Chapter 2

Poisson Random Graph

2.1 Introduction

In this chapter, we will start studying the theory of *random graph models*. This topic is one of the most important topics in the field of networks, and a large fraction of the research has been devoted to it by the network scientists. Let us start with understanding the meaning of the term 'random graph model'. Consider the problem of spread of an epidemic or some information on networks; maybe you want to study how a rumour spreads in social networks. In particular, suppose that you want to study whether having more triangles in the network accelerates the spread of the rumour or slows it down. One way to do this is to take a real-world network with a high density of triangles, and then simulate the spread of rumour on it. However, this doesn't allow you to *vary* the density of triangles to systematically study the effect of number of triangles on the spread. Random graph models (or just random graphs for short) provide precisely this type of facility. In a nutshell, random graphs are mathematical models of networks in which certain properties of interest could be fixed leaving the network otherwise random. They are useful when we want to mimic only certain properties of the real world networks such as the degree distribution or the density of triangles. It is important to keep in mind that these models don't try to explain *why* networks possess a particular property. They just let you reproduce that property so that you can work with it.

Random graphs are usually defined in terms of a probability distribution over an *ensemble* of graphs. Ensemble is simply the set of all possible graphs given the parameters and constraints. For example, a set of all graphs with 10 vertices and 5 edges is an ensemble. If we specify the probability of drawing each graph from the ensemble, we have the probability distribution defined over it.

We will start with a very simple random graph model called $G(n, m)$. In this model, one starts with n isolated vertices, and m edges are distributed randomly between the total $\binom{n}{2}$ pairs. Thus, there are total $\binom{\binom{n}{2}}{m}$ members in the ensemble, and $G(n, m)$ is actually specified by defining the uniform probability distribution on the ensemble. In other words, $G(n, m)$ is a set of graphs with n vertices and m edges so that each graph in the set (or ensemble) equally likely.

2.2 Erdős-Rényi model

A much more flexible model is the model studied by two Hungarian mathematicians Paul Erdős and Alfred Rényi in late 1950s. The Erdős-Rényi graph is constructed by starting with n isolated vertices, and then each pair is connected by an **undirected** edge with probability p. Because of this definition, the model is also known as $G(n, p)$. In effect, we have an ensemble of graphs with n vertices with a certain probability distribution defined over it. Evidently, the number of edges in this model is not fixed unlike $G(n, m)$, and can vary anywhere from 0 (no edges present) to $\binom{n}{2}$ (all possible edges present). Let us try to see what is the probability distribution over the ensemble. Consider a graph G in the ensemble with exactly m edges. The probability of generating G using $G(n, p)$ is:

$$P(G) = p^m (1-p)^{\binom{n}{2}-m} \tag{2.1}$$

Thus, the probability distribution over the ensemble is not uniform, again unlike $G(n, m)$, and depending upon the value of p, some graphs have much higher probability of appearance than the others. (Question: what happens for $p = 0.5$?)

Though the total number of edges in the graph is not fixed, we can ask what is the average or mean number of edges in the graph. This could be calculated by averaging the value of m over the probability distribution over the ensemble.

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m p^m (1-p)^{\binom{n}{2}-m} \tag{2.2}$$

In principle, one could perform this sum and obtain the desired result. However, we could reduce much of the efforts by using the fact that the expectation value operator is linear. Since each edge is present with probability p and absent with probability $1 - p$, each edge is a Bernoulli random variable. Thus, the total number of edges in a graph is given by:

$$e = \sum_{i=0}^{\binom{n}{2}} e_i \tag{2.3}$$

where e_i represents the value for i^{th} edge (0 if absent, 1 if present). Taking expectation of both the sides, and using linearity of the expectation,

$$\langle m \rangle = \mathbb{E}(e) = \mathbb{E}\left(\sum_{i=1}^{\binom{n}{2}} e_i\right) = \sum_{i=1}^{\binom{n}{2}} \mathbb{E}(e_i) = \sum_{i=1}^{\binom{n}{2}} p = \binom{n}{2} p \tag{2.4}$$

We can rewrite this expression in terms of the average degree c of the graph. A given vertex is connected to each of the remaining $(n - 1)$ vertices with probability p, and hence the mean degree of the vertex is $c = (n - 1)p$. This provides us with a useful expression:

$$p = \frac{c}{n-1} \tag{2.5}$$

Many times we are interested in the case of the **sparse graphs** in which the average degree c remains constant as the size of the graph increases. Then for the ER graph, from Eq.(2.5) we must have $p \to 0$ as $n \to \infty$. Henceforth, we will tacitly assume that the graph under consideration is sparse unless otherwise stated. We can also rewrite the average number of edges in terms of c using Eq(2.4) and Eq(2.5):

$$\langle m \rangle = \binom{n}{2} \frac{c}{n-1} = \frac{n(n-1)}{2} \times \frac{c}{n-1} = \frac{cn}{2} \tag{2.6}$$

2.2.1 Degree distribution

The degree distribution p_k of a network gives the probability that a randomly chosen vertex in the network has degree k. For the $G(n, p)$, the probability for a given vertex to be connected to some fixed k vertices, and of not being connected to the remaining $n - 1 - k$ vertices is $p^k(1 - p)^{n-1-k}$. However, this set of k vertices can be chosen in $\binom{n}{k}$ ways. Thus, the total probability that the vertex has degree k is:

$$p_k = \binom{n}{k} p^k (1-p)^{n-1-k} \tag{2.7}$$

This expression could be further simplified by writing p in terms of c and taking limit $n \to \infty$ as given below.

$$p_k = \lim_{n\to\infty} \binom{n}{k} \left(\frac{c}{n-1}\right)^k \left(1 - \frac{c}{n-1}\right)^{n-1-k} \tag{2.8}$$

Consider the first term,

$$\lim_{n\to\infty}\binom{n}{k} = \lim_{n\to\infty}\frac{n!}{k!(n-k)!} = \lim_{n\to\infty}\frac{(n-k)!(n-k+1)(n-k+2)\cdots n}{k!(n-k)!}$$
$$= \lim_{n\to\infty}\frac{(n-k+1)(n-k+2)\cdots n}{k!} = \lim_{n\to\infty}\frac{n^k}{k!} \tag{2.9}$$

Now consider the last term in Eq(2.8),

$$\lim_{n\to\infty}\left(1-\frac{c}{n-1}\right)^{n-1-k} = \lim_{n\to\infty}\left(1-\frac{c}{n-1}\right)^{n-1} = e^{-c} \tag{2.10}$$

Using Eqs (2.9) and (2.10) in (2.8), we get:

$$p_k = \lim_{n\to\infty}\frac{n^k}{k!}\frac{c^k}{(n-1)^k}e^{-c}$$

Hence,

$$p_k = e^{-c}\frac{c^k}{k!} \tag{2.11}$$

This is the standard Poisson distribution with mean c. In a nutshell, the Erdős-Rényi graph has a Poisson degree distribution, a distribution that is sharply peaked around the average c. Because of this, all the vertices in the ER graph has very similar degree values. As we will see, this makes the model completely unsatisfactory when it comes to using the model as a model of real networks. Now you also know why the model is also called the Poisson Random Graph!

2.3 Giant component in the ER graph

Consider the $G(n,p)$ for $p=0$ and for $p=1$. In the first case, there are no edges at all in the graph, all the vertices are isolated, and so the size of the largest component in the graph is 1. In the second case, all possible edges in the graph are present, and the size of the largest component is n. When n is large, the size of the largest component is much larger for $p=1$ than for $p=0$. However, apart from this quantitative difference, there also exists an important qualitative difference between the two cases: In the first case, the size of the largest component is independent of n, whereas in the second case the size varies as n. In the jargon of physics, we say that in the first case the size is *intensive*, whereas in the second case it is *extensive*. We can re-frame the whole thing in terms of the average degree c since $c=(n-1)p$. In the first case, $c=0$, the size of the largest component is 1, and in the second case $c=n-1$, and the size of the largest component is n.

The next obvious question is how the transition between these two regimes happens as c is increased? Does is so happen that as c is increased from 0, the largest component somehow starts becoming more and more extensive? In other words, does it happen that the size of the largest component varies as $n^{f(c)}$ for some function f with $f(0)=1$ and $f(n-1)=1$, and with f taking values between 0 and 1 for in between c values? Turns out that something much more interesting happens for the ER graph. Note that here we always assume that $n\to\infty$. For the ER graph, as we increase c from 0 to 1, the largest component in the graph always has a constant size independent of n. But as c crosses 1, suddenly this size becomes proportional to n so that as the graph size increases, the size of the largest component also increases proportionally. Such component whose size varies as n in the network when $n\to\infty$, is called the **giant component**.

2.3.1 Emergence of the giant component

Let u denote the fraction of vertices in the graph which **do not** belong to the giant component (This also means that u is the probability that a randomly selected vertex in the graph does not belong to the giant component). Then the fraction of vertices that belong to the giant component (GC) is $S=1-u$. Our aim to find out the value of S as a function of the average degree c. Consider a vertex i that does not belong to the GC. This means that it should not be connected to the GC via any other vertex j in the graph, and there are $n-1$ such vertices. For this to happen, only two possibilities exist:

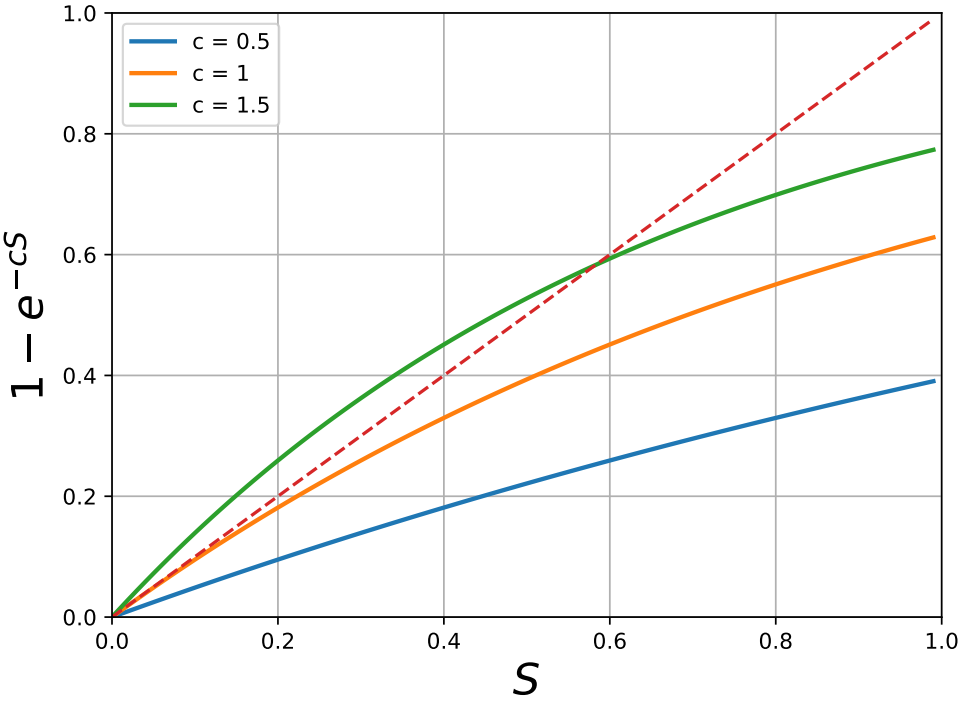

Figure 2.1

1. i is not connected to j. This happens with probability $1-p$. Or,
2. i is connected to j but j itself is not a part of the GC. This happens with probability pu

Hence, the total probability that vertex i does not belong the GC via j is $1-p+pu$. Then the probability that i is not connected to the GC via any other $n-1$ vertices is $(1-p+pu)^{n-1}$. However, we already know that this should be equal to u. Thus, we started to find out u, and then we got the expression in terms of u itself! This might sound like a chicken-egg problem, but a moment's reflection shows that this simply leads to a self consistent expression for u (and hence for S), and in the limit $n \to \infty$ we can simplify it as follows.

$$\begin{aligned} u &= \lim_{n\to\infty} (1-p+pu)^{n-1} \\ u &= \lim_{n\to\infty} \left(1 - \frac{c}{n-1} + \frac{c}{n-1}u\right)^{n-1} \\ u &= \lim_{n\to\infty} \left(1 - \frac{c}{n-1}(1-u)\right)^{n-1} = e^{-c(1-u)} \end{aligned} \tag{2.12}$$

Using $u = 1 - S$, this leads to:

$$1 - S = e^{-cS} \tag{2.13}$$

or rearranging,

$$S = 1 - e^{-cS} \tag{2.14}$$

This is the equation first derived by Erdős and Rényi for the size of the giant component in $G(n,p)$. Unfortunately, we cannot write S as an explicit function of c, and hence, for a given value of c, one must resort to numerical methods to find the corresponding value(s) of S. Another way to visualize the solutions of Eq(2.14) is to use a graphical method. Fig. 2.1 shows the plot of the R.H.S. of Eq(2.14) vs S for different

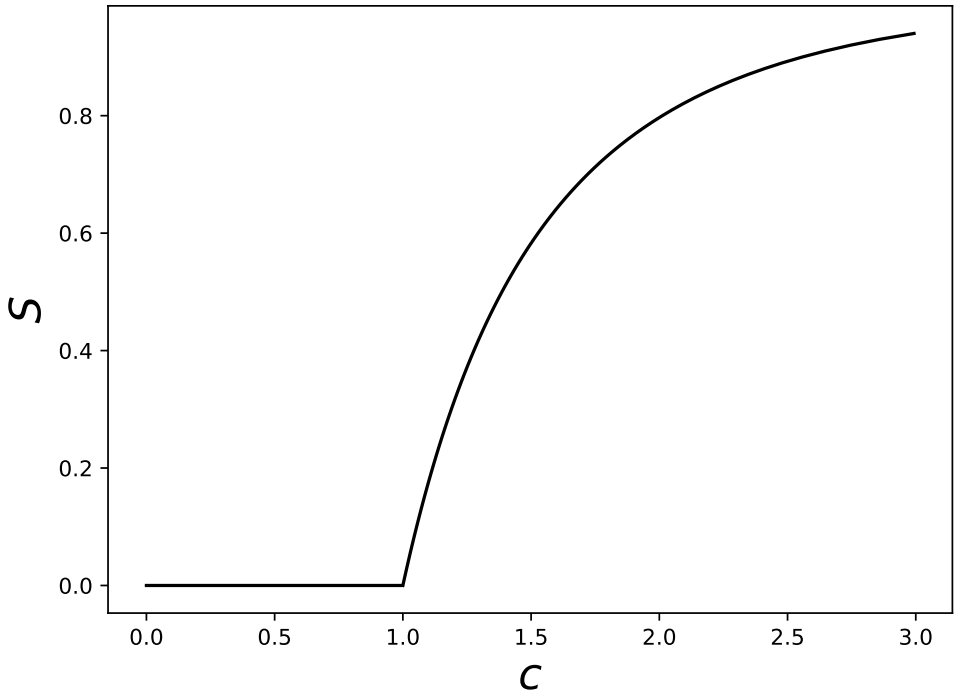

Figure 2.2

c. The dashed line $y = x$. The figure shows that for $c > 1$, this line intersects the curve $1 - e^{-cS}$ at two points, whereas for $c < 1$, there is only one intersection at $S = 0$. Hence, for $c < 1$, certainly the giant component won't exist, and for $c > 1$ the giant component **can** exist because the solution $S > 0$ also exists. The transition between these two regimes happens for $c = 1$, and the corresponding curve is a tangent to the line $S = S$ in the figure. This could also be seen by calculating the derivative of the curve at $S = 0$, and equating it to 1:

$$\begin{aligned} \left[\frac{d}{dS}(1 - e^{-cS})\right]_{S=0} &= 1 \\ [ce^{-cS}] = 1 \implies c &= 1 \end{aligned} \tag{2.15}$$

Probably the best numerical method to find the actual value of S for a given value of c, is to use iteration: start with some guess value S_0 of S, and plug it in the R.H.S. of Eq(2.14) to get a better approximation S_1, and then use this S_1 to get S_2 using the same equation, and so on until the convergence. For $c > 1$, this iteration converges only to the solution $S > 0$ when the initial guess $S_0 \in (0, 1)$, but since we know that $S = 0$ is always a solution, this is fine. Plotting these solutions as a function of c, we get the Fig. 2.2.

The figure shows the interesting fact that we mentioned earlier: at $c = 1$, suddenly the giant component appears, and after that with further increase in c increases the size of the GC. This phenomenon of a sudden qualitative change in a system's properties is known as a *phase-transition* in physics, and so we say that for ER graph, at $c = 1$, a phase-transition occurs that leads to appearance of a component whose size is *extensive*.

2.3.2 The core-periphery argument

Now we turn to a more subtle issue about the existence of a giant component. In the above analysis, we have shown that for $c > 1$, a giant component can exist in the Erdős-Rényi graph, but not that it **must** exist! Since in the regime $c > 1$ the Eq(2.14) has two solutions ($S = 0$ and $S > 0$), there is a possibility that the

giant component may or may not exist in the graph. Now we give an argument, known as the core-periphery argument, to show that this is not the case, and a giant component must exist for $c > 1$.

So consider the ER graph for $c > 1$. In the limit $n \to \infty$, there must exist a small connected set of vertices. Call this set the set $\mathcal{A}$. Let us divide this set $\mathcal{A}$ into two parts:

1. **Core:** The set of vertices in $\mathcal{A}$ which are connected only to the vertices in $\mathcal{A}$
2. **Periphery:** The set of vertices in $\mathcal{A}$ which are also connected to vertices outside $\mathcal{A}$

Now, let us enlarge our set by including in $\mathcal{A}$ all those vertices in the remaining graph which are connected to the vertices in the periphery of $\mathcal{A}$. Doing this gives a new periphery and a new core because the vertices that were part of the periphery are now part of the new core, and the vertices that just joined the set now form a new periphery. The question is, how big is this new periphery? If the original set $\mathcal{A}$ has size s, there are $n - s$ vertices outside the periphery. Any vertex in the old periphery is attached to each of the $n - s$ vertices outside with equal probability $p = c/n$. Hence the expected number of connections this one vertex has to the outside is:

$$\lim_{n\to\infty} (n - s) \times \frac{c}{n} = c \tag{2.16}$$

Thus, the size of the new periphery is c times the size of the old periphery. Now if $c > 1$, the size of the periphery increases by a factor of c (i.e. the size of the new set becomes $(1 + c)$ times its old size), and the set will eventually include a finite fraction of vertices in the network. Whereas if $c < 1$, each new periphery will have a size smaller than the previous periphery and so the set won't be able to include a finite fraction of vertices in the network. So in summary, when $c > 1$, we do see that the giant component **must** exist, and the solution $S = 0$ is never realized.

2.3.3 Uniqueness of the giant component

So far we have learned two things about the giant component in $G(n, p)$. First, it can exist only for $c > 1$, and we also saw how to calculate its size. Then, we showed that it must exist for $c > 1$ using the core-periphery argument. Now, we will show that when giant component exists, it is unique i.e. there can not exist more than one giant components.

Suppose, on the contrary that two GCs exist in the ER graph, and let S_1 and S_2 be the fractions of vertices in them. If n is the total size of the graph, this means that the numbers of vertices in the two GCs are $S_1 n$ and $S_2 n$ respectively. If we add even a single edge between these two components, they will get merged into a single component. Hence, as $n \to \infty$, there should not exist any edge between the two components of sizes $S_1 n$ and $S_2 n$. The total number of possible edges between these components is of course the product of their sizes, which is $S_1 S_2 n^2$. Since each edge is absent with probability $1 - c/(n - 1)$, the probability that there is no edge between the components is:

$$q = \left(1 - \frac{c}{n-1}\right)^{S_1 S_2 n^2} \tag{2.17}$$

Taking logarithm of both sides,

$$\ln q = S_1 S_2 n^2 \ln\left(1 - \frac{c}{n-1}\right) \tag{2.18}$$

Now consider the series expansion:

$$\ln(1 + x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots \tag{2.19}$$

when x is small, we can neglect all higher order terms in this expression and we get:

$$\ln(1 + x) \approx x \tag{2.20}$$

Thus, since $c/(n - 1)$ is very small, the Eq(2.18) could be approximated as:

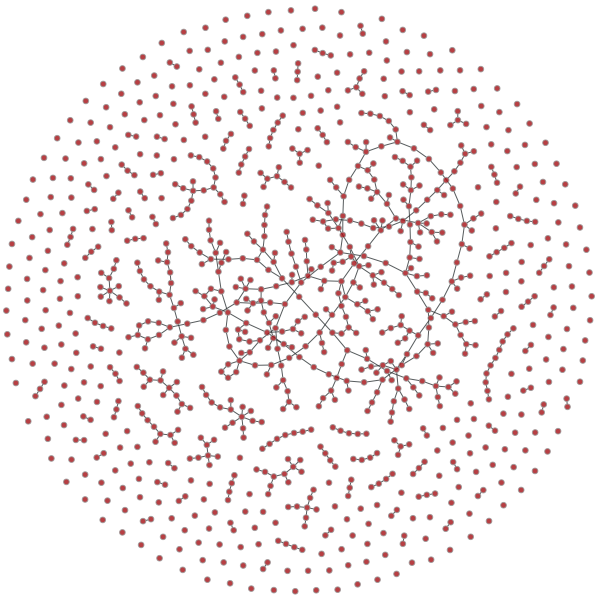

Figure 2.3

$$\ln q \approx S_1 S_2 n^2 \left(-\frac{c}{n-1}\right) \approx -S_1 S_2 cn$$
$$\therefore q \approx e^{-S_1 S_2 cn} \tag{2.21}$$

In the limit $n \to \infty$, this probability goes to 0, and hence the two components cannot remain separate. This shows that when the giant component exists, it is unique.

2.4 Small components in the ER graph

We have seen that the GC exists for the Erdős-Rényi graph for $c > 1$. However, the fraction of vertices S in the giant component is always less than 1 for any finite value of c (See Fig. 2.2). Hence, there are vertices in the graph which are not part of the GC. We might ask what is the structure of the remaining network then? The answer is that the rest of the network is made up of *small components*. Here the term *small* should not be taken literally, and by the phrase *small* we mean that the sizes of these components remain constant in the limit $n \to \infty$ unlike the GC whose size scales as n. Here we will only calculate the average size of a small component to which a randomly chosen vertex belongs for any given value of c in the limit $n \to \infty$.

We can calculate this average by noticing an interesting fact about the small components: **small components are trees!**. To see this, consider a small component of size s. To hold these s vertices together, we need at least $s-1$ edges out of total possible $\binom{s}{2}$ edges. Thus, the possible number of extra edges in this component is: $\binom{s}{2} - (s-1)$. Each of these edges is present with equal probability $p = \frac{c}{n-1}$, and hence the average number of extra edges in this component is:

$$\text{Average number of extra edges} = \lim_{n\to\infty} \left(\binom{s}{2} - (s-1)\right) \times \frac{c}{n-1} = 0 \tag{2.22}$$

Hence, the average number of extra edges in this component goes to 0, and the component is a tree. Notice that this argument doesn't hold for the giant component since the size s is not finite in that case. Fig. 2.3 shows a realization of $G(n,p)$ for $n = 1000$ and $c = 1.2$, and as you can see, all the small components are indeed trees.

Now consider a vertex i in a small component of size s, let k be the degree of this vertex. Since the component is tree, the vertex i is connected to subgraphs of sizes $t_1, t_2, \cdots t_k$. Thus, the size of the component to which this vertex belongs can be written as :

$$s = 1 + \sum_{j=1}^{k} t_j \tag{2.23}$$

First, let us take average of this over the vertices with degree k:

$$\langle s \rangle_k = 1 + \sum_{j=1}^{k} \langle t_j \rangle \tag{2.24}$$

Where t_j is the average size of the j^{th} subgraph connected to vertex i. However, the subgraphs connected to vertex i could have been labeled using a different order, but that should not change the average size of the j^{th} component. This means that all the average sizes $\langle t_j \rangle$ must be all same equal to $\langle t \rangle$. Hence,

$$\langle s \rangle_k = 1 + k\langle t \rangle \tag{2.25}$$

Now we need to average this quantity over the degree distribution p_k^{small} of the vertices in the small components. Notice that the degree distribution of the vertices in the small components cannot be same as the degree distribution of the whole network. The primary reason is that the vertices with higher degrees have higher chance of being connected to the giant component, and so the vertices in the small components tend to have smaller degrees. Thus, we have:

$$\langle s \rangle = \sum_{k=0}^{\infty} s_k p_k^{\text{small}} = \sum_{k=0}^{\infty} (1 + k\langle t \rangle) p_k^{\text{small}} = \sum_{k=0}^{\infty} p_k^{\text{small}} + \langle t \rangle \sum_{k=0}^{\infty} k p_k^{\text{small}} = 1 + \langle t \rangle \langle k \rangle_{\text{small}} \tag{2.26}$$

where $\langle k \rangle_{\text{small}}$ is the average degree of a vertex in a small component, and we have used the fact that p_k^{small} is normalized to unity. There are in total $(1-S)n$ vertices in small components. Now consider a vertex v in a small component. This vertex is connected to each of the remaining $(1-S)n - 1$ vertices with the same probability $p = \frac{c}{n-1}$. Hence, its average degree is:

$$\langle k \rangle_{\text{small}} = \lim_{n \to \infty} ((1-S)n - 1) \times \frac{c}{n-1} = (1-S)c \tag{2.27}$$

That is, the average degree of a vertex in a small component is $(1-S)$ times the average degree of the network, a reasonable fact. Putting this value in Eq(2.26), we get:

$$\langle s \rangle = 1 + \langle t \rangle (1-S)c \tag{2.28}$$

Now we only want to find out the value of $\langle t \rangle$. To do this, we use a beautiful trick from statistical physics known as the *cavity method*. Again consider the same component with size s and the vertex i with degree k in it. Recall that the quantity $\langle t \rangle$ is the average size of a subgraph connected to this vertex i. Now suppose, we remove this vertex i and the edges connected to it from the graph. Since the small component is a tree, the removal of i will break the component into k different components. However, in the limit of $n \to \infty$, all statistical properties of the graph would be preserved, and hence the average size of any of these new components is equal to the average size of any small component. In other words, $\langle s \rangle = \langle t \rangle$, a remarkable result! Using this in Eq(2.28), we finally get:

$$\langle s \rangle = 1 + \langle s \rangle (1-S)c \tag{2.29}$$

Rearranging, we get:

$$\langle s \rangle = \frac{1}{1 - c + cS} \tag{2.30}$$

Remember that this is not the average size of a small component; it is the average size of the component to which a randomly chosen vertex belongs. Larger components have higher probability

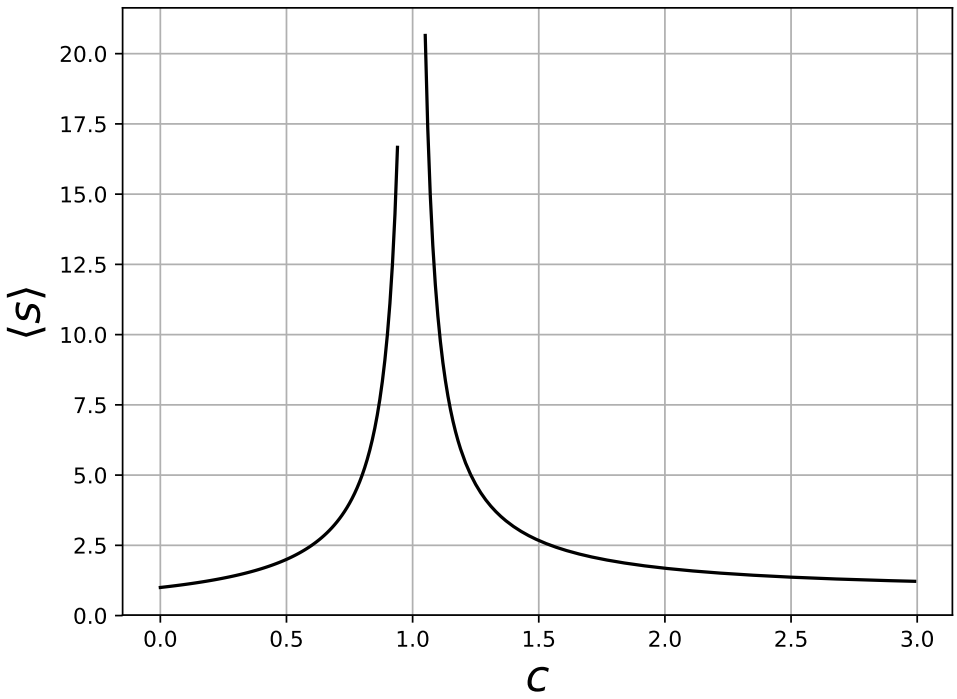

Figure 2.4

for a vertex belonging to them, and so the $\langle s \rangle$ is really a biased average, biased towards larger components. However, this turns out to be a more useful quantity than the average size of a small component.

Fig. 2.4 shows the variation of the average size of a small component with c as given by Eq(2.30). When $c < 1$, we know that there is no giant component, and hence $S = 0$ in this regime. Then according to Eq(2.30), we get $\langle s \rangle = 1/(1 - c)$. When $c > 1$, we have to first calculate the corresponding value of S by solving Eq(2.14), and then substituting values of c and S in (2.30). Observe the divergence at $c = 1$, the point at which the giant component appears. This divergence could be understood as follows: when $c < 1$, there is no giant component, and only small components exist. As we start increasing c, the sizes of these components start increasing and so does their average size. At $c = 1$, these small components merge together to form the giant component, and so their average size diverges. For $c > 1$, giant component is present, and so the expression for $\langle s \rangle$ takes into account only the small components. Then in this regime, with increase in c more and more vertices join the giant component, and so the average size of a small component again decreases.

2.5 Path lengths in the ER graph

We have already discussed that for almost all real-world networks, the average vertex to vertex distance is very small compared with the size of the graph. We would like to see whether the ER graph offers some explanation of this. Turns out that, in spite of being completely random, the average path length of the ER graph varies roughly as $\ln n$, and since ln is an extremely slowly varying function of its argument, the actual increase is remarkably slow. Here, we instead of the average path length, concentrate on the diameter of the network. Recall that the diameter is the longest distance in the network, or the 'longest shortest path'. As we will see now, the diameter of the ER graph also varies roughly as $\ln n$, and so this offers some explanation of the *small-world effect*.

Let us first do an approximate calculation for the diameter of the ER graph. Consider a particular vertex

in the graph. Now imagine that starting from this vertex, we start enlarging a set containing this vertex by going outwards along the edges. Since this vertex is connected on average to c other vertices, and each of those neighbours is also connected to c neighbours on average, we see that in two steps we reach c^2 vertices on average. Continuing this argument, in s steps, we should reach around c^s vertices. Of course this cannot go on indefinitely, and the number of vertices reached should become equal to n for some value of s. This value of s is the maximum distance from the vertex that we started with, and hence should be comparable to the diameter D of the network. In other words,

$$c^D \approx n \implies D \approx= \frac{\ln n}{\ln c} \tag{2.31}$$

Thus we see that the diameter indeed varies as $\ln n$. However, this calculation is too crude. First, we are really calculating the *radius* of the network, not the diameter. More importantly, saying that the number of new vertices reach at step s is c^s is a good approximation only when this number is sufficiently smaller than n. When the number is large, we will probably overcount since many vertices will be counted several times as they could be reached via different paths from the starting vertex. To overcome this, let us modify our approximate argument. Instead of considering a single vertex as the source, let us consider two vertices i and j as sources, and now imagine going outwards from these two sources for s and t steps respectively. In these many steps, we will reach around c^s and c^t vertices respectively provided that these numbers are small compared to n.

So after s and t steps, we would really have two "balls" of vertices around i and j with radii s and t respectively. What if there is at least one edge between the "surfaces" of these two balls? A vertex on the surface of the first ball is at a distance s from i, and a vertex on the surface of the second ball is at distance t from j. Hence, if there is an edge between any vertex on the surface of the first ball and any vertex on the surface of the second ball, the distance d_{ij} between i and j would be $s+t+1$. In other words we can write the following probability equation,

$$\mathbb{P}(d_{ij} > s+t+1) = \mathbb{P}(\text{There is no edge between the surfaces}) \tag{2.32}$$

Now what is the value of the expression on the R.H.S.? Since the number of vertices at distance s from a chosen vertex is around c^s, the numbers of vertices on the surfaces of the two balls are c^s and c^t respectively. Hence, the total number of possible edges between the surfaces is $c^s \times c^t = c^{s+t}$. Each of these edges is present with probability c/n, and hence the probability that all of these are absent is:

$$\mathbb{P}(d_{ij} > s+t+1) = \left(1 - \frac{c}{n}\right)^{c^{s+t}} \tag{2.33}$$

Using a shorthand $s+t+1 = l$, this can be rewritten as:

$$\mathbb{P}(d_{ij} > l) = \left(1 - \frac{c}{n}\right)^{c^{l-1}} \tag{2.34}$$

To simplify this using the fact that $\frac{c}{n}$ is small, we use our usual trick of taking logarithms of both the sides. We have:

$$\ln \mathbb{P}(d_{ij} > l) = c^{l-1} \ln\left(1 - \frac{c}{n}\right) \approx c^{l-1}\left(-\frac{c}{n}\right) = -\frac{c^l}{n} \tag{2.35}$$

Hence,

$$\mathbb{P}(d_{ij} > l) = \exp\left(-\frac{c^l}{n}\right) \tag{2.36}$$

The diameter D of the network is that value of l for which no distance is greater than it. In other words, $\mathbb{P}(d_{ij} > D) = 0$. Thus, we must have:

$$\lim_{n\to\infty} \exp\left(-\frac{c^D}{n}\right) = 0 \tag{2.37}$$

This is possible only if c^D increases faster than n, say as $an^{1+\epsilon}$ with a constant, and $\epsilon \to 0$ from above. Notice that we can keep the numbers c^s and c^t small compared with n in this calculation as was required.

For example, since $c^D = c^{s+t+1} = an^{1+\epsilon}$, we have $c^s c^t = \frac{a}{c} n^{1+\epsilon}$, and so we can make both c^s and c^t vary as $n^{\frac{1+\epsilon}{2}}$ or roughly as $\sqrt{n}$. Then, taking logarithm of both sides, we have:

$$D \ln c = \ln a + \ln n \implies D = \frac{\ln a}{\ln c} + \frac{\ln n}{\ln c} \implies D = A + \frac{\ln n}{\ln c} \tag{2.38}$$

where A is a constant i.e. its value does not depend on n. Thus, after this more accurate calculation, we realize that apart from the additive constant A, we get the same expression that we got by using the crude calculation, and indeed the diameter of the ER graph varies as $\ln n$.

2.6 Comparison with the real-world networks

We now turn to the issue of comparing the properties of ER graph with the empirical networks. We have already seen that the model provides a reasonably satisfying explanation of the component structures and the path lengths observed in real world graphs. However, there are number of features observed in the real-world networks which are at all not reproduced by the model, and so it fails miserably at these fronts.

Let us start with the fact that in most real world networks, we see a large number of triangles. For example, in the context of social networks, it is highly likely that two of your friends are also friends with each other. The average probability of two neighbours being neighbours with each other is just the clustering coefficient of the graph. Since in the ER graph each edge is equally likely to be present, the clustering coefficient is just $p = c/(n-1)$. If we assume the ER model to be a good model for the social network of all humans on earth, and if we assume that each one of us has around 1000 acquintances, and using the fact that there are around 10^9 humans on earth, then according to this model, the clustering coefficient of our graph would be:

$$C = \frac{1000}{10^9} \approx 10^{-6} \tag{2.39}$$

This value is by orders of magnitude smaller than the actual value; though we don't know the actual value for sure, we do know that it can't be this small, and hence the ER model completely fails to explain the observed clustering in the real-world networks.

Also, we have seen before that in the real-world networks, degree values of the connected vertices are usually correlated (or anti-correlated). We have seen that this tendency can be measured using the *degree assortativity coefficient*. But in the ER model, edges are placed completely independently of each other and with equal probability. Because of this, the degree values of the connected vertices in the ER graph have no correlation at all, and so this is another property that ER model fails to explain.

One can point out many other properties of the real-world graphs which are not present in the ER graph, but there is one property that deserves a longer discussion: the *degree distribution*. As we already know, the degree distribution p_k is the fraction of vertices in the graph with degree k, and it turns out to be the most important property in terms of its effect on the processes on networks. We have seen before that the degree distribution in many real networks has a heavy tail, meaning that many vertices in these graphs have a fairly small degree, but there exist a small number of vertices with extremely high degrees. Recall that these networks are called *scale-free networks* (also recall the associated controversy!). Since the degrees of the vertices in the ER graph are Poisson distributed, and Poisson distribution being a fairly peaked distribution around its average value, all the vertices in the ER graph have values close to each other. Thus, ER graph fails to explain the observed scale-free nature of many real networks, and this turns out to be huge shortcoming of the model.

Fortunately, it is possible to construct a random graph model in which the degree distribution can have an arbitrary shape, and in the next chapter, we will start the study of one such model.