble average of a property is often a good guide. Third, it can be shown that the distribution of values for many network measures is sharply peaked, becoming concentrated more and more narrowly around the ensemble average as the size of the network becomes large, so that in the large n limit essentially all values one is likely to encounter are very close to the mean.

Some properties of the random graph $G(n,m)$ are straightforward to calculate: obviously the average number of edges is m, for instance, and the average degree is $\langle k \rangle = 2m/n$. Unfortunately, other properties are not so easy to calculate, and most mathematical work has actually been conducted on a slightly different model that is considerably easier to handle. This model is called

[1] It would in theory be perfectly possible, however, to create a variant of the model with multiedges or self-edges, or both.

$G(n, p)$. In $G(n, p)$ we fix not the number but the *probability* of edges between vertices. Again we have n vertices, but now we place an edge between each distinct pair with independent probability p. In this network the number of edges is not fixed. Indeed it is possible that the network could have no edges at all, or could have edges between every distinct pair of vertices. (For most values of p these are not likely outcomes, but they could happen.)

Again, the technical definition of the random graph is not in terms of a single network, but in terms of an ensemble, a probability distribution over all possible networks. To be specific, $G(n, p)$ is the ensemble of networks with n vertices in which each simple graph G appears with probability

$$P(G) = p^m (1-p)^{\binom{n}{2}-m}, \tag{12.2}$$

where m is the number of edges in the graph, and non-simple graphs have probability zero.

$G(n, p)$ was first studied, to this author's knowledge, by Solomonoff and Rapoport [303], but it is most closely associated with the names of Paul Erdős and Alfréd Rényi, who published a celebrated series of papers about the model in the late 1950s and early 1960s [105–107]. If you read scientific papers on this subject, you will sometimes find the model referred to as the "Erdős–Rényi model" or the "Erdős–Rényi random graph" in honor of their contribution. It is also sometimes called the "Poisson random graph" or the "Bernoulli random graph," names that refer to the distributions of degrees and edges in the model. And sometimes the model is referred to simply as "the" random graph—there are many random graph models, but $G(n, p)$ is the most fundamental and widely studied of them, so if someone is talking about a random graph but doesn't bother to mention which one, they are probably thinking of this one.

In this chapter we describe the basic mathematics of the random graph $G(n, p)$, focusing particularly on the degree distribution and component sizes, which are two of the model's most illuminating characteristics. The techniques we develop in this chapter will also prove useful for some of the more complex models examined later in the book.

12.2 MEAN NUMBER OF EDGES AND MEAN DEGREE

Let us start our study of the random graph $G(n, p)$ with a very simple calculation, the calculation of the expected number of edges in our model network. We have said that the number of edges in the model is not fixed, but we can calculate its mean or expectation value as follows. The number of graphs with exactly n vertices and m edges is equal to the number of ways of picking the positions of the edges from the $\binom{n}{2}$ distinct vertex pairs. Each of these graphs appears with the same probability $P(G)$, given by Eq. (12.2), and hence the total probability of drawing a graph with m edges from our ensemble is

$$P(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}, \tag{12.3}$$

which is just the standard binomial distribution. Then the mean value of m is

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} mP(m) = \binom{n}{2} p. \tag{12.4}$$

This result comes as no surprise. The expected number of edges between any individual pair of vertices is just equal to the probability p of an edge between the same vertices, and Eq. (12.4) thus says merely that the expected total number of edges in the network is equal to the expected number p between any pair of vertices, multiplied by the number of pairs.

We can use this result to calculate the mean degree of a vertex in the random graph. As pointed out in the previous section, the mean degree in a graph with exactly m edges is $\langle k \rangle = 2m/n$, and hence the mean degree in $G(n, p)$ is

$$\langle k \rangle = \sum_{m=0}^{\binom{n}{2}} \frac{2m}{n} P(m) = \frac{2}{n} \binom{n}{2} p = (n-1)p, \tag{12.5}$$

where we have used Eq. (12.4) and the fact that n is constant. The mean degree of a random graph is often denoted c in the literature, and we will adopt this convention here also, writing

$$c = (n-1)p. \tag{12.6}$$

This result is also unsurprising. It says that the expected number of edges connected to a vertex is equal to the expected number p between the vertex and any other vertex, multiplied by the number $n-1$ of other vertices.

12.3 DEGREE DISTRIBUTION

Only slightly more taxing is the calculation of the degree distribution of $G(n, p)$. A given vertex in the graph is connected with independent probability p to each of the $n-1$ other vertices. Thus the probability of being connected to a particular k other vertices and not to any of the others is $p^k(1-p)^{n-1-k}$. There are $\binom{n-1}{k}$ ways to choose those k other vertices, and hence the total probability of being connected to exactly k others is

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \tag{12.7}$$

which is a binomial distribution again. In other words, $G(n,p)$ has a binomial degree distribution.

In many cases we are interested in the properties of large networks, so that n can be assumed to be large. Furthermore, as discussed in Section 6.9, many networks have a mean degree that is approximately constant as the network size becomes large. (For instance, the typical number of friends a person has does not depend strongly on the total number of people in the world.) In such a case Eq. (12.7) simplifies as follows.

Equation (12.6) tells us that $p = c/(n-1)$ will become vanishingly small as $n \to \infty$, which allows us to write

$$\begin{aligned} \ln\left[(1-p)^{n-1-k}\right] &= (n-1-k)\ln\left(1-\frac{c}{n-1}\right) \\ &\simeq -(n-1-k)\frac{c}{n-1} \simeq -c, \end{aligned} \tag{12.8}$$

where we have expanded the logarithm as a Taylor series, and the equalities become exact as $n \to \infty$. Taking exponentials of both sizes, we thus find that $(1-p)^{n-1-k} = \mathrm{e}^{-c}$ in the large-n limit. Also for large n we have

$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!\,k!} \simeq \frac{(n-1)^k}{k!}, \tag{12.9}$$

and thus Eq. (12.7) becomes

$$p_k = \frac{(n-1)^k}{k!} p^k \mathrm{e}^{-c} = \frac{(n-1)^k}{k!}\left(\frac{c}{n-1}\right)^k \mathrm{e}^{-c} = \mathrm{e}^{-c}\frac{c^k}{k!}, \tag{12.10}$$

in the limit of large n.

Equation (12.10) is the Poisson distribution: in the limit of large n, $G(n,p)$ has a Poisson degree distribution. This is the origin of the name *Poisson random graph,* which we will use occasionally to distinguish this model from some of the more sophisticated random graphs in the following chapter that don't in general have Poisson degree distributions.

12.4 CLUSTERING COEFFICIENT

A very simple quantity to calculate for the Poisson random graph is the clustering coefficient. Recall that the clustering coefficient C is a measure of the transitivity in a network (Section 7.9) and is defined as the probability that two network neighbors of a vertex are also neighbors of each other. In a random graph the probability that *any* two vertices are neighbors is exactly the same—all such probabilities are equal to $p = c/(n-1)$. Hence

$$C = \frac{c}{n-1}. \tag{12.11}$$

This is one of several respects in which the random graph differs sharply from most from real-world networks, many of which have quite high clustering coefficients—see Table 8.1—while Eq. (12.11) tends to zero in the limit $n \to \infty$ if the mean degree c stays fixed. This discrepancy is discussed further in Section 12.8.

12.5 GIANT COMPONENT

Consider the Poisson random graph $G(n,p)$ for $p = 0$. In this case there are no edges in the network at all and it is completely disconnected. Each vertex is an island on its own; the network has n separate components of exactly one vertex each.

In the opposite limit, when $p = 1$, every possible edge in the network is present and the network is an n-vertex clique in the technical sense of the word (see Section 7.8.1) meaning that every vertex is connected directly to every other. In this case, all the vertices are connected together in a single component that spans the entire network.

Now let us focus on the size of the largest component in the network in each of these cases. In the first case ($p = 0$) the largest component has size 1. In the second ($p = 1$) the largest component has size n. Apart from the second being much larger than the first, there is an important qualitative difference between these two cases: in the first case the size of the largest component is independent of the number of vertices n in the network; in the second it is proportional to n, or *extensive* in the jargon of theoretical physics. In the first case, the largest component will stay the same size if we make the network larger, but in the second it will grow with the network.

The distinction between these two cases is an important one. In many applications of networks it is crucial that there be a component that fills most of the network. For instance, in the Internet it is important that there be a path through the network from most computers to most others. If there were not, the network wouldn't be able to perform its intended role of providing computer-to-computer communications for its users. Moreover, as discussed in Section 8.1, most networks do in fact have a large component that fills most of the network. We can gain some useful insights about what is happening in such networks by considering how the components in our random graph behave. Although the random graph is a very simple network model and doesn't provide an accurate representation of the Internet or other real-world networks, we will see that when trying to understand the world it can be very helpful to study such simplified models.

So let us consider the largest component of our random graph, which, as we have said, has constant size 1 when $p = 0$ and extensive size n when $p = 1$. An interesting question to ask is how the transition between these two extremes occurs if we construct random graphs with gradually increasing values of p, starting at 0 and ending up at 1. We might guess, for instance, that the size of the largest component somehow increases gradually with p, becoming extensive only in the limit where $p = 1$. In reality, however, something much more interesting happens. As we will see, the size of the largest component undergoes a sudden change, or *phase transition*, from constant size to extensive size at one particular special value of p. Let us take a look at this transition.

A network component whose size grows in proportion to n we call a *giant component*. We can calculate the size of the giant component in the Poisson random graph exactly in the limit of large network size $n \to \infty$ as follows. We denote by u the average fraction of vertices in the random graph that do *not* belong to the giant component. Thus if there is no giant component in our graph, we will have $u = 1$, and if there is a giant component we will have $u < 1$. Alternatively, we can regard u as the probability that a randomly chosen vertex in the graph does not belong to the giant component.

For a vertex i not to belong to the giant component it must not be connected to the giant component via any other vertex. That means that for every other vertex j in the graph either (a) i is not connected to j by an edge, or (b) i is connected to j but j is itself not a member of the giant component. The probability of outcome (a) is simply $1 - p$, the probability of not having an edge between i and j, and the probability of outcome (b) is pu, where the factor of p is the probability of having an edge and the factor of u is the probability that vertex j doesn't belong to the giant component.[2] Thus the total probability of not being connected to the giant component via vertex j is $1 - p + pu$.

Then the total probability of not being connected to the giant component

[2] We need to be a little careful here: u here should really be the probability that j is not connected to the giant component via any of its connections other than the connection to i. However, it turns out that in the limit of large system size this probability is just equal to u. For large n the probability of not being connected to the giant component via any of the $n - 2$ vertices other than i is not significantly smaller than the probability for all $n - 1$ vertices.

via any of the $n-1$ other vertices in the network is

$$u = (1-p+pu)^{n-1} = \left[1 - \frac{c}{n-1}(1-u)\right]^{n-1}, \tag{12.12}$$

where we have used Eq. (12.6). Now we take logs of both sides thus:

$$\begin{aligned}\ln u &= (n-1)\ln\left[1 - \frac{c}{n-1}(1-u)\right] \\ &\simeq -(n-1)\frac{c}{n-1}(1-u) = -c(1-u),\end{aligned} \tag{12.13}$$

where the approximate equality becomes exact in the limit of large n. Taking exponentials of both sides, we then find that

$$u = \mathrm{e}^{-c(1-u)}. \tag{12.14}$$

But if u is the fraction of vertices not in the giant component, then the fraction of vertices that are in the giant component is $S = 1-u$. Eliminating u in favor of S then gives us

$$S = 1 - \mathrm{e}^{-cS}. \tag{12.15}$$

This equation, which was first given by Erdős and Rényi in 1959 [105], tells us the size of the giant component as a fraction of the size of the network in the limit of large network size, for any given value of the mean degree c. Unfortunately, though the equation is very simple it doesn't have a simple solution for S in closed form.[3] We can however get a good feeling for its behavior from a graphical solution. Consider Fig. 12.1. The three curves show the function $y = 1 - \mathrm{e}^{-cS}$ for different values of c. Note that S can take only values from zero to one, so only this part of the curve is shown. The dashed line in the

[3] One can write a closed-form solution in terms of the *Lambert W-function*, which is defined as the solution to the equation $W(z)\mathrm{e}^{W(z)} = z$. In terms of this function the size of the giant component is

$$S = 1 + \frac{W(-c\mathrm{e}^{-c})}{c},$$

where we take the principal branch of the W-function. This expression may have some utility for numerical calculations and series expansions, but it is not widely used. Alternatively, although we cannot write a simple solution for S as a function of c, we can write a solution for c as a function of S. Rearranging Eq. (12.15) for c gives

$$c = -\frac{\ln(1-S)}{S},$$

which can be useful, for instance, for plotting purposes. (We can make a plot of S as a function of c by first making a plot of c as a function of S and then swapping the axes.)

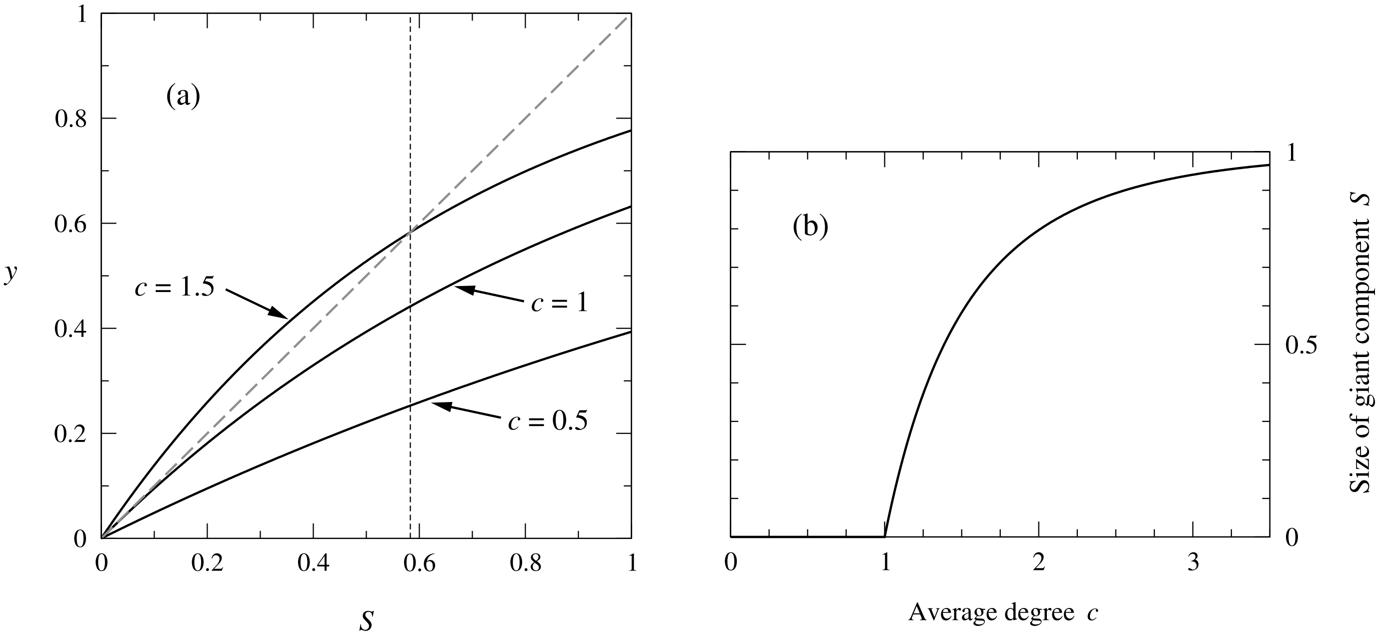


Figure 12.1: Graphical solution for the size of the giant component. (a) The three curves in the left panel show $y = 1 - \mathrm{e}^{-cS}$ for values of c as marked, the diagonal dashed line shows $y = S$, and the intersection gives the solution to Eq. (12.15), $S = 1 - \mathrm{e}^{-cS}$. For the bottom curve there is only one intersection, at $S = 0$, so there is no giant component, while for the top curve there is a solution at $S = 0.583\ldots$ (vertical dashed line). The middle curve is precisely at the threshold between the regime where a non-trivial solution for S exists and the regime where there is only the trivial solution $S = 0$. (b) The resulting solution for the size of the giant component as a function of c.

figure is the function $y = S$. Where line and curve cross we have $S = 1 - \mathrm{e}^{-cS}$ and the corresponding value of S is a solution to Eq. (12.15).

As the figure shows, depending on the value of c there may be either one solution for S or two. For small c (bottom curve in the figure) there is just one solution at $S = 0$, which implies that there is no giant component in the network. (You can confirm for yourself that $S = 0$ is a solution directly from Eq. (12.15).) On the other hand, if c is large enough (top curve) then there are two solutions, one at $S = 0$ and one at $S > 0$. Only in this regime can there be a giant component.

The transition between the two regimes corresponds to the middle curve in the figure and falls at the point where the gradient of the curve and the gradient of the dashed line match at $S = 0$. That is, the transition takes place when

$$\frac{\mathrm{d}}{\mathrm{d}S}\left(1 - \mathrm{e}^{-cS}\right) = 1, \tag{12.16}$$

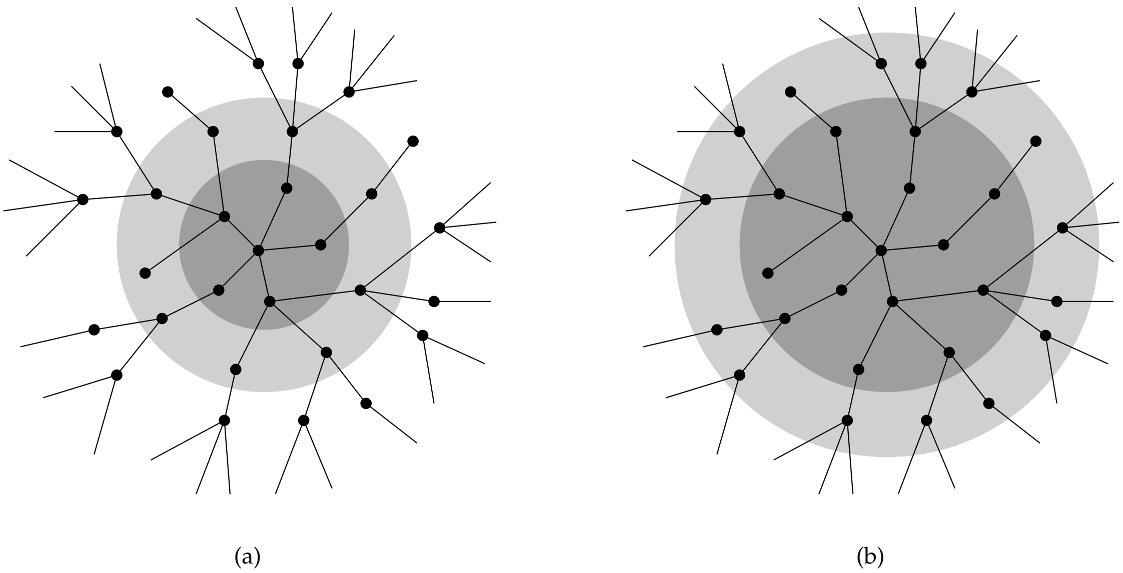


Figure 12.2: Growth of a vertex set in a random graph. (a) A set of vertices (inside the gray circles) consists of a core (dark gray) and a periphery (lighter). (b) If we grow the set by adding to it those vertices immediately adjacent to the periphery, then the periphery vertices become a part of the new core and a new periphery is added.

or

$$ce^{-cS} = 1. \tag{12.17}$$

Setting $S = 0$ we then deduce that the transition takes place at $c = 1$.

In other words, the random graph can have a giant component only if $c > 1$. At $c = 1$ and below we have $S = 0$ and there is no giant component.

This does not entirely solve the problem, however. Technically we have proved that there can be no giant component for $c \leq 1$, but not that there has to be a giant component at $c > 1$—in the latter regime there are two solutions for S, one of which is the solution $S = 0$ in which there is no giant component. So which of these solutions is the correct one that describes the true size of the giant component?

In answering this question, we will see another way to think about the formation of the giant component. Consider the following process. Let us find a small set of connected vertices somewhere in our network—say a dozen or so, as shown in Fig. 12.2a. In the limit of large $n \to \infty$ such a set is bound to exist somewhere in the network, so long as $c > 0$. We will divide the set into its *core* and its *periphery*. The core is the vertices that have connections only to other vertices in the set—the darker gray region in the figure. The *periphery* is the vertices that have at least one neighbor outside the set—the lighter gray.

Now imagine enlarging our set by adding to it all those vertices that are immediate neighbors, connected by at least one edge to the set—Fig. 12.2b. Now the old periphery is part of the core and there is a new periphery consisting of the vertices just added. How big is this new periphery? We don't know for certain, but we know that each vertex in the old periphery is connected with independent probability p to every other vertex. If there are s vertices in our set, then there are $n - s$ vertices outside the set, and the average number of connections a vertex in the periphery has to outside vertices is

$$p(n-s) = c\frac{n-s}{n-1} \simeq c, \tag{12.18}$$

where the equality becomes exact in the limit $n \to \infty$. This means that the average number of immediate neighbors of the set—the size of the new periphery when we grow the set—is c times the size of the old periphery.

We can repeat this argument, growing the set again and again, and each time the average size of the periphery will increase by another factor of c. Thus if $c > 1$ the average size of the periphery will grow exponentially. On the other hand, if $c < 1$ it will shrink exponentially and eventually dwindle to zero. Furthermore, if it grows exponentially our connected set of vertices will eventually form a component comparable in size to the whole network—a giant component—while if it dwindles the set will only ever have finite size and no giant component will form.

So we see that indeed we expect a giant component if (and only if) $c > 1$. And when there is a giant component the size of that giant component will be given by the larger solution to Eq. (12.15). This now allows us to calculate the size of the giant component for all values of c. (For $c > 1$ we have to solve for the larger solution of Eq. (12.15) numerically, since there is no exact solution, but this is easy enough to do.) The results are shown in Fig. 12.1. As the figure shows, the size of the giant component grows rapidly from zero as the value of c passes 1, and tends towards $S = 1$ as c becomes large.

12.6 SMALL COMPONENTS

In this section we look at the properties of random graphs from a different point of view, the point of view of the non-giant components. We have seen that in a random graph with $c > 1$ there exists a giant component that fills an extensive fraction of the network. That fraction is typically less than 100%, however. What is the structure of the remainder of the network? The answer is that it is made up of many small components whose average size is constant and doesn't increase with the size of the network.

The first step in demonstrating this result and shedding light on the structure of the small components is to show that there is only one giant component in a random graph, and hence that all other components are "non-giant" components. This is fairly easy to establish. Suppose that there were two or more giant components in a random graph. Take any two giant components, which have size $S_1 n$ and $S_2 n$, where S_1 and S_2 are the fractions of the network filled by each. The number of distinct pairs of vertices (i, j), where i is in the first giant component and j is in the second, is just $S_1 n \times S_2 n = S_1 S_2 n^2$. Each of these pairs is connected by an edge with probability p, or not with probability $1 - p$. For the two giant components to be separate components we require that there be zero edges connecting them together, which happens with probability q given by

$$q = (1-p)^{S_1 S_2 n^2} = \left(1 - \frac{c}{n-1}\right)^{S_1 S_2 n^2}, \tag{12.19}$$

where we have made use of Eq. (12.6).

Taking logs of both sides and going to the limit $n \to \infty$, we then find

$$\begin{aligned} \ln q &= S_1 S_2 \lim_{n\to\infty} \left[n^2 \ln\left(1 - \frac{c}{n-1}\right)\right] = S_1 S_2 \left[-c(n+1) + \tfrac{1}{2}c^2\right] \\ &= cS_1 S_2 \left[-n + \left(\tfrac{1}{2}c - 1\right)\right], \end{aligned} \tag{12.20}$$

where we have dropped terms of order $1/n$. Taking the exponential again, we get

$$q = q_0 \, \mathrm{e}^{-cS_1 S_2 n}, \tag{12.21}$$

where $q_0 = \mathrm{e}^{c(c/2-1)S_1 S_2}$, which is independent of n if c is constant. Thus, for constant c, the probability that the two giant components are really separate components dwindles exponentially with increasing n, and in the limit of large n will vanish altogether. In a large random graph, therefore, there is only the very tiniest of probabilities that we will have two giant components, and for infinite n the probability is formally zero and it will never happen.

Given then that there is only one giant component in our random graph and that in most situations it does not fill the entire network, it follows that there must also be some non-giant components, i.e., components whose size does not increase in proportion to the size of the network. These are the *small components*.

12.6.1 SIZES OF THE SMALL COMPONENTS

The small components can, in general, come in various different sizes. We can calculate the distribution of these sizes as follows.

The basic quantity we focus on is the probability π_s that a randomly chosen vertex belongs to a small component of size exactly s vertices total. Note that if there is a giant component in our network then some vertices do not belong to a small component of any size and hence π_s is not normalized to unity. The sum of π_s over all sizes s is equal to the fraction of vertices that are not in the giant component. That is,

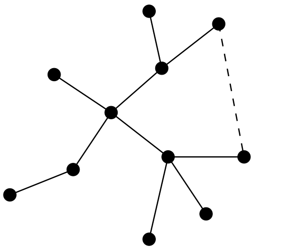
$$\sum_{s=0}^{\infty} \pi_s = 1 - S, \tag{12.22}$$

where S is, as before, the fraction of vertices in the giant component.

Recall that a tree is a graph or subgraph that has no loops—see Section 6.7.

The crucial insight that allows us to calculate π_s is that the small components are trees, as we can see by the following argument. Consider a small component of s vertices that takes the form of a tree. A tree of s vertices contains $s-1$ edges, as shown in Section 6.7, and this is the smallest number of edges that is needed to connect this many vertices together. If we add another edge to our component then we will create a loop, since we will be adding a new path between two vertices that are already connected (see figure). In a Poisson random graph the probability of such edge being present is the same as for any other edge, $p = c/(n-1)$. The total number of places where we could add such an extra edge to the component is given by the number of distinct pairs of vertices minus the number that are already connected by an edge, or

$$\binom{s}{2} - (s-1) = \tfrac{1}{2}(s-1)(s-2), \tag{12.23}$$



If we add an edge (dashed) to a tree we create a loop.

and the total number of extra edges in the component is $\frac{1}{2}(s-1)(s-2)c/(n-1)$. Assuming that s increases more slowly than $\sqrt{n}$ (and we will shortly see that it does), this probability tends to zero in the limit $n \to \infty$, and hence there are no loops in the component and the component is a tree.

We can use this observation to calculate the probability π_s as follows. Consider a vertex i in a small component of a random graph, as depicted in Fig. 12.3. Each of i's edges leads to a separate subgraph—the shaded regions in the figure—and because the whole component is a tree we know that these subgraphs are not connected to one another, other than via vertex i, since if they were there would be a loop in the component and it would not be a tree. Thus the size of the component to which i belongs is the sum of the sizes of the subgraphs reachable along each of its edges, plus 1 for vertex i itself. To put

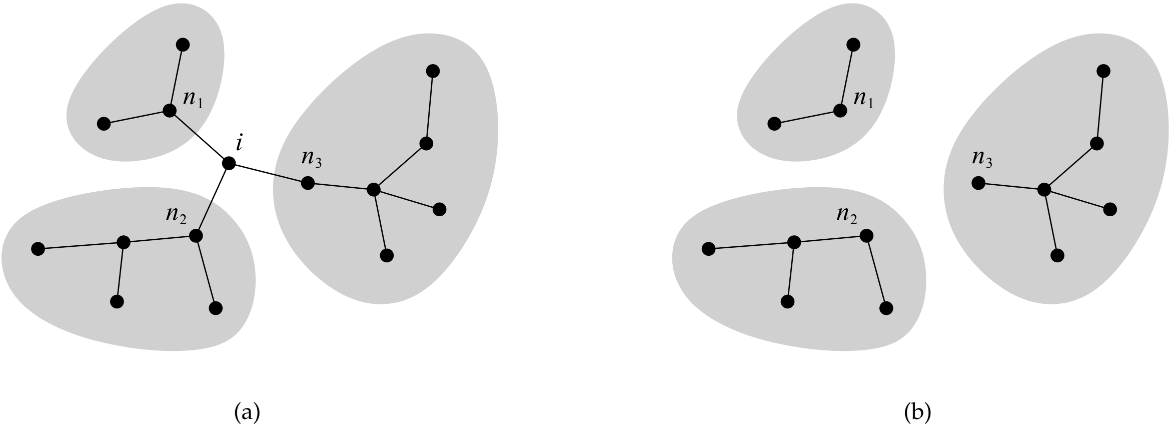


Figure 12.3: The size of one of the small components in a random graph. (a) The size of the component to which a vertex i belongs is the sum of the number of vertices in each of the subcomponents (shaded regions) reachable via i's neighbors n_1, n_2, n_3, plus one for i itself. (b) If vertex i is removed the subcomponents become components in their own right.

that another way, vertex i belongs to a component of size s if the sizes of the subgraphs to which its neighbors $n_1, n_2, \ldots$ belong sum to $s - 1$.

Bearing this in mind, consider now a slightly modified network, the network in which vertex i is completely removed, along with all its edges.[4] This network is still a random graph with the same value of p—each possible edge is still present with independent probability p—but the number of vertices has decreased by one, from n to $n - 1$. In the limit of large n, however, this decrease is negligible. The average properties, such as size of the giant component and size of the small components will be indistinguishable for random graphs with sizes n and $n - 1$, but the same p.

In this modified network, what were previously the subgraphs of our small component are now separate small components in their own right. And since the network has the same average properties as the original network for large n, that means that the probability that neighbor n_1 belongs to a small component of size s_1 (or a subgraph of size s_1 in the original network) is itself given by π_{s_1}. We can use this observation to develop a self-consistent expression for

[4] In the statistical physics literature, this trick of removing a vertex is called a *cavity method*. Cavity methods are used widely in the solution of all kinds of physics problems and are a powerful method for many calculations on lattices and in low-dimensional spaces as well as on networks [218].

the probability π_s.

Suppose that vertex i has degree k. As we have said, the probability that neighbor n_1 belongs to a small component of size s_1 when i is removed from the network is π_{s_1}. So the probability $P(s|k)$ that vertex i belongs to a small component of size s, given that its degree is k, is the probability that its k neighbors belong to small components of sizes $s_1, \ldots, s_k$—which is $\prod_{j=1}^{k} \pi_{s_j}$—and that those sizes add up to $s - 1$:

$$P(s|k) = \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta\bigl(s-1, \textstyle\sum_j s_j\bigr), \tag{12.24}$$

where $\delta(m, n)$ is the Kronecker delta.

To get π_s, we now just average $P(s|k)$ over the distribution p_k of the degree thus:

$$\begin{aligned} \pi_s = \sum_{k=0}^{\infty} p_k P(s|k) &= \sum_{k=0}^{\infty} p_k \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta\bigl(s-1, \textstyle\sum_j s_j\bigr) \\ &= \mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta\bigl(s-1, \textstyle\sum_j s_j\bigr), \end{aligned} \tag{12.25}$$

where we have made use of Eq. (12.10) for the degree distribution of the random graph.

This expression would be easy to evaluate if it were not for the delta function: one could separate the terms in the product, distribute them among the individual summations, and complete the sums in closed form. With the delta function, however, it is difficult to see how the sum can be completed.

Luckily there is a trick for problems like these, a trick that we will use many times in the rest of this book. We introduce a *generating function* or *z-transform*, defined by

$$h(z) = \pi_1 z + \pi_2 z^2 + \pi_3 z^3 + \ldots = \sum_{s=1}^{\infty} \pi_s z^s. \tag{12.26}$$

This generating function is a polynomial or series in z whose coefficients are the probabilities π_s. It encapsulates all of the information about the probability distribution in a single function. Given $h(z)$ we can recover the probabilities by differentiating:

$$\pi_s = \frac{1}{s!} \frac{\mathrm{d}^s h}{\mathrm{d}z^s}\bigg|_{z=0}. \tag{12.27}$$

Thus $h(z)$ is a complete representation of our probability distribution and if we can calculate it, then we can calculate π_s. We will look at generating functions in more detail in the next section, but for now let us complete the present calculation.

We can calculate $h(z)$ by substituting Eq. (12.25) into Eq. (12.26), which gives

$$\begin{aligned}
h(z) &= \sum_{s=1}^{\infty} z^s \mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta\bigl(s-1, \textstyle\sum_j s_j\bigr) \\
&= \mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] z^{1+\sum_j s_j} \\
&= z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} z^{s_j} \right] \\
&= z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \left[\sum_{s=1}^{\infty} \pi_s z^s \right]^k = z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \bigl[h(z)\bigr]^k \\
&= z \exp\bigl[c\bigl(h(z)-1\bigr)\bigr].
\end{aligned} \tag{12.28}$$

Thus we have a simple, self-consistent equation for $h(z)$ that eliminates the awkward delta function of (12.25).

Unfortunately, like the somewhat similar Eq. (12.15), this equation doesn't have a known closed-form solution for $h(z)$, but that doesn't mean the expression is useless. In fact we can calculate many useful things from it without solving for $h(z)$ explicitly. For example, we can calculate the mean size of the component to which a randomly chosen vertex belongs, which is given by

$$\langle s \rangle = \frac{\sum_s s\pi_s}{\sum_s \pi_s} = \frac{h'(1)}{1-S}, \tag{12.29}$$

where $h'(z)$ denotes the first derivative of $h(z)$ with respect to its argument and we have made use of Eqs. (12.22) and (12.26). (The denominator in this expression is necessary because π_s is not normalized to 1.)

From Eq. (12.28) we have

$$\begin{aligned}
h'(z) &= \exp\bigl[c\bigl(h(z)-1\bigr)\bigr] + czh'(z)\exp\bigl[c\bigl(h(z)-1\bigr)\bigr] \\
&= \frac{h(z)}{z} + ch(z)h'(z),
\end{aligned} \tag{12.30}$$

or, rearranging,

$$h'(z) = \frac{h(z)}{z[1-ch(z)]}, \tag{12.31}$$

and thus

$$h'(1) = \frac{h(1)}{1-ch(1)}. \tag{12.32}$$

But $h(1) = \sum_s \pi_s = 1 - S$, from Eqs. (12.22) and (12.26), so that

$$h'(1) = \frac{1-S}{1-c+cS}. \tag{12.33}$$

And so the average size $\langle s \rangle$ of Eq. (12.29) becomes

$$\langle s \rangle = \frac{1}{1-c+cS}. \tag{12.34}$$

When $c < 1$ and there is no giant component, this gives simply $\langle s \rangle = 1/(1-c)$. When there is a giant component, the behavior is more complicated, because we have to solve for S first before finding the value of $\langle s \rangle$, but the calculation can still be done. We first solve Eq. (12.15) for S and then substitute into Eq. (12.34).

It's interesting to note that Eq. (12.34) diverges when $c = 1$. (At this point $S = 0$, so the denominator vanishes.) Thus, if we slowly increase the mean degree c of our network from some small initial value less than 1, the average size of the component to which a vertex belongs gets bigger and bigger and finally becomes infinite exactly at the point where the giant component appears. For $c > 1$ Eq. (12.34) measures only the sizes of the non-giant components and the equation tells us that these get smaller again above $c = 1$. Thus the general picture we have is in one in which the small components get larger up to $c = 1$, where they diverge and the giant component appears, then smaller again as the giant component grows larger. Figure 12.4 shows a plot of $\langle s \rangle$ as a function of c with the divergence clearly visible.

Although the random graph is certainly not a realistic model of most networks, this general picture of the component structure of the network turns out to be a good guide to the behavior of networks in the real world. If a network has a low density of edges then typically it consists only of small components, but if the density is becomes enough then a single large component forms, usually accompanied by many separate small ones. Moreover, the small components tend on average to be smaller if the largest component is very large. This is a good example of the way in which simple models of networks can give us a feel for how more complicated real-world systems should behave in general.

12.6.2 AVERAGE SIZE OF A SMALL COMPONENT

A further important point to notice about Eq. (12.34) is that the average size of the small components does not grow with the number of vertices n. The typical size of the small components in a random graph remains constant as

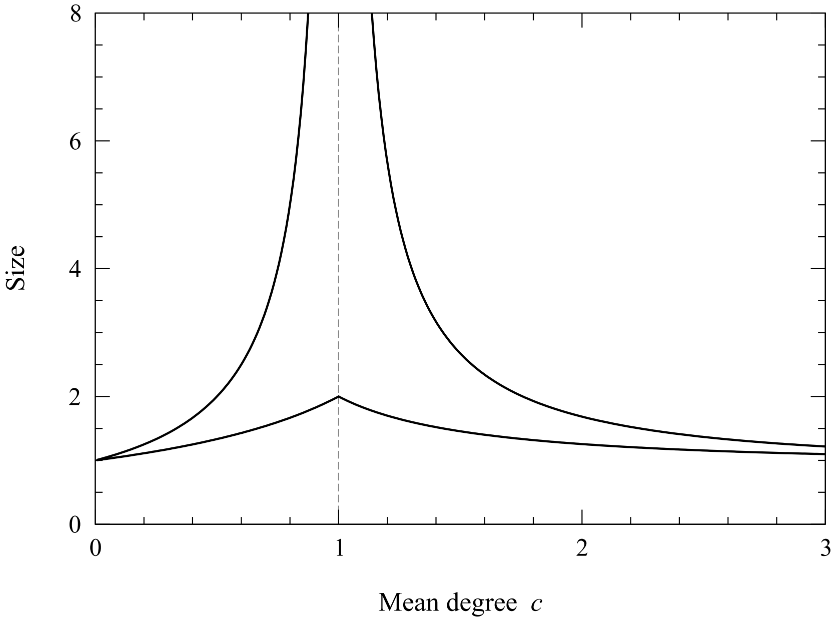


Figure 12.4: Average size of the small components in a random graph. The upper curve shows the average size $\langle s \rangle$ of the component to which a randomly chosen vertex belongs, calculated from Eq. (12.34). The lower curve shows the overall average size R of a component, calculated from Eq. (12.40). The dotted vertical line marks the point $c = 1$ at which the giant component appears. Note that, as discussed in the text, the upper curve diverges at this point but the lower one does not.

the graph gets larger. We must, however, be a little careful with these statements. Recall that π_s is the probability that a randomly chosen vertex belongs to a component of size s, and hence $\langle s \rangle$ as calculated here is not strictly the average size of a component, but the average size of the component to which a randomly chosen vertex belongs. Because larger components have more vertices in them, the chances of landing on them when we choose a random vertex is larger, in proportion to their size, and hence $\langle s \rangle$ is a biased estimate of the actual average component size. To get a correct figure for the average size of a component we need to make a slightly different calculation.

Let n_s be the actual number of components of size s in our random graph. Then the number of vertices that belong to components of size s is sn_s and hence the probability of a randomly chosen vertex belonging to such a component is

$$\pi_s = \frac{sn_s}{n}. \tag{12.35}$$

The average size of a component, which we will denote R, is

$$R = \frac{\sum_s s n_s}{\sum_s n_s} = \frac{n \sum_s \pi_s}{n \sum_s \pi_s / s} = \frac{1 - S}{\sum_s \pi_s / s}, \tag{12.36}$$

where we have made use of Eq. (12.22). The remaining sum we can again evaluate using our generating function by noting that

$$\int_0^1 \frac{h(z)}{z}\,\mathrm{d}z = \sum_{s=1}^{\infty} \pi_s \int_0^1 z^{s-1}\,\mathrm{d}z = \sum_{s=1}^{\infty} \frac{\pi_s}{s}. \tag{12.37}$$

A useful expression for $h(z)/z$ can be obtained by rearranging Eq. (12.31) to yield

$$\frac{h(z)}{z} = \bigl[1 - c h(z)\bigr] \frac{\mathrm{d}h}{\mathrm{d}z}, \tag{12.38}$$

and hence we find that

$$\begin{aligned} \sum_{s=1}^{\infty} \frac{\pi_s}{s} &= \int_0^1 \bigl[1 - c h(z)\bigr] \frac{\mathrm{d}h}{\mathrm{d}z} \mathrm{d}z = \int_0^{1-S} (1 - c h)\,\mathrm{d}h \\ &= 1 - S - \tfrac{1}{2} c (1 - S)^2, \end{aligned} \tag{12.39}$$

where we have used $h(1) = \sum_s \pi_s = 1 - S$ for the upper integration limit.

Substituting this result into Eq. (12.36), we find that the average component size is

$$R = \frac{2}{2 - c + cS}. \tag{12.40}$$

As with Eq. (12.34), this expression is independent of n, so the average size of a small component indeed does not grow as the graph becomes large.

On the other hand, R does not diverge at $c = 1$ as $\langle s \rangle$ does. At $c = 1$, with $S = 0$, Eq. (12.40) gives just $R = 2$. The reason for this is that, while the largest component in the network for $c = 1$ does become infinite in the limit of large n, so also does the total number of components. So the average size of a component is the ratio of two diverging quantities. Depending on the nature of the divergences, such a ratio could be infinite itself, or zero, or finite but nonzero in the special case where the two divergences have the same asymptotic form. In this instance the latter situation holds—both quantities are diverging linearly with n—and the average component size remains finite. A plot of R is included in Fig. 12.4 for comparison with $\langle s \rangle$.

12.6.3 THE COMPLETE DISTRIBUTION OF COMPONENT SIZES

So far we have calculated the average size of a small component in the random graph, but not the individual probabilities π_s that specify the complete distribution of sizes. In principle, we should be able to calculate the π_s by solving

Eq. (12.28) for the generating function $h(z)$ and then differentiating according to Eq. (12.27) to get π_s. Unfortunately we cannot follow this formula in practice because, as mentioned above, Eq. (12.28) does not have a known solution.

Remarkably, however, it turns out that we can still calculate the values of the individual π_s, by an alternative route. The calculations involve some more advanced mathematical techniques and if you are not particularly interested in the details it will do no harm to skip this section. If you're interested in this rather elegant development, however, read on.

To calculate an explicit expression for the probabilities π_s of the component sizes we make use of a beautiful result from the theory of complex variables, the *Lagrange inversion formula*. The Lagrange inversion formula is a formula that allows the explicit solution of equations of the form

$$f(z) = z\phi(f(z)) \tag{12.41}$$

for the unknown function $f(z)$, where $\phi(f)$ is a known function which at $f = 0$ is finite, non-zero, and differentiable.

Equation (12.41) has precisely the form of the equation for our generating function, Eq. (12.28). What's more, the Lagrange formula gives a solution for $f(z)$ in terms of the coefficients of the series expansion of $f(z)$ in powers of z, which is precisely what we want in the present case, since the coefficients are the probabilities π_s, which is what we want to calculate. The Lagrange formula is thus perfectly suited to the problem in hand. Here we first derive the general form of the formula then apply it to the current problem.[5]

Let us write the function $f(z)$ in Eq. (12.41) as a series expansion thus:

$$f(z) = \sum_{s=1}^{\infty} a_s z^s, \tag{12.42}$$

The coefficient a_s in this expansion is given explicitly by

$$a_s = \frac{1}{s!}\frac{\mathrm{d}^s f}{\mathrm{d}z^s}\bigg|_{z=0} = \frac{1}{s!}\left[\frac{\mathrm{d}^{s-1}}{\mathrm{d}z^{s-1}}\left(\frac{\mathrm{d}f}{\mathrm{d}z}\right)\right]_{z=0}. \tag{12.43}$$

Cauchy's formula for the nth derivative of a function $g(z)$ at $z = z_0$ says that

$$\frac{\mathrm{d}^n g}{\mathrm{d}z^n}\bigg|_{z=z_0} = \frac{n!}{2\pi\mathrm{i}}\oint \frac{g(z)}{(z-z_0)^{n+1}}\,\mathrm{d}z, \tag{12.44}$$

[5] The formula derived here is not the *most* general form of the Lagrange inversion formula. It is adequate for the particular problem we are interested in solving, but the full Lagrange inversion formula is even more powerful, and can solve a broader range of problems. For details, see Wilf [329].

where the integral is around a contour that encloses z_0 in the complex plane but encloses no poles in $g(z)$. We will use an infinitesimal circle around z_0 as our contour.

Applying Cauchy's formula to (12.43) with $g(z) = f'(z)$, $z_0 = 0$, and $n = s-1$, we get

$$a_s = \frac{1}{2\pi i s} \oint \frac{1}{z^s} \frac{\mathrm{d}f}{\mathrm{d}z}\,\mathrm{d}z = \frac{1}{2\pi i s} \oint \frac{\mathrm{d}f}{z^s}, \tag{12.45}$$

where the second integral is now around a contour in f rather than z. In this equation we are now thinking of z as being a function of f, $z = z(f)$, rather than the other way around. We are perfectly entitled to do this—knowing either quantity specifies the value of the other.[6]

It will be important later that the contour followed by f surrounds the origin, so let us pause for a moment to demonstrate that it does. Our choice of contour for z in the first integral of Eq. (12.45) is an infinitesimal circle around the origin. Expanding Eq. (12.41) to leading order around the origin, we find that

$$f(z) = z\phi(f(0)) + \mathrm{O}(z^2) = z\phi(0) + \mathrm{O}(z^2), \tag{12.46}$$

where we have made use of the fact that $f(0) = 0$, which is easily seen from Eq. (12.41) given that $\phi(f)$ is non-zero and finite at $f = 0$ by hypothesis. In the limit of small $|z|$ where the terms of order z^2 can be neglected, Eq. (12.46) implies that f traces a contour about the origin if z does, since the two are proportional to one another.

We now rearrange our original equation, Eq. (12.41), to give the value of z in terms of f thus

$$z(f) = \frac{f}{\phi(f)}, \tag{12.47}$$

and then substitute into Eq. (12.45) to get

$$a_s = \frac{1}{2\pi i s} \oint \frac{[\phi(f)]^s}{f^s}\,\mathrm{d}f. \tag{12.48}$$

Since, as we have said, the contour encloses the origin, this expression can be written in terms of a derivative evaluated at the origin by again making use of Cauchy's formula, Eq. (12.44):

$$a_s = \frac{1}{s!} \left[\frac{\mathrm{d}^{s-1}}{\mathrm{d}f^{s-1}} [\phi(f)]^s \right]_{f=0}. \tag{12.49}$$

[6] The situation gets complicated if $z(f)$ is many-valued for some f, i.e., if $f(z)$ is non-monotonic. In our case, however, where the coefficients in the expansion of $f(z)$ are necessarily all non-negative because they are probabilities, $f(z)$ is monotonically increasing and no such problems arise.

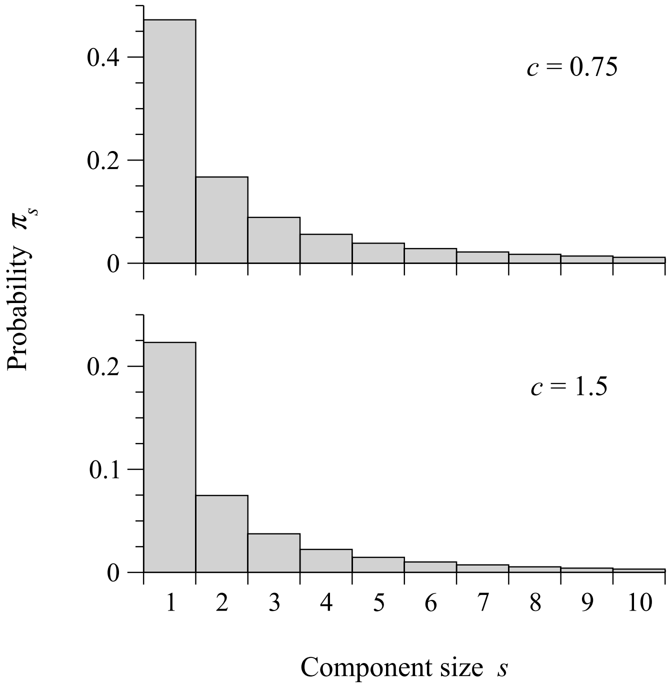


Figure 12.5: Sizes of small components in the random graph. This plot shows the probability π_s that a randomly chosen vertex belongs to a small component of size s in a Poisson random graph with $c = 0.75$ (top), which is in the regime where there is no giant component, and $c = 1.5$ (bottom), where there is a giant component.

This is the Lagrange inversion formula. This remarkably simple formula gives us, in effect, a complete series solution to Eq. (12.41).

To apply the formula to the current problem, of the component size distribution for the random graph, we set $f(z) \to h(z)$ and $\phi(f) \to \mathrm{e}^{c(h-1)}$. Then the coefficients π_s of $h(z)$ are given by

$$\pi_s = \frac{1}{s!}\left[\frac{\mathrm{d}^{s-1}}{\mathrm{d}h^{s-1}}\mathrm{e}^{sc(h-1)}\right]_{h=0} = \frac{\mathrm{e}^{-sc}(sc)^{s-1}}{s!}. \qquad (12.50)$$

These are the probabilities that a randomly chosen vertex belongs to a small component of size s in a random graph with mean degree c. Figure 12.5 shows the shape of π_s as a function of s for two different values of c. As the plot shows, the distribution is heavily skewed, with many components of small size and only a few larger ones.

12.7 PATH LENGTHS

In Sections 3.6 and 8.2 we discussed the small-world effect, the observation that the typical lengths of paths between vertices in networks tend to be short. Most people find the small-world effect surprising upon first learning about it.

We can use the random graph model to shed light on how the effect arises by examining the behavior of the network diameter in the model.

See Section 6.10.1 for a discuss of geodesic distances and diameters.

Recall that the diameter of a network is the longest geodesic distance between any two vertices in the same component of the network. As we now show, the diameter of a random graph varies with the number n of vertices as $\ln n$. Since $\ln n$ is typically a relatively small number even when n is large, this offers some explanation of the small-world effect, although it also leaves some questions open, as discussed further below.

The basic idea behind the estimation of the diameter of a random graph is simple. As discussed in Section 12.5, the average number of vertices s steps away from a randomly chosen vertex in a random graph is c^s. Since this number grows exponentially with s it doesn't take very many such steps before the number of vertices reached is equal to the total number of vertices in the whole network; this happens when $c^s \simeq n$ or equivalently $s \simeq \ln n / \ln c$. At this point, roughly speaking, every vertex is within s steps of our starting point, implying that the diameter of the network is approximately $\ln n / \ln c$.

Although the random graph is, as we have said, not an accurate model of most real-world networks, this is, nonetheless, believed to be the basic mechanism behind the small-world effect in most networks: the number of vertices within distance s of a particular starting point grows exponentially s and hence the diameter is logarithmic in n. We discuss the comparison with real-world networks in more detail below.

The argument above is only approximate. It's true that there are on average c^s vertices s steps away from any starting point so long as s is small. But once c^s becomes comparable with n the result has to break down since clearly the number of vertices at distance s cannot exceed the number of vertices in the whole graph. (Indeed it cannot exceed the number in the giant component.)

One way to deal with this problem is to consider two different starting vertices i and j. The average numbers of vertices s and t steps from them respectively will then be equal to c^s and c^t so long as we stay in the regime where both these numbers are much less than n. In the following calculation we consider only configurations in which both remain smaller than order n in the limit $n \to \infty$ so as to satisfy this condition.

The situation we consider is depicted in Fig. 12.6, with the two vertices i and j each surrounded by a "ball" or neighborhood consisting of all vertices with distances up to and including s and t respectively. If there is an edge between the "surface" (i.e., most distant vertices) of one neighborhood and the surface of the other, as depicted by the dashed line, then it is straightforward to show that there is also an edge between the surfaces of any pair of neighborhoods with larger s or t (or both). Turning that statement around, if there

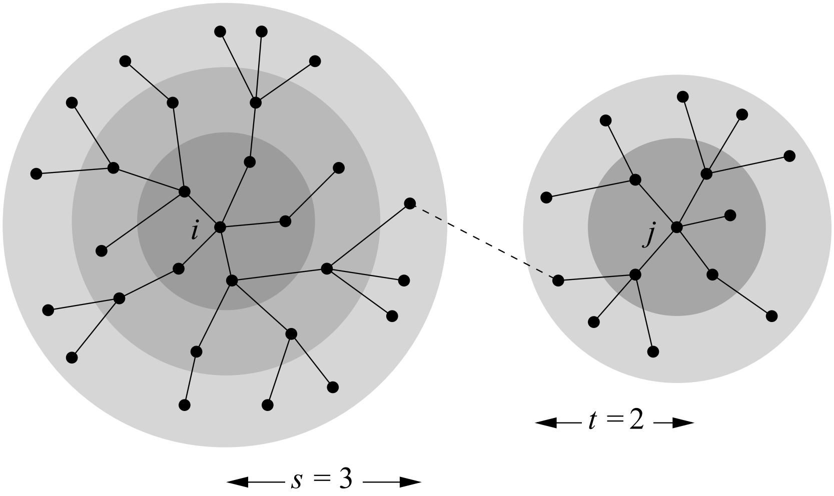


Figure 12.6: Neighborhoods of two vertices in a random graph. In the argument given in the text we consider the sets of vertices within distances s and t respectively of two randomly chosen vertices i and j. If there is an edge between any vertex on the surface of one neighborhood and any vertex on the surface of the other (dashed line), then there is a path between i and j of length $s + t + 1$.

is no edge between the surfaces of our neighborhoods, then there is also no edge between any smaller neighborhoods, which means that the shortest path between i and j must have length greater than $s + t + 1$. The reverse is also trivially true, that a shortest path longer than $s + t + 1$ implies there is no edge between our surfaces. Thus the absence of an edge between the surfaces is a necessary and sufficient condition for the distance d_{ij} between i and j to be greater than $s + t + 1$. This in turn implies that the probability $P(d_{ij} > s + t + 1)$ is equal to the probability that there is no edge between the two surfaces.

There are on average $c^s \times c^t$ pairs of vertices such that one lies on each surface, and each pair is connected with probability $p = c/(n-1) \simeq c/n$ (assuming n to be large) or not with probability $1 - p$. Hence $P(d_{ij} > s + t + 1) = (1-p)^{c^{s+t}}$. Defining for convenience $\ell = s + t + 1$, we can also write this as

$$P(d_{ij} > \ell) = (1-p)^{c^{\ell-1}} = \left(1 - \frac{c}{n}\right)^{c^{\ell-1}}. \qquad (12.51)$$

Taking logs of both sides, we find

$$\ln P(d_{ij} > \ell) = c^{\ell-1} \ln\left(1 - \frac{c}{n}\right) \simeq -\frac{c^{\ell}}{n}, \qquad (12.52)$$

where the approximate inequality becomes exact as $n \to \infty$. Thus in this limit

$$P(d_{ij} > \ell) = \exp\left(-\frac{c^{\ell}}{n}\right). \tag{12.53}$$

The diameter of the network is the smallest value of ℓ such that $P(d_{ij} > \ell)$ is zero, i.e., the value such that no matter which pair of vertices we happen to pick there is zero chance that they will be separated by a greater distance. In the limit of large n, Eq. (12.53) will tend to zero only if c^{ℓ} grows faster than n, meaning that our smallest value of ℓ is the value such that $c^{\ell} = an^{1+\epsilon}$ with a constant and $\epsilon \to 0$ from above. Note that we can, as promised, achieve this while keeping both c^s and c^t smaller than order n, so that our argument remains valid.

Rearranging for ℓ, we now find our expression for the diameter:

$$\ell = \frac{\ln a}{\ln c} + \lim_{\epsilon \to 0} \frac{(1+\epsilon)\ln n}{\ln c} = A + \frac{\ln n}{\ln c}, \tag{12.54}$$

where A is a constant.[7] Apart from the constant, this is the same result as we found previously using a rougher argument. The constant is known—it has a rather complicated value in terms of the Lambert W-function [114]—but for our purposes the important point is that it is (asymptotically) independent of n. Thus the diameter indeed increases only slowly with n, as $\ln n$, making it relatively small in large random graphs.

The logarithmic dependence of the diameter on n offers some explanation of the small-world effect of Section 3.6. Even in a network such as the acquaintance network of the entire world, with nearly seven billion inhabitants (at the time of writing), the value of $\ln n / \ln c$ can be quite small. Supposing each person to have about a thousand acquaintances,[8] we would get

$$\ell = \frac{\ln n}{\ln c} = \frac{\ln 6 \times 10^9}{\ln 1000} = 3.3\ldots, \tag{12.55}$$

which is easily small enough to account for the results of, for example, the small-world experiments of Milgram and others [93, 219, 311].

[7] There are still some holes in our argument. In particular, we have assumed that the product of the numbers of vertices on the surface of our two neighborhoods is c^{s+t} when in practice this is only the average value and there will in general be some variation. Also the calculation should really be confined to the giant component, since the longest path always falls in the giant component in the limit of large n. For a careful treatment of these issues see, for instance, Fernholz and Ramachandran [114].

[8] This appears to be a reasonable figure. Bernard *et al.* [36] estimated the typical number of acquaintances for people in the United States to be about 2000—see Section 3.2.1.

On the other hand, although this calculation gives us some insight into the nature of the small-world effect, this cannot be the entire explanation. There are clearly many things wrong with the random graph as a model of real social networks, as we now discuss.

12.8 PROBLEMS WITH THE RANDOM GRAPH

The Poisson random graph is one of the best studied models of networks. In the half century since its first proposal it has given us a tremendous amount of insight into the expected structure of networks of all kinds, particularly with respect to component sizes and network diameters. The fact that it is both simple to describe and straightforward to study using analytic methods makes it an excellent tool for investigating all sorts of network phenomena. We will return to the random graph many times in the remainder of this book to help us understand the way networks behave.

The random graph does, however, have some severe shortcomings as a network model. There are many ways in which it is completely unlike the real-world networks we have seen in the previous chapters. One clear problem is that it shows essentially no transitivity or clustering. In Section 12.4 we saw at the clustering coefficient of a random graph is $C = c/(n-1)$, which tends to zero in the limit of large n. And even for the finite values of n appropriate to real-world networks the value of C in the random graph is typically very small. For the acquaintance network of the human population of the world, with its $n \simeq 7$ billion people, each having about 1000 acquaintances [175], a random graph with the same n and c would have a clustering coefficient of

$$C \simeq \frac{1000}{7\,000\,000\,000} \simeq 10^{-7}. \tag{12.56}$$

Whether the clustering coefficient of the real acquaintance network is 0.01 or 0.5 hardly matters. (It is probably somewhere in between.) Either way it is clear that the random graph and the true network are in strong disagreement.[9]

The random graph also differs from real-world networks in many other ways. For instance, there is no correlation between the degrees of adjacent vertices—necessarily so, since the edges are placed completely at random. The degrees in real networks, by contrast, are usually correlated, as discussed in Section 8.7. Many, perhaps most, real-world networks also show grouping of

[9] This disagreement, highlighted particularly by Watts and Strogatz [323], was one of the observations that prompted the current wave of interest in the properties of networks in the mathematical sciences, starting in the late 1990s.

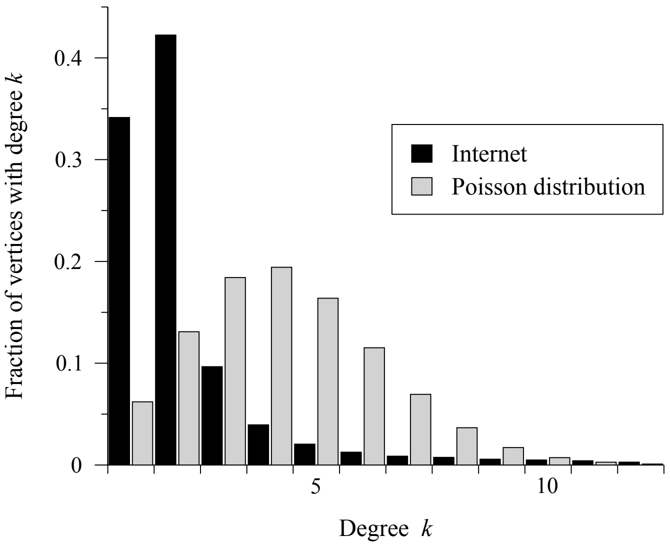


Figure 12.7: Degree distribution of the Internet and a Poisson random graph. The dark bars in this plot show the fraction of vertices with the given degrees in the network representation of the Internet at the level of autonomous systems. The lighter bars represent the same measure for a random graph with the same average degree as the Internet. Even though the two distributions have the same averages, it is clear that they are entirely different in shape.

their vertices into "communities," as discussed on Section 11.2.1, but random graphs have no such structure. And there are many other examples of interesting structure in real networks that is absent from the random graph.

However, perhaps the most significant respect in which the properties of random graphs diverge from those of real-world networks is the shape of their degree distribution. As discussed in Section 8.3, real networks typically have right-skewed degree distributions, with most vertices having low degree but with a small number of high-degree "hubs" in the tail of the distribution. The random graph on the other hand has a Poisson degree distribution, Eq. (12.10), which is not right-skewed to any significant extent. Consider Fig. 12.7, for example, which shows a histogram of the degree distribution of the Internet (darker bars), measured at the level of autonomous systems (Section 2.1.1). The right-skewed form is clearly visible in this example. On the same figure we show the Poisson degree distribution of a random graph (lighter bars) with the same average degree c as the Internet example. Despite having the same averages, the two distributions are clearly entirely different. It turns out that this difference has a profound effect on all sorts of properties of the network—we will see many examples in this book. This makes the Poisson random graph inadequate to explain many of the interesting phenomena we see in networks today, including resilience phenomena, epidemic spreading processes, percolation, and many others.

Luckily it turns out to be possible to generalize the random graph model to allow for non-Poisson degree distributions. This development, which leads to some of the most beautiful results in the mathematics of networks, is described in the next chapter.

PROBLEMS

12.1 Consider the random graph $G(n,p)$ with mean degree c.

a) Show that in the limit of large n the expected number of triangles in the network is $\frac{1}{6}c^3$. This means that the number of triangles is constant, neither growing nor vanishing in the limit of large n.

b) Show that the expected number of connected triples in the network (as defined on page 200) is $\frac{1}{2}nc^2$.

c) Hence calculate the clustering coefficient C, as defined in Eq. (7.41), and confirm that it agrees for large n with the value given in Eq. (12.11).

12.2 Consider the random graph $G(n,p)$ with mean degree c.

a) Argue that the probability that a vertex of degree k belongs to a small component is $(1-S)^k$, where S is the fraction of the network occupied by the giant component.

b) Thus, using Bayes' theorem (or otherwise) show that the fraction of vertices in small components that have degree k is $\mathrm{e}^{-c}c^k(1-S)^{k-1}/k!$.

12.3 Starting from the generating function $h(z)$ defined in Eq. (12.26), or otherwise, show that

a) the mean-square size of the component in a random graph to which a randomly chosen vertex belongs is $1/(1-c)^3$ in the regime where there is no giant component;

b) the mean-square size of a randomly chosen component in the same regime is $1/[(1-c)(1-\frac{1}{2}c)]$.

Note that both quantities diverge at the phase transition where the giant component appears.

12.4 In Section 7.8.2 we introduced the idea of a bicomponent. A vertex in a random graph belongs to a bicomponent if two or more of its neighbors belong to the giant component of the network (since the giant component completes a loop between those neighbors forming a bicomponent). In principle, a vertex can also be in a bicomponent if two or more of its neighbors belong to the same small component, but in practice this never happens, since that would imply that the small component in question contained a loop and, as we have seen, the small components in a random graph are trees and so have no loops.

a) Show that the fraction of vertices in a random graph that belong to a bicomponent is $S_2 = (1 - cu)(1 - u)$, where u is defined by Eq. (12.14).

b) Show that this expression can be rewritten as $S_2 = S + (1 - S)\ln(1 - S)$, where S is the size of the giant component.

c) Hence argue that the random graph contains a giant bicomponent whenever it contains an ordinary giant component.

12.5 The *cascade model* is a simple mathematical model of a directed acyclic graph, sometimes used to model food webs. We take n vertices labeled $i = 1 \ldots n$ and place an undirected edge between each distinct pair with independent probability p, just as in the ordinary random graph. Then we add directions to the edges such that each edge runs from the vertex with numerically higher label to the vertex with lower label. This ensures that all directed paths in the network run from higher to lower labels and hence that the network is acyclic, as discussed in Section 6.4.2.

a) Show that the average in-degree of vertex i in the ensemble of the cascade model is $\langle k_i^{\text{in}} \rangle = (n - i)p$ and the average out-degree is $\langle k_i^{\text{out}} \rangle = (i - 1)p$.

b) Show that the expected number of edges that connect to vertices i and lower from vertices above i is $(ni - i^2)p$.

c) Assuming n is even, what are the largest and smallest values of this quantity and where do they occur?

In a food web this expected number of edges from high- to low-numbered vertices is a rough measure of energy flow and the cascade model predicts that energy flow will be largest in the middle portions of a food web and smallest at the top and bottom.

12.6 We can make a simple random graph model of a network with clustering or transitivity as follows. We take n vertices and go through each distinct trio of three vertices, of which there are $\binom{n}{3}$, and with independent probability p we connect the members of the trio together using three edges to form a triangle, where $p = c/\binom{n-1}{2}$ with c a constant.

a) Show that the mean degree of a vertex in this model network is $2c$.

b) Show that the degree distribution is

$$p_k = \begin{cases} \mathrm{e}^{-c} c^{k/2}/(k/2)! & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

c) Show that the clustering coefficient, Eq. (7.41), is $C = 1/(2c + 1)$.

d) Show that when there is a giant component in the network its expected size S as a fraction of network size satisfies $S = 1 - \mathrm{e}^{-cS(2-S)}$.

e) What is the value of the clustering coefficient when the giant component fills half of the network?

CHAPTER 13

RANDOM GRAPHS WITH GENERAL DEGREE DISTRIBUTIONS

This chapter describes more sophisticated random graph models that mimic networks with arbitrary degree distributions

IN THE previous chapter we looked at the classic random graph model, in which pairs of vertices are connected at random with uniform probabilities. Although this model has proved tremendously useful as a source of insight into the structure of networks, it also has, as described in Section 12.8, a number of serious shortcomings. Chief among these is its degree distribution, which follows the Poisson distribution and is quite different from the degree distributions seen in most real-world networks. In this chapter we show how we can create more sophisticated random graph models, which incorporate arbitrary degree distributions and yet are still exactly solvable for many of their properties in the limit of large network size.

The fundamental mathematical tool that we will use to derive the results of this chapter is the probability generating function. We have already seen in Section 12.6 one example of a generating function, which was useful in the calculation of the distribution of component sizes in the Poisson random graph. We begin this chapter with a more formal introduction to generating functions and to some of their properties which will be useful in later calculations. Readers interested in pursuing the mathematics of generating functions further may like to look at the book by Wilf [329].[1]

[1] Professor Wilf has generously made his book available for free in electronic form. You can download it from `www.math.upenn.edu/~wilf/DownldGF.html`.