INDEX

Figure 3.0 (cover image)

Erdős Number

The Hungarian mathematician Pál Erdős authored hundreds of research papers, many of them in collaboration with other mathematicians. His relentless collaborative approach to mathematics inspired the *Erdős Number*, which works like this: Erdős' Erdős number is 0. Erdős' coauthors have Erdős number 1. Those who have written a paper with someone with Erdős number 1 have Erdős number 2, and so on. If there is no chain of coauthorships connecting someone to Erdős, then that person's Erdős number is infinite. Many famous scientists have low Erdős numbers: Albert Einstein has Erdős Number 2 and Richard Feynman has 3. The image shows the collaborators of Pál Erdős, as drawn in 1970 by Ronald Graham, one of Erdős' close collaborators. As Erdős' fame rose, this image has achieved an iconic status.

PDF V32, 08.09.2014

INTRODUCTION

Imagine organizing a party for a hundred guests who initially do not know each other [1]. Offer them wine and cheese and you will soon see them chatting in groups of two to three. Now mention to Mary, one of your guests, that the red wine in the unlabeled dark green bottles is a rare vintage, much better than the one with the fancy red label. If she shares this information only with her acquaintances, your expensive wine appears to be safe, as she only had time to meet a few others so far.

The guests will continue to mingle, however, creating subtle paths between individuals that may still be strangers to each other. For example, while John has not yet met Mary, they have both met Mike, so there is an invisible path from John to Mary through Mike. As time goes on, the guests will be increasingly interwoven by such elusive links. With that the secret of the unlabeled bottle will pass from Mary to Mike and from Mike to John, escaping into a rapidly expanding group.

To be sure, when all guests had gotten to know each other, everyone would be pouring the superior wine. But if each encounter took only ten minutes, meeting all ninety-nine others would take about sixteen hours. Thus, you could reasonably hope that a few drops of your fine wine would be left for you to enjoy once the guests are gone.

Yet, you would be wrong. In this chapter we show you why. We will see that the party maps into a classic model in network science called the random network model. And random network theory tells us that we do not have to wait until *all* individuals get to know each other for our expensive wine to be in danger. Rather, soon after each person meets at least one other guest, an invisible network will emerge that will allow the information to reach all of them. Hence in no time everyone will be enjoying the better wine.

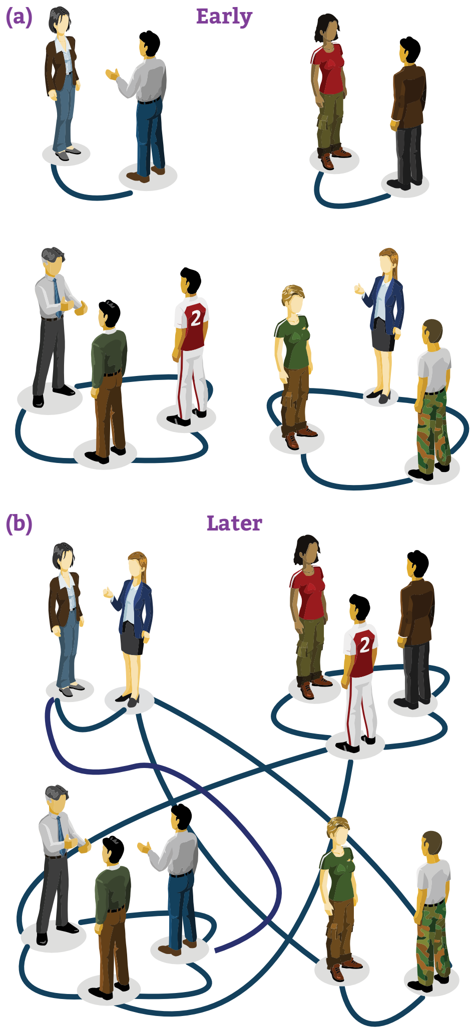


Figure 3.1
From a Cocktail Party to Random Networks

The emergence of an acquaintance network through random encounters at a cocktail party.

(a) Early on the guests form isolated groups.

(b) As individuals mingle, changing groups, an invisible network emerges that connects all of them into a single network.

THE RANDOM NETWORK MODEL

Network science aims to build models that reproduce the properties of real networks. Most networks we encounter do not have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web. Rather, at first inspection they look as if they were spun randomly (Figure 2.4). Random network theory embraces this apparent randomness by constructing and characterizing networks that are *truly random*.

From a modeling perspective a network is a relatively simple object, consisting of only nodes and links. The real challenge, however, is to decide where to place the links between the nodes so that we reproduce the complexity of a real system. In this respect the philosophy behind a random network is simple: We assume that this goal is best achieved by placing the links randomly between the nodes. That takes us to the definition of a random network (BOX 3.1):

A random network consists of N nodes where each node pair is connected with probability p.

To construct a random network we follow these steps:

1) Start with N isolated nodes.

2) Select a node pair and generate a random number between 0 and 1. If the number exceeds p, connect the selected node pair with a link, otherwise leave them disconnected.

3) Repeat step (2) for each of the $N(N$-1)/2 node pairs.

The network obtained after this procedure is called a *random graph* or a *random network*. Two mathematicians, Pál Erdős and Alfréd Rényi, have played an important role in understanding the properties of these networks. In their honor a random network is called the *Erdős-Rényi network* (BOX 3.2).

BOX 3.1

DEFINING RANDOM NETWORKS

There are two definitions of a random network:

***G*(*N, L*) Model**
N labeled nodes are connected with L randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9].

***G*(*N, p*) Model**
Each pair of N labeled nodes is connected with probability p, a model introduced by Gilbert [10].

Hence, the *G(N, p)* model fixes the probability p that two nodes are connected and the *G(N, L)* model fixes the total number of links L. While in the *G(N, L)* model the average degree of a node is simply $<k> = 2L/N$, other network characteristics are easier to calculate in the *G(N, p)* model. Throughout this book we will explore the *G*(*N, p*) model, not only for the ease that it allows us to calculate key network characteristics, but also because in real networks the number of links rarely stays fixed.

BOX 3.2

RANDOM NETWORKS: A BRIEF HISTORY

Anatol Rapoport (1911-2007), a Russian immigrant to the United States, was the first to study random networks. Rapoport's interests turned to mathematics after realizing that a successful career as a concert pianist would require a wealthy patron. He focused on mathematical biology at a time when mathematicians and biologists hardly spoke to each other. In a paper written with Ray Solomonoff in 1951 [11], Rapoport demonstrated that if we increase the average degree of a network, we observe an abrupt transition from disconnected nodes to a graph with a giant component.

The study of random networks reached prominence thanks to the fundamental work of Pál Erdős and Alfréd Rényi (Figure 3.2). In a sequence of eight papers published between 1959 and 1968 [2-9], they merged probability theory and combinatorics with graph theory, establishing *random graph theory*, a new branch of mathematics [2].

The random network model was independently introduced by Edgar Nelson Gilbert (1923-2013) [10] the same year Erdős and Rényi published their first paper on the subject. Yet, the impact of Erdős and Rényi's work is so overwhelming that they are rightly considered the founders of random graph theory.

> *"A mathematician is a device for turning coffee into theorems"*
>
> *Alfréd Rényi (a quote often attributed to Erdős)*

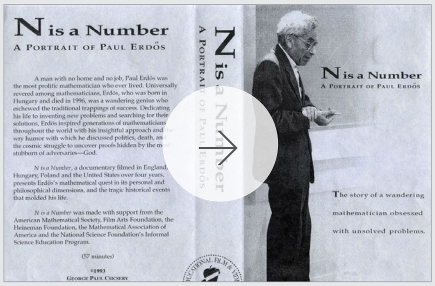
Figure 3.2

(a) Pál Erdős (1913-1996)

Hungarian mathematician known for both his exceptional scientific output and eccentricity. Indeed, Erdős published more papers than any other mathematician in the history of mathematics. He co-authored papers with over five hundred mathematicians, inspiring the concept of *Erdős number*. His legendary personality and profound professional impact has inspired two biographies [12, 13] and a documentary [14] (Online Resource 3.1).

(b) Alfréd Rényi (1921-1970)

Hungarian mathematician with fundamental contributions to combinatorics, graph theory, and number theory. His impact goes beyond mathematics: The Rényi entropy is widely used in chaos theory and the random network theory he co-developed is at the heart of network science. He is remembered through the hotbed of Hungarian mathematics, the Alfréd Rényi Institute of Mathematics in Budapest.



N is a Number

A Portrait of Paul Erdős

A man with no home and no job, Paul Erdös was the most prolific mathematician who ever lived. Universally revered among mathematicians, Erdös, who was born in Hungary and died in 1996, was a wandering genius who eschewed the traditional trappings of success. Dedicating his life to inventing new problems and searching for their solutions, Erdös inspired generations of mathematicians throughout the world with his insightful approach and the wry humor with which he discussed politics, death, and the cosmic struggle to uncover proofs hidden by the most stubborn of adversaries—God.

N is a Number, a documentary filmed in England, Hungary, Poland and the United States over four years, presents Erdős's mathematical quest in its personal and philosophical dimensions, and the tragic historical events that molded his life.

N is a Number was made with support from the American Mathematical Society, Film Arts Foundation, the Heineman Foundation, the Mathematical Association of America and the National Science Foundation's Informal Science Education Program.

(57 minutes)

©1993
George Paul Csicsery

N is a Number
A Portrait of Paul Erdős

N is a Number
A Portrait of Paul Erdős

The story of a wandering mathematician obsessed with unsolved problems.

Online Resource 3.1

N is a Number: A Portrait of Paul Erdős

The 1993 biographical documentary of Pál Erdős, directed by George Paul Csicsery, offers a glimpse into Erdős' life and scientific impact [14].

SECTION 3.3

NUMBER OF LINKS

Each random network generated with the same parameters *N*, *p* looks slightly different (Figure 3.3). Not only the detailed wiring diagram changes between realizations, but so does the number of links *L*. It is useful, therefore, to determine how many links we expect for a particular realization of a random network with fixed *N* and *p*.

The probability that a random network has exactly *L* links is the product of three terms:

1) The probability that *L* of the attempts to connect the *N*(*N*-1)/2 pairs of nodes have resulted in a link, which is p^L.

2) The probability that the remaining *N*(*N*-1)/2 - *L* attempts have not resulted in a link, which is $(1\text{-}p)^{N(N\text{-}1)/2\text{-}L}$.

3) A combinational factor,

$$\binom{\frac{N(N\text{-}1)}{2}}{L}, \tag{3.0}$$

counting the number of different ways we can place *L* links among *N*(*N*-1)/2 node pairs.

We can therefore write the probability that a particular realization of a random network has exactly *L* links as

$$p_L = \binom{\frac{N(N\text{-}1)}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}. \tag{3.1}$$

As (3.1) is a binomial distribution (BOX 3.3), the expected number of links in a random graph is

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2}. \tag{3.2}$$

Hence <L> is the product of the probability p that two nodes are connected and the number of pairs we attempt to connect, which is $L_{max} = N(N - 1)/2$ (CHAPTER 2).

Using (3.2) we obtain the average degree of a random network

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1). \tag{3.3}$$

Hence <k> is the product of the probability p that two nodes are connected and (N-1), which is the maximum number of links a node can have in a network of size N.

In summary the number of links in a random network varies between realizations. Its expected value is determined by N and p. If we increase p a random network becomes denser: The average number of links increase linearly from <L> = 0 to $L_{\rm max}$ and the average degree of a node increases from <k> = 0 to <k> = N-1.

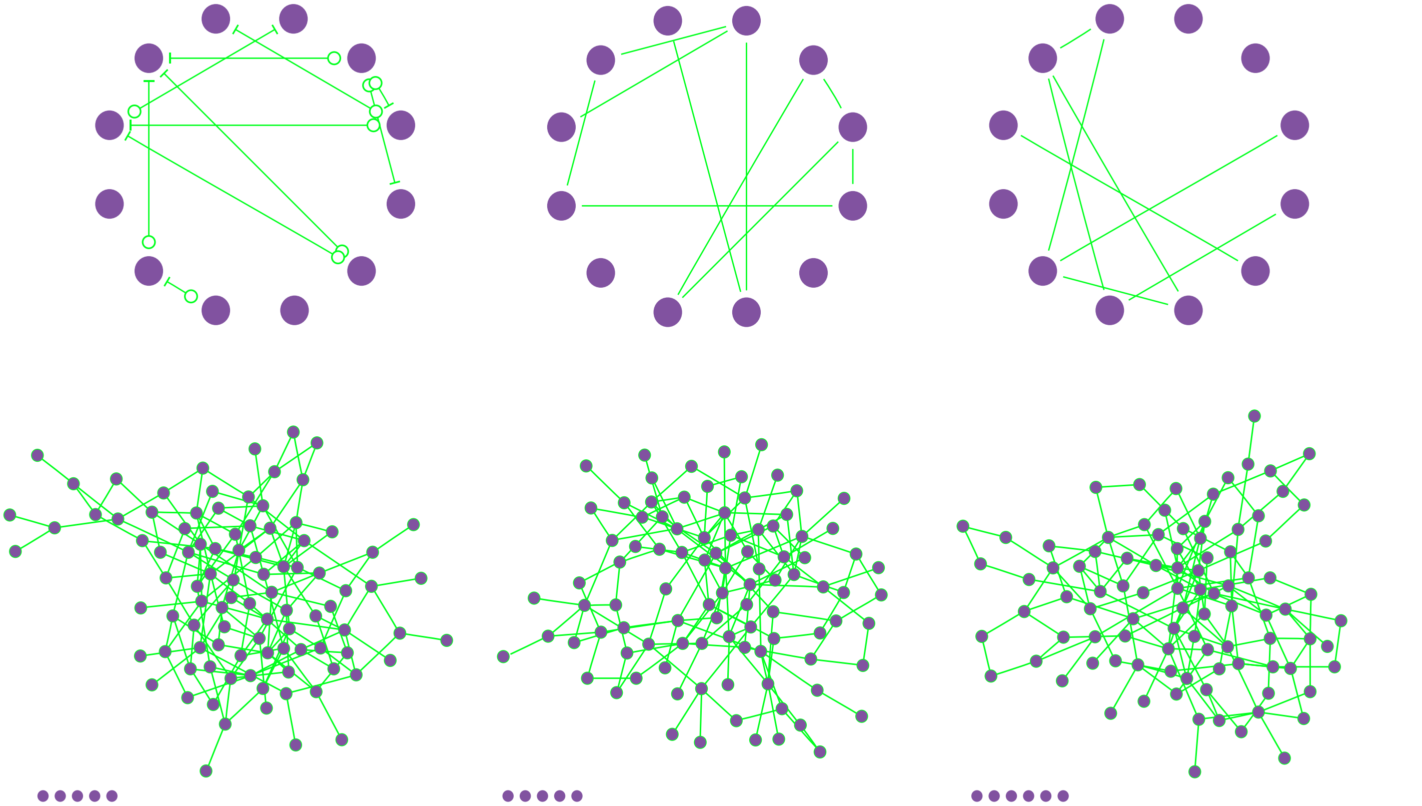


Figure 3.3

Random Networks are Truly Random

Top Row

Three realizations of a random network generated with the same parameters p=1/6 and N=12. Despite the identical parameters, the networks not only look different, but they have a different number of links as well (L=10, 10, 8).

Bottom Row

Three realizations of a random network with p=0.03 and N=100. Several nodes have degree k=0, shown as isolated nodes at the bottom.

BOX 3.3

BINOMIAL DISTRIBUTION: MEAN AND VARIANCE

If we toss a fair coin N times, tails and heads occur with the same probability $p = 1/2$. The binomial distribution provides the probability p_x that we obtain exactly x heads in a sequence of N throws. In general, the binomial distribution describes the number of successes in N independent experiments with two possible outcomes, in which the probability of one outcome is p, and of the other is 1-p.

The binomial distribution has the form

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}.$$

The mean of the distribution (first moment) is

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np. \tag{3.4}$$

Its second moment is

$$\langle x^2 \rangle = \sum_{x=0}^{N} x^2 p_x = p(1-p)N + p^2N^2, \tag{3.5}$$

providing its standard deviation as

$$\sigma_x = \left(\langle x^2 \rangle - \langle x \rangle^2\right)^{\frac{1}{2}} = [p(1-p)N]^{\frac{1}{2}}. \tag{3.6}$$

Equations (3.4) - (3.6) are used repeatedly as we characterize random networks.

DEGREE DISTRIBUTION

In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links (Figure 3.3). These differences are captured by the degree distribution, p_k, which is the probability that a randomly chosen node has degree k. In this section we derive p_k for a random network and discuss its properties.

BINOMIAL DISTRIBUTION

In a random network the probability that node i has exactly k links is the product of three terms [15]:

- The probability that k of its links are present, or p^k.

- The probability that the remaining (N-1-k) links are missing, or (1-p)$^{N\text{-}1\text{-}k}$.

- The number of ways we can select k links from N- 1 potential links a node can have, or

$$\binom{N-1}{k}.$$

Consequently the degree distribution of a random network follows the binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \quad (3.7)$$

The shape of this distribution depends on the system size N and the probability p (Figure 3.4). The binomial distribution (BOX 3.3) allows us to calculate the network's average degree <k>, recovering (3.3), as well as its second moment <k^2> and variance σ_k (Figure 3.4).

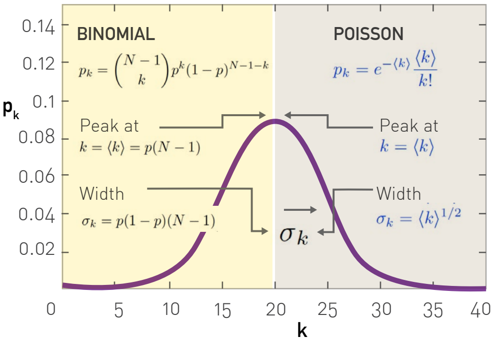


Figure 3.4

Binomial vs. Poisson Degree Distribution

The exact form of the degree distribution of a random network is the binomial distribution (left half). For $N \gg$ <k> the binomial is well approximated by a Poisson distribution (right half). As both formulas describe the same distribution, they have the identical properties, but they are expressed in terms of different parameters: The binomial distribution depends on p and N, while the Poisson distribution has only one parameter, <k>. It is this simplicity that makes the Poisson form preferred in calculations.

POISSON DISTRIBUTION

Most real networks are sparse, meaning that for them $<k> \ll N$ (Table 2.1). In this limit the degree distribution (3.7) is well approximated by the Poisson distribution (ADVANCED TOPICS 3.A)

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}, \tag{3.8}$$

which is often called, together with (3.7), the *degree distribution of a random network.*

The binomial and the Poisson distribution describe the same quantity, hence they have similar properties (Figure 3.4):

- Both distributions have a peak around $<k>$. If we increase p the network becomes denser, increasing $<k>$ and moving the peak to the right.

- The width of the distribution (dispersion) is also controlled by p or $<k>$. The denser the network, the wider is the distribution, hence the larger are the differences in the degrees.

When we use the Poisson form (3.8), we need to keep in mind that:

- The exact result for the degree distribution is the binomial form (3.7), thus (3.8) represents only an approximation to (3.7) valid in the $<k> \ll N$ limit. As most networks of practical importance are sparse, this condition is typically satisfied.

- The advantage of the Poisson form is that key network characteristics, like $<k>$, $<k^2>$ and σ_k , have a much simpler form (Figure 3.4), depending on a single parameter, $<k>$.

- The Poisson distribution in (3.8) does not explicitly depend on the number of nodes N. Therefore, (3.8) predicts that the degree distribution of networks of different sizes but the same average degree $<k>$ are indistinguishable from each other (Figure 3.5).

In summary, while the Poisson distribution is only an approximation to the degree distribution of a random network, thanks to its analytical simplicity, it is the preferred form for p_k. Hence throughout this book, unless noted otherwise, we will refer to the Poisson form (3.8) as the degree distribution of a random network. Its key feature is that its properties are independent of the network size and depend on a single parameter, the average degree $<k>$.

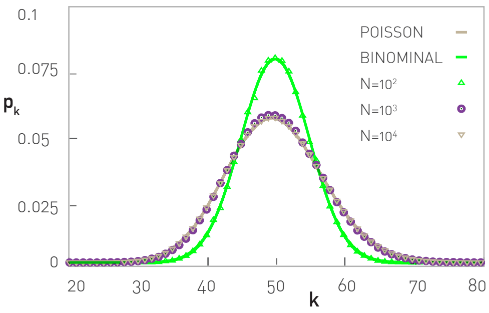


Figure 3.5

Degree Distribution is Independent of the Network Size

The degree distribution of a random network with $<k> = 50$ and $N = 10^2, 10^3, 10^4$.

Small Networks: Binomial

For a small network ($N = 10^2$) the degree distribution deviates significantly from the Poisson form (3.8), as the condition for the Poisson approximation, $N \gg <k>$, is not satisfied. Hence for small networks one needs to use the exact binomial form (3.7) (green line).

Large Networks: Poisson

For larger networks ($N = 10^3, 10^4$) the degree distribution becomes indistinguishable from the Poisson prediction (3.8), shown as a continuous grey line. Therefore for large N the degree distribution is independent of the network size. In the figure we averaged over 1,000 independently generated random networks to decrease the noise.

SECTION 3.5

REAL NETWORKS ARE NOT POISSON

As the degree of a node in a random network can vary between 0 and *N*-1, we must ask, how big are the differences between the node degrees in a particular realization of a random network? That is, can high degree nodes coexist with small degree nodes? We address these questions by estimating the size of the largest and the smallest node in a random network.

Let us assume that the world's social network is described by the random network model. This random society may not be as far fetched as it first sounds: There is significant randomness in whom we meet and whom we choose to become acquainted with.

Sociologists estimate that a typical person knows about 1,000 individuals on a first name basis, prompting us to assume that $<k> \approx 1{,}000$. Using the results obtained so far about random networks, we arrive to a number of intriguing conclusions about a random society of $N \simeq 7 \times 10^9$ of individuals (ADVANCED TOPICS 3.B):

- The most connected individual (the largest degree node) in a random society is expected to have $k_{max} = 1{,}185$ acquaintances.

- The degree of the least connected individual is $k_{min} = 816$, not that different from k_{max} or $<k>$.

- The dispersion of a random network is $\sigma_k = <k>^{1/2}$, which for $<k> = 1{,}000$ is $\sigma_k = 31.62$. This means that the number of friends a typical individual has is in the $<k> \pm \sigma_k$ range, or between 968 and 1,032, a rather narrow window.

Taken together, in a random society all individuals are expected to have a comparable number of friends. Hence if people are randomly connected to each other, we lack outliers: There are no highly popular individuals, and no one is left behind, having only a few friends. This suprising conclusion is a consequence of an important property of random networks: *in a large random network the degree of most nodes is in the narrow vicinity of* $<k>$

(BOX 3.4).

This prediction blatantly conflicts with reality. Indeed, there is extensive evidence of individuals who have considerably more than 1,185 acquaintances. For example, US president Franklin Delano Roosevelt's appointment book has about 22,000 names, individuals he met personally [16, 17]. Similarly, a study of the social network behind Facebook has documented numerous individuals with 5,000 Facebook friends, the maximum allowed by the social networking platform [18]. To understand the origin of these discrepancies we must compare the degree distribution of real and random networks.

In Figure 3.6 we show the degree distribution of three real networks, together with the corresponding Poisson fit. The figure documents systematic differences between the random network predictions and the real data:

- The Poisson form significantly underestimates the number of high degree nodes. For example, according to the random network model the maximum degree of the Internet is expected to be around 20. In contrast the data indicates the existence of routers with degrees close to 10^3.

- The spread in the degrees of real networks is much wider than expected in a random network. This difference is captured by the dispersion σ_k (Figure 3.4). If the Internet were to be random, we would expect σ_k = 2.52. The measurements indicate $\sigma_{internet}$ = 14.14, significantly higher than the random prediction. These differences are not limited to the networks shown in Figure 3.6, but all networks listed in Table 2.1 share this property.

In summary, the comparison with the real data indicates that the random network model does not capture the degree distribution of real networks. In a random network most nodes have comparable degrees, forbidding hubs. In contrast, in real networks we observe a significant number of highly connected nodes and there are large differences in node degrees. We will resolve these differences in CHAPTER 4.

BOX 3.4

WHY ARE HUBS MISSING?

To understand why hubs, nodes with a very large degree, are absent in random networks, we turn to the degree distribution (3.8).

We first note that the $1/k!$ term in (3.8) significantly decreases the chances of observing large degree nodes. Indeed, the Stirling approximation

$$k! \sim \left[\sqrt{2\pi k}\right]\left(\frac{k}{e}\right)^k$$

allows us rewrite (3.8) as

$$p_k = \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}}\left(\frac{e\langle k \rangle}{k}\right)^k. \qquad (3.9)$$

For degrees $k > e<k>$ the term in the parenthesis is smaller than one, hence for large k both k-dependent terms in (3.9), i.e. $1/\sqrt{k}$ and $(e<k>/k)^k$ decrease rapidly with increasing k. Overall (3.9) predicts that in a random network the chance of observing a hub decreases faster than exponentially.

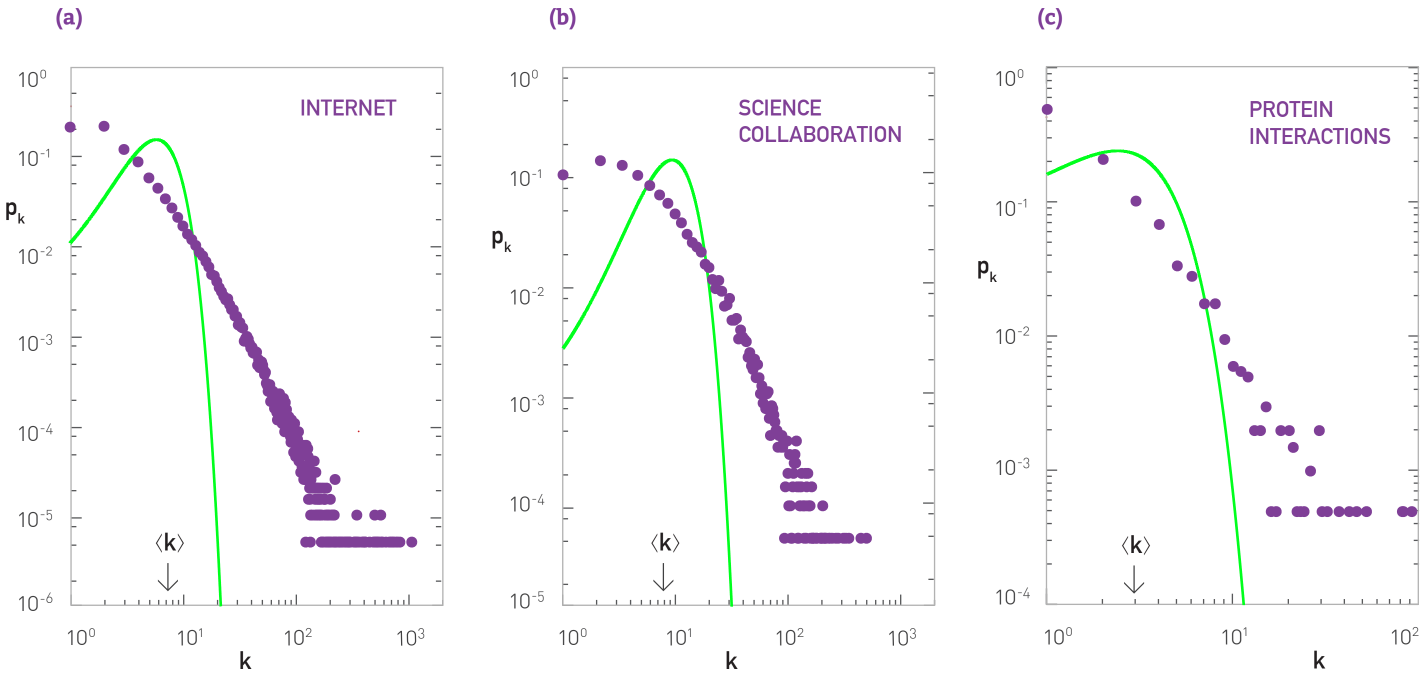


Figure 3.6

Degree Distribution of Real Networks

The degree distribution of the (a) Internet, (b) science collaboration network, and (c) protein interaction network (Table 2.1). The green line corresponds to the Poisson prediction, obtained by measuring <*k*> for the real network and then plotting (3.8). The significant deviation between the data and the Poisson fit indicates that the random network model underestimates the size and the frequency of the high degree nodes, as well as the number of low degree nodes. Instead the random network model predicts a larger number of nodes in the vicinity of <*k*> than seen in real networks.

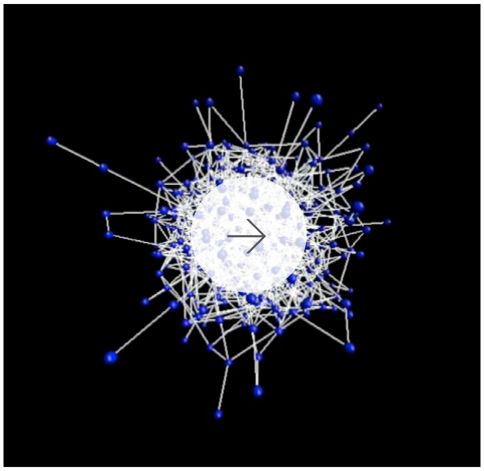
SECTION 3.6

THE EVOLUTION OF A RANDOM NETWORK

The cocktail party we encountered at the beginning of this chapter captures a dynamical process: Starting with N isolated nodes, the links are added gradually through random encounters between the guests. This corresponds to a gradual increase of p, with striking consequences on the network topology (Online Resource 3.2). To quantify this process, we first inspect how the size of the largest connected cluster within the network, N_G, varies with $<k>$. Two extreme cases are easy to understand:

- For $p = 0$ we have $<k> = 0$, hence all nodes are isolated. Therefore the largest component has size $N_G = 1$ and $N_G/N \rightarrow 0$ for large N.

- For $p = 1$ we have $<k>= N-1$, hence the network is a complete graph and all nodes belong to a single component. Therefore $N_G = N$ and $N_G/N = 1$.



Online Resource 3.2
Evolution of a Random Network

A video showing the change in the structure of a random network with increasing p. It vividly illustrates the absence of a giant component for small p and its sudden emergence once p reaches a critical value.

One would expect that the largest component grows gradually from $N_G = 1$ to $N_G = N$ if $<k>$ increases from 0 to N-1. Yet, as Figure 3.7a indicates, this is not the case: N_G/N remains zero for small $<k>$, indicating the lack of a large cluster. Once $<k>$ exceeds a critical value, N_G/N increases, signaling the rapid emergence of a large cluster that we call the *giant component*. Erdős and Rényi in their classical 1959 paper predicted that the condition for the emergence of the giant component is [2]

$$\langle k \rangle = 1. \tag{3.10}$$

In other words, we have a giant component if and only if each node has on average more than one link (ADVANCED TOPICS 3.C).

The fact that we need at least one link per node to observe a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node. It is somewhat counterintuitive, however, that one link is *sufficient* for its emergence.

We can express (3.10) in terms of p using (3.3), obtaining

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}, \qquad (3.11)$$

Therefore the larger a network, the smaller p is sufficient for the giant component.

The emergence of the giant component is only one of the transitions characterizing a random network as we change $<k>$. We can distinguish four topologically distinct regimes (Figure 3.7a), each with its unique characteristics:

Subcritical Regime: $0 < <k> < 1$ ($p < \frac{1}{N}$, Figure 3.7b).

For $<k> = 0$ the network consists of N isolated nodes. Increasing $<k>$ means that we are adding $N<k> = pN(N\text{-}1)/2$ links to the network. Yet, given that $<k> < 1$, we have only a small number of links in this regime, hence we mainly observe tiny clusters (Figure 3.7b).

We can designate at any moment the largest cluster to be the giant component. Yet in this regime the relative size of the largest cluster, N_G/N, remains zero. The reason is that for $<k> < 1$ the largest cluster is a tree with size $N_G \sim \ln N$, hence its size increases much slower than the size of the network. Therefore $N_G/N \simeq \ln N/N \rightarrow 0$ in the $N\rightarrow\infty$ limit.

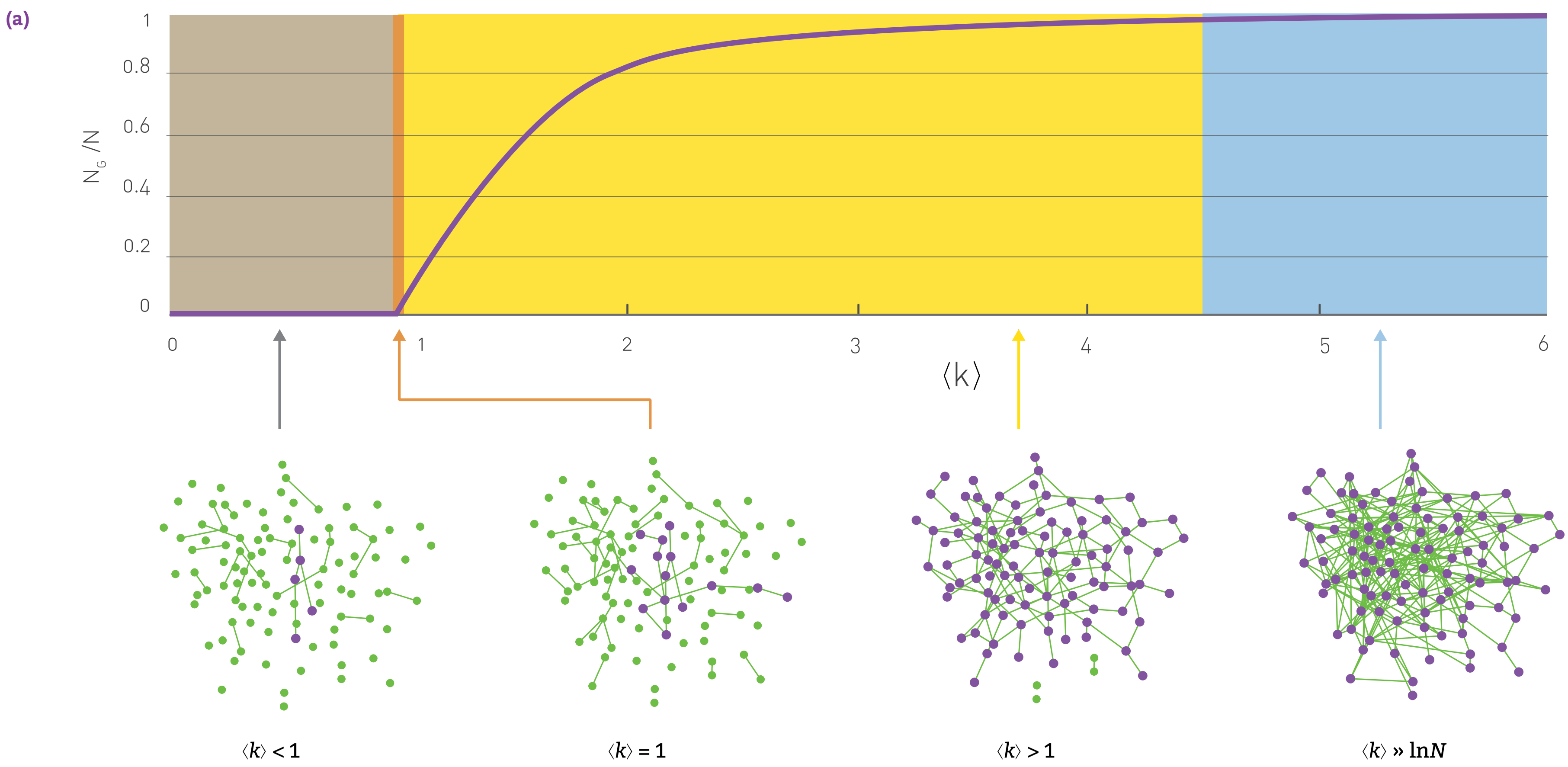
In summary, in the subcritical regime the network consists of numerous tiny components, whose size follows the exponential distribution (3.35). Hence these components have comparable sizes, lacking a clear winner that we could designate as a giant component.

Critical Point: $<k> = 1$ ($p = \frac{1}{N}$, Figure 3.7c).

The critical point separates the regime where there is not yet a giant component ($<k> < 1$) from the regime where there is one ($<k> > 1$). At this point the relative size of the largest component is still zero (Figure 3.7c). Indeed, the size of the largest component is $N_G \sim N^{2/3}$. Consequently N_G grows much slower than the network's size, so its relative size decreases as $N_G/N \sim N^{-1/3}$ in the $N\rightarrow\infty$ limit.

Note, however, that in absolute terms there is a significant jump in the size of the largest component at $<k> = 1$. For example, for a random network with $N = 7 \times 10^9$ nodes, comparable to the globe's social network, for $<k> < 1$ the largest cluster is of the order of $N_G \simeq \ln N = \ln (7 \times 10^9) \simeq 22.7$. In contrast at $<k> = 1$ we expect $N_G \sim N^{2/3} = (7 \times 10^9)^{2/3} \simeq 3 \times 10^6$, a jump of about five orders of magnitude. Yet, both in the subcritical regime and at the critical point the largest component contains only a vanishing fraction of the total number of nodes in the network.

In summary, at the critical point most nodes are located in numerous small components, whose size distribution follows (3.36). The power law form indicates that components of rather different sizes coexist. These numerous small components are mainly trees, while the giant component



(b) Subcritical Regime

- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the largest cluster: $N_G \sim \ln N$
- The clusters are trees

(c) Critical Point

- No giant component
- Cluster size distribution: $p_s \sim s^{-3/2}$
- Size of the largest cluster: $N_G \sim N^{3/2}$
- The clusters may contain loops

(d) Supercritical Regime

- Single giant component
- Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
- Size of the giant component: $N_G \sim (p - p_c)N$
- The small clusters are trees
- Giant component has loops

(e) Connected Regime

- Single giant component
- No isolated nodes or clusters
- Size of the giant component: $N_G = N$
- Giant component has loops

Figure 3.7

Evolution of a Random Network

(a) The relative size of the giant component in function of the average degree <*k*> in the Erdős-Rényi model. The figure illustrates the phase tranisition at <*k*> = 1, responsible for the emergence of a giant component with nonzero N_G.

(b-e) A sample network and its properties in the four regimes that characterize a random network.

may contain loops. Note that many properties of the network at the critical point resemble the properties of a physical system undergoing a phase transition (ADVANCED TOPICS 3.F).

Supercritical Regime: $<k> > 1$ ($p > \frac{1}{N}$, Figure 3.7d).

This regime has the most relevance to real systems, as for the first time we have a giant component that looks like a network. In the vicinity of the critical point the size of the giant component varies as

$$N_G / N \sim \langle k \rangle - 1, \tag{3.12}$$

or

$$N_G \sim (p - p_c)N, \tag{3.13}$$

where p_c is given by (3.11). In other words, the giant component contains a finite fraction of the nodes. The further we move from the critical point, a larger fraction of nodes will belong to it. Note that (3.12) is valid only in the vicinity of $<k> = 1$. For large $<k>$ the dependence between N_G and $<k>$ is nonlinear (Figure 3.7a).

In summary in the supercritical regime numerous isolated components coexist with the giant component, their size distribution following (3.35). These small components are trees, while the giant component contains loops and cycles. The supercritical regime lasts until all nodes are absorbed by the giant component.

Connected Regime: $\langle k \rangle > \ln N$ ($p > \frac{\ln N}{N}$, Figure 3.7e).

For sufficiently large p the giant component absorbs all nodes and components, hence $N_G \simeq N$. In the absence of isolated nodes the network becomes connected. The average degree at which this happens depends on N as (ADVANCED TOPIC 3.E)

$$\langle k \rangle = \ln N. \tag{3.14}$$

Note that when we enter the connected regime the network is still relatively sparse, as $\ln N / N \rightarrow 0$ for large N. The network turns into a complete graph only at $<k> = N - 1$.

In summary, the random network model predicts that the emergence of a network is not a smooth, gradual process: The isolated nodes and tiny components observed for small $<k>$ collapse into a giant component through a phase transition (ADVANCED TOPICS 3.F). As we vary $<k>$ we encounter four topologically distinct regimes (Figure 3.7).

The discussion offered above follows an empirical perspective, fruitful if we wish to compare a random network to real systems. A different perspective, with its own rich behavior, is offered by the mathematical literature (BOX 3.5).

BOX 3.5

NETWORK EVOLUTION IN GRAPH THEORY.

In the random graph literature it is often assumed that the connection probability $p(N)$ scales as N^z, where z is a tunable parameter between -∞ and 0 [15]. In this language Erdős and Rényi discovered that as we vary z, key properties of random graphs appear quite suddenly.

A graph has a given property Q if the probability of having Q approaches 1 as $N \to \infty$. That is, for a given z either almost every graph has the property Q or almost no graph has it. For example, for z less than -3/2 almost all graphs contain only isolated nodes and pairs of nodes connected by a link. Once z exceeds -3/2, most networks will contain paths connecting three or more nodes (Figure 3.8).

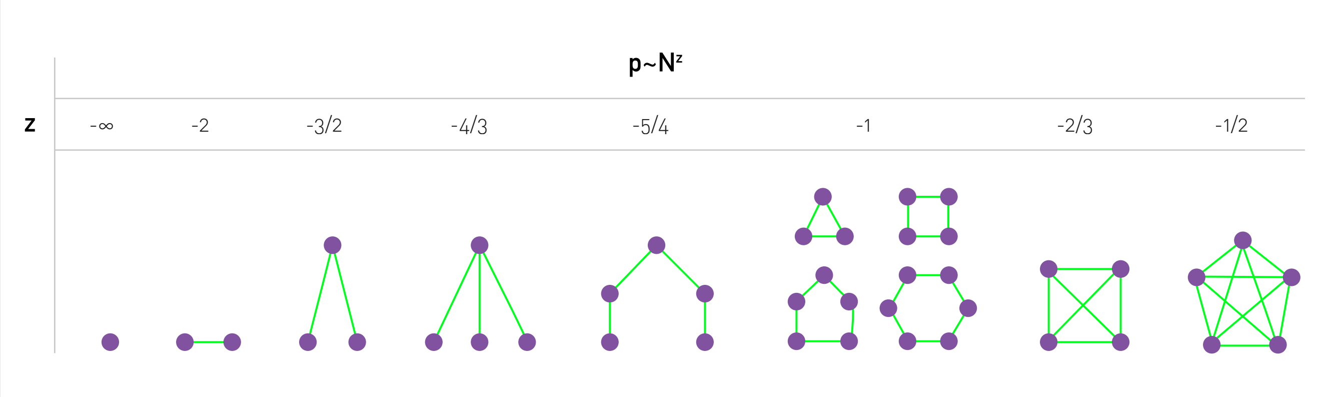


Figure 3.8
Evolution of a Random Graph

The threshold probabilities at which different subgraphs appear in a random graph, as defined by the exponent z in the $p(N) \sim N^z$ relationship. For $z < -3/2$ the graph consists of isolated nodes and edges. When z passes -3/2 trees of order 3 appear, while at $z = -4/3$ trees of order 4 appear. At $z = 1$ trees of all orders are present, together with cycles of all orders. Complete subgraphs of order 4 appear at $z =-2/3$, and as z increases further, complete subgraphs of larger and larger order emerge. After [19].

REAL NETWORKS ARE SUPERCRITICAL

Two predictions of random network theory are of direct importance for real networks:

1) Once the average degree exceeds $\langle k \rangle = 1$, a giant component should emerge that contains a finite fraction of all nodes. Hence only for $\langle k \rangle > 1$ the nodes organize themselves into a recognizable network.

2) For $\langle k \rangle > \ln N$ all components are absorbed by the giant component, resulting in a single connected network.

Do real networks satisfy the criteria for the existence of a giant component, i.e. $\langle k \rangle > 1$? And will this giant component contain all nodes for $\langle k \rangle > \ln N$, or will we continue to see some disconnected nodes and components? To answer these questions we compare the structure of a real network for a given $\langle k \rangle$ with the theoretical predictions discussed above.

The measurements indicate that real networks extravagantly exceed the $\langle k \rangle = 1$ threshold. Indeed, sociologists estimate that an average person has around 1,000 acquaintances; a typical neuron is the human brain has about 7,000 synapses; in our cells each molecule takes part in several chemical reactions.

This conclusion is supported by Table 3.1, that lists the average degree of several undirected networks, in each case finding $\langle k \rangle > 1$. Hence the average degree of real networks is well beyond the $\langle k \rangle = 1$ threshold, implying that they all have a giant component. The same is true for the reference networks listed in Table 3.1.

Let us now turn to the second prediction, inspecting if we have single component (i.e. if $\langle k \rangle > \ln N$), or if the network is fragmented into multiple components (i.e. if $\langle k \rangle < \ln N$). For social networks the transition between the supercritical and the fully connected regime should be at $\langle k \rangle > \ln(7 \times 10^9) \approx 22.7$. That is, if the average individual has more than two dozens acquaintances, then a random society must have a single component, leav-

NETWORK	N	L	⟨k⟩	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,439	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

Table 3.1

Are Real Networks Connected?

The number of nodes N and links L for the undirected networks of our reference network list of Table 3.1, shown together with $\langle k \rangle$ and $\ln N$. A giant component is expected for $\langle k \rangle > 1$ and all nodes should join the giant component for $\langle k \rangle > \ln N$. While for all networks $\langle k \rangle > 1$, for most $\langle k \rangle$ is under the $\ln N$ threshold (see also Figure 3.9).

ing no individual disconnected. With <k> ≈ 1,000 this condition is clearly satisfied. Yet, according to Table 3.1 many real networks do not obey the fully connected criteria. Consequently, according to random network theory these networks should be fragmented into several disconnected components. This is a disconcerting prediction for the Internet, indicating that some routers should be disconnected from the giant component, being unable to communicate with other routers. It is equally problematic for the power grid, indicating that some consumers should not get power. These predictions are clearly at odds with reality.

In summary, we find that most real networks are in the supercritical regime (Figure 3.9). Therefore these networks are expected to have a giant component, which is in agreement with the observations. Yet, this giant component should coexist with many disconnected components, a prediction that fails for several real networks. Note that these predictions should be valid only if real networks are accurately described by the Erdős-Rényi model, i.e. if real networks are random. In the coming chapters, as we learn more about the structure of real networks, we will understand why real networks can stay connected despite failing the $k > \ln N$ criteria.

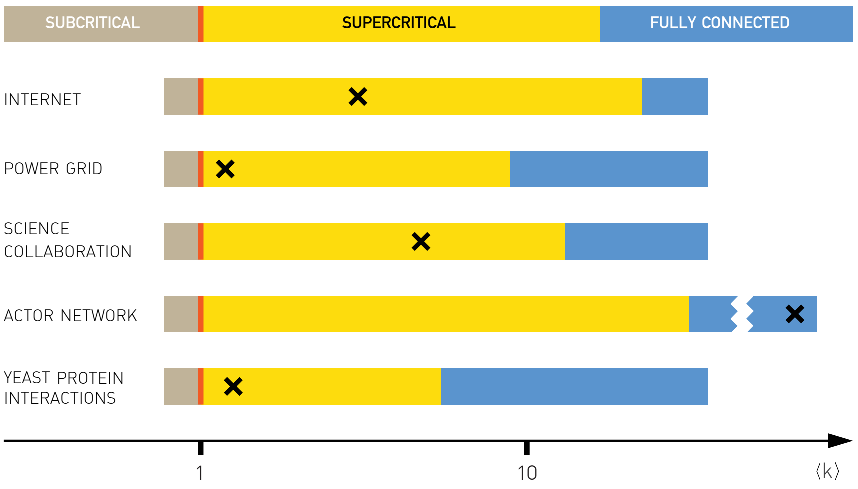


Figure 3.9

Most Real Networks are Supercritical

The four regimes predicted by random network theory, marking with a cross the location (<k>) of the undirected networks listed in Table 3.1. The diagram indicates that most networks are in the supercritical regime, hence they are expected to be broken into numerous isolated components. Only the actor network is in the connected regime, meaning that all nodes are part of a single giant component. Note that while the boundary between the subcritical and the supercritical regime is always at <k> = 1, the boundary between the supercritical and the connected regime is at $\ln N$, which varies from system to system.

SMALL WORLDS

The *small world phenomenon*, also known as *six degrees of separation*, has long fascinated the general public. It states that if you choose any two individuals anywhere on Earth, you will find a path of at most six acquaintances between them (Figure 3.10). The fact that individuals who live in the same city are only a few handshakes from each other is by no means surprising. The small world concept states, however, that even individuals who are on the opposite side of the globe can be connected to us via a few acquaintances.

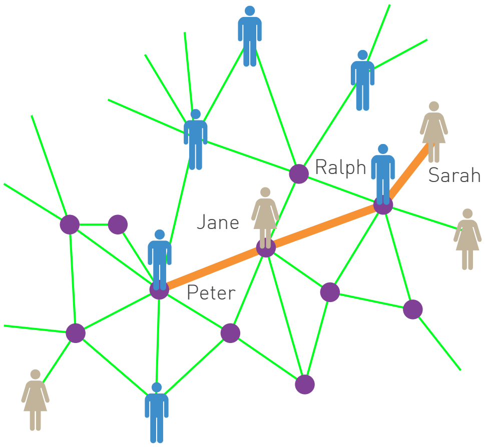


Figure 3.10

Six Deegree of Separation

According to six degrees of separation two individuals, anywhere in the world, can be connected through a chain of six or fewer acquaintances. This means that while Sarah does not know Peter, she knows Ralph, who knows Jane and who in turn knows Peter. Hence Sarah is three handshakes, or three degrees from Peter. In the language of network science six degrees, also called the small world property, means that the distance between any two nodes in a network is unexpectedly small.

In the language of network science the small world phenomenon implies that *the distance between two randomly chosen nodes in a network is short.* This statement raises two questions: What does short (or small) mean, i.e. short compared to what? How do we explain the existence of these short distances?

Both questions are answered by a simple calculation. Consider a random network with average degree $<k>$. A node in this network has on average:

$<k>$ nodes at distance one (d=1).
$<k>^2$ nodes at distance two (d=2).
$<k>^3$ nodes at distance three (d =3).
...
$<k>^d$ nodes at distance d.

For example, if $<k> \approx 1{,}000$, which is the estimated number of acquaintences an individual has, we expect 10^6 individuals at distance two and about a billion, i.e. almost the whole earth's population, at distance three from us.

To be precise, the expected number of nodes up to distance d from our starting node is

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}. \qquad (3.15)$$

N(*d*) must not exceed the total number of nodes, *N*, in the network. Therefore the distances cannot take up arbitrary values. We can identify the maximum distance, d_{max}, or the network's diameter by setting

$$N(d_{max}) \approx N. \tag{3.16}$$

Assuming that <*k*> » 1, we can neglect the (-1) term in the nominator and the denominator of (3.15), obtaining

$$\langle k \rangle^{d_{max}} \approx N. \tag{3.17}$$

Therefore the diameter of a random network follows

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle}, \tag{3.18}$$

which represents the mathematical formulation of the small world phenomenon. The key, however is its interpretation:

- As derived, (3.18) predicts the scaling of the network diameter, d_{max}, with the size of the system, *N*. Yet, for most networks (3.18) offers a better approximation to the average distance between two randomly chosen nodes, <*d*>, than to d_{max} (Table 3.2). This is because d_{max} is often dominated by a few extreme paths, while <*d*> is averaged over all node pairs, a process that supresses the fluctuations. Hence typically the small world property is defined by

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}, \tag{3.19}$$

  describing the dependence of the average distance in a network on *N* and <*k*>.

- In general ln*N* « *N*, hence the dependence of <*d*> on ln*N* implies that the distances in a random network are *orders of magnitude smaller than the size of the network*. Consequently by *small* in the "small world phenomenon" we mean that *the average path length or the diameter depends logarithmically on the system size*. Hence, "small" means that <*d*> is proportional to ln*N*, rather than *N* or some power of *N* (Figure 3.11).

- The 1/ln <*k*> term implies that the denser the network, the smaller is the distance between the nodes.

- In real networks there are systematic corrections to (3.19), rooted in the fact that the number of nodes at distance *d* > <*d*> drops rapidly (ADVANCED TOPICS 3.F).

Let us illustrate the implications of (3.19) for social networks. Using *N* ≈ 7 ×10⁹ and <*k*> ≈ 10³, we obtain

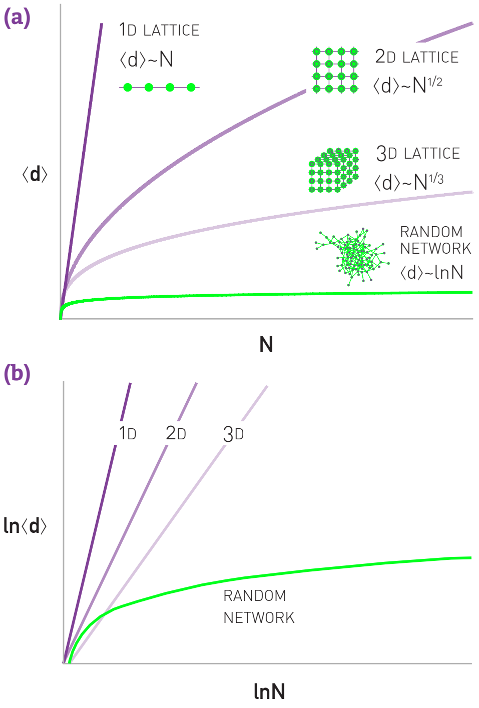


Figure 3.11

Why are Small Worlds Surprising?

Much of our intuition about distance is based on our experience with regular lattices, which do not display the small world property:

1D: For a one-dimensional lattice (a line of length *N*) the diameter and the average path length scale linearly with *N*: d_{max}~<*d*> ~*N*.

2D: For a square lattice d_{max}~<*d*> ~ $N^{1/2}$.

3D: For a cubic lattice d_{max}~<*d*> ~ $N^{1/3}$.

4D: In general, for a *d*-dimensional lattice d_{max} ~ <*d*> ~ $N^{1/d}$.

These polynomial dependences predict a much faster increase with *N* than (3.19), indicating that in lattices the path lengths are significantly longer than in a random network. For example, if the social network would form a square lattice (2*D*), where each individual knows only its neighbors, the average distance between two individuals would be roughly $(7 \times 10^9)^{1/2}$ = 83,666. Even if we correct for the fact that a person has about 1,000 acquaintances, not four, the average separation will be orders of magnitude larger than predicted by (3.19).

(a) The figure shows the predicted *N*-dependence of <*d*> for regular and random networks on a linear scale.
(b) The same as in **(a)**, but shown on a log-log scale.

$$\langle d \rangle \approx \frac{\ln 7 \times 10^9}{\ln(10^3)} = 3.28. \quad (3.20)$$

Therefore, all individuals on Earth should be within three to four handshakes of each other [20]. The estimate (3.20) is probably closer to the real value than the frequently quoted six degrees (BOX 3.7).

Much of what we know about the small world property in random networks, including the result (3.19), is in a little known paper by Manfred Kochen and Ithiel de Sola Pool [20], in which they mathematically formulated the problem and discussed in depth its sociological implications. This paper inspired the well known Milgram experiment (BOX 3.6), which in turn inspired the *six-degrees of separation* phrase.

While discovered in the context of social systems, the small world property applies beyond social networks (BOX 3.6). To demonstrate this in Table 3.2 we compare the prediction of (3.19) with the average path length <*d*> for several real networks, finding that despite the diversity of these systems and the significant differences between them in terms of *N* and <*k*>, (3.19) offers a good approximation to the empirically observed <*d*>.

In summary the small world property has not only ignited the public's imagination (BOX 3.8), but plays an important role in network science as well. The small world phenomena can be reasonably well understood in the context of the random network model: It is rooted in the fact that the number of nodes at distance *d* from a node increases exponentially with *d*. In the coming chapters we will see that in real networks we encounter systematic deviations from (3.19), forcing us to replace it with more accurate predictions. Yet the intuition offered by the random network model on the origin of the small world phenomenon remains valid.

NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81
Actor Network	702,388	29,397,908	83,71	3,91	14	3,04
Citation Network	449,673	4,707,958	10.43	11,21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Table 3.2
Six Degrees of Separation

The average distance <*d*> and the maximum distance $d_{\max}$ for the ten reference networks. The last column provides <*d*> predicted by (3.19), indicating that it offers a reasonable approximation to the measured <*d*>. Yet, the agreement is not perfect - we will see in the next chapter that for many real networks (3.19) needs to be adjusted. For directed networks the average degree and the path lengths are measured along the direction of the links.

BOX 3.6

19 DEGREES OF SEPARATION

How many clicks do we need to reach a randomly chosen document on the Web? The difficulty in addressing this question is rooted in the fact that we lack a complete map of the WWW—we only have access to small samples of the full map. We can start, however, by measuring the WWW's average path length in samples of increasing sizes, a procedure called *finite size scaling*. The measurements indicate that the average path length of the WWW increases with the size of the network as [21]

$$\langle d \rangle \approx 0.35 + 0.89 \ln N.$$

In 1999 the WWW was estimated to have about 800 million documents [22], in which case the above equation predicts <*d*>≈18.69. In other words in 1999 two randomly chosen documents were on average 19 clicks from each other, a result that became known as *19 degrees of separation*. Subsequent measurements on a sample of 200 million documents found <*d*>≈16 [23], in good agreement with the <*d*>≈17 prediction. Currently the WWW is estimated to have about trillion nodes ($N \sim 10^{12}$), in which case the formula predicts <*d*>≈25. Hence <*d*> is not fixed but as the network grows, so does the distance between two documents.

The average path length of 25 is much larger than the proverbial six degrees (BOX 3.7). The difference is easy to understand: The WWW has smaller average degree and larger size than the social network. According to (3.19) both of these differences increase the Web's diameter.

BOX 3.7

SIX DEGREES: EXPERIMENTAL CONFIRMATION

The first empirical study of the small world phenomena took place in 1967, when Stanley Milgram, building on the work of Pool and Kochen [20], designed an experiment to measure the distances in social networks [24, 25]. Milgram chose a stock broker in Boston and a divinity student in Sharon, Massachusetts as *targets*. He then randomly selected residents of Wichita and Omaha, sending them a letter containing a short summary of the study's purpose, a photograph, the name, address and information about the target person. They were asked to forward the letter to a friend, relative or acquantance who is most likely to know the target person.

Within a few days the first letter arrived, passing through only two links. Eventually 64 of the 296 letters made it back, some, however, requiring close to a dozen intermediates [25]. These completed chains allowed Milgram to determine the number of individuals required to get the letter to the target (Figure 3.12a). He found that the median number of intermediates was 5.2, a relatively small number that was remarkably close to Frigyes Karinthy's 1929 insight (BOX 3.8).

Milgram lacked an accurate map of the full acquaintance network, hence his experiment could not detect the true distance between his study's participants. Today Facebook has the most extensive social network map ever assembled. Using Facebook's social graph of May 2011, consisting of 721 million active users and 68 billion symmetric friendship links, researchers found an average distance 4.74 between the users (Figure 3.12b). Therefore, the study detected only 'four degrees of separation' [18], closer to the prediction of (3.20) than to Milgram's six degrees [24, 25].

> *"I asked a person of intelligence how many steps he thought it would take, and he said that it would require 100 intermediate persons, or more, to move from Nebraska to Sharon."*
>
> *Stanley Milgram, 1969*

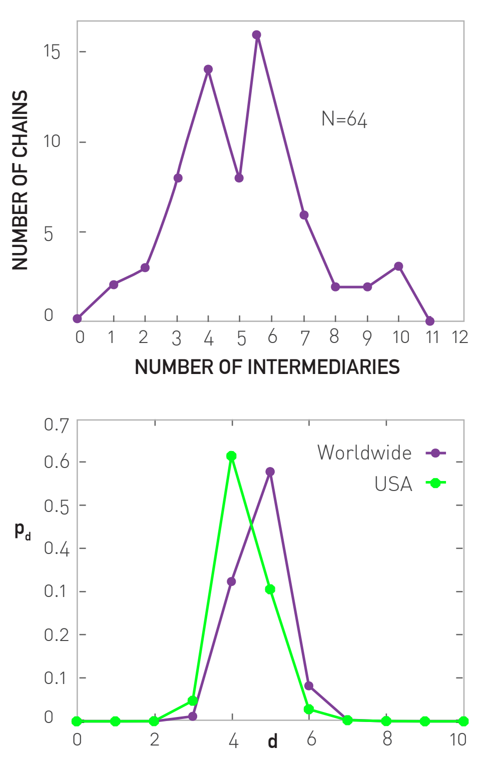


Figure 3.12

Six Degrees? From Milgram to Facebook

(a) In Milgram's experiment 64 of the 296 letters made it to the recipient. The figure shows the length distribution of the completed chains, indicating that some letters required only one intermediary, while others required as many as ten. The mean of the distribution was 5.2, indicating that on average six 'handshakes' were required to get a letter to its recipient. The playwright John Guare renamed this 'six degrees of separation' two decades later. After [25].

(b) The distance distribution, p_d , for all pairs of Facebook users worldwide and within the US only.Using Facebook's N and L (3.19) predicts the average degree to be approximately 3.90, not far from the reported four degrees. After [18].

BOX 3.8

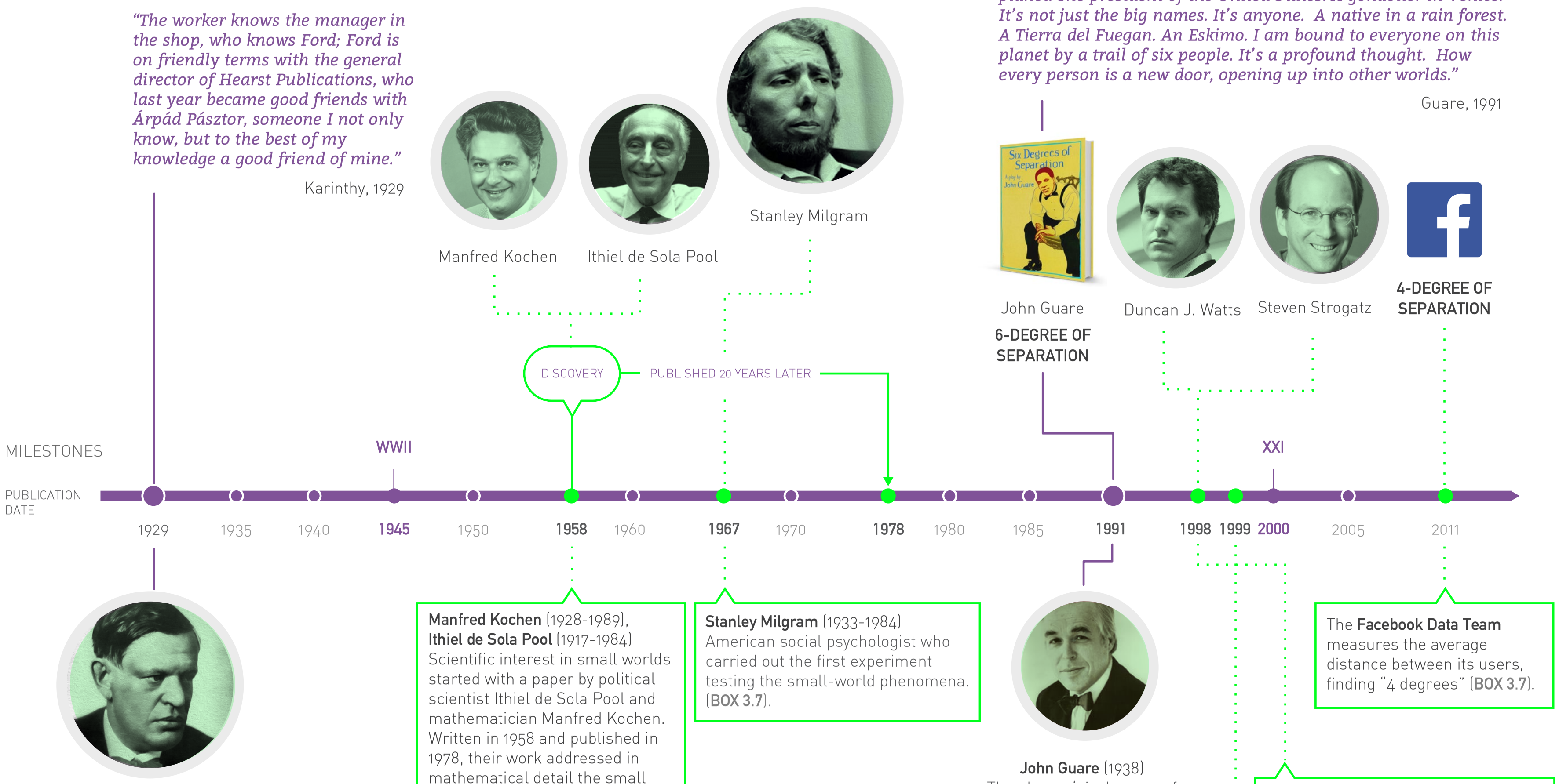
19 DEGREES OF THE WWW

"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Árpád Pásztor, someone I not only know, but to the best of my knowledge a good friend of mine."

Karinthy, 1929

"Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people. It's a profound thought. How every person is a new door, opening up into other worlds."

Guare, 1991



Frigyes Karinthy (1887-1938)
Hungarian writer, journalist and playwright, the first to describe the small world property. In his short story entitled 'Láncszemek' (Chains) he links a worker in Ford's factory to himself [26, 27].

Manfred Kochen (1928-1989), **Ithiel de Sola Pool** (1917-1984)
Scientific interest in small worlds started with a paper by political scientist Ithiel de Sola Pool and mathematician Manfred Kochen. Written in 1958 and published in 1978, their work addressed in mathematical detail the small world effect, predicting that most individuals can be connected via two to three acquaintances. Their paper inspired the experiments of Stanley Milgram.

Stanley Milgram (1933-1984)
American social psychologist who carried out the first experiment testing the small-world phenomena. (BOX 3.7).

19 Degrees of the WWW
Measurements on the WWW indicate that the separation between two randomly chosen documents is 19 [21] (Box 3.6).

John Guare (1938)
The phrase 'six degrees of separation' was introduced by the playwright John Guare, who used it as the title of his Broadway play [28].

The **Facebook Data Team** measures the average distance between its users, finding "4 degrees" (BOX 3.7).

Duncan J. Watts (1971), **Steven Strogatz** (1959)
A new wave of interest in small worlds followed the study of Watts and Strogatz, finding that the small world property applies to natural and technological networks as well [29].

CLUSTERING COEFFICIENT

The degree of a node contains no information about the relationship between a node's neighbors. Do they all know each other, or are they perhaps isolated from each other? The answer is provided by the local clustering coefficient C_i, that measures the density of links in node *i*'s immediate neighborhood: $C_i = 0$ means that there are no links between *i*'s neighbors; $C_i = 1$ implies that each of the *i*'s neighbors link to each other (SECTION 2.10).

To calculate C_i for a node in a random network we need to estimate the expected number of links L_i between the node's k_i neighbors. In a random network the probability that two of *i*'s neighbors link to each other is *p*. As there are $k_\mathrm{i}(k_\mathrm{i}$ - 1)/2 possible links between the k_i neighbors of node *i*, the expected value of L_i is

$$\langle L_i \rangle = p\frac{k_i(k_i - 1)}{2}. \tag{3.20}$$

Thus the local clustering coefficient of a random network is

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}. \tag{3.21}$$

Equation (3.21) makes two predictions:

(1) For fixed <*k*>, the larger the network, the smaller is a node's clustering coefficient. Consequently a node's local clustering coefficient C_i is expected to decrease as 1/*N*. Note that the network's average clustering coefficient, <*C*> also follows (3.21).

(2) The local clustering coefficient of a node is independent of the node's degree.

To test the validity of (3.21) we plot <*C*>/<*k*> in function of *N* for several undirected networks (Figure 3.13a). We find that <*C*>/<k> does not decrease as N^{-1}, but it is largely independent of *N*, in violation of the prediction (3.21) and point (1) above. In Figure 3.13b-d we also show the dependency of C on the node's degree k_i for three real networks, finding that $C(k)$ systematically decreases with the degree, again in violation of (3.21) and point (2).

In summary, we find that the random network model does not capture the clustering of real networks. Instead real networks have a much higher clustering coefficient than expected for a random network of similar N and L. An extension of the random network model proposed by Watts and Strogatz [29] addresses the coexistence of high $<C>$ and the small world property (BOX 3.9). It fails to explain, however, why high-degree nodes have a smaller clustering coefficient than low-degree nodes. Models explaining the shape of $C(k)$ are discussed in Chapter 9.

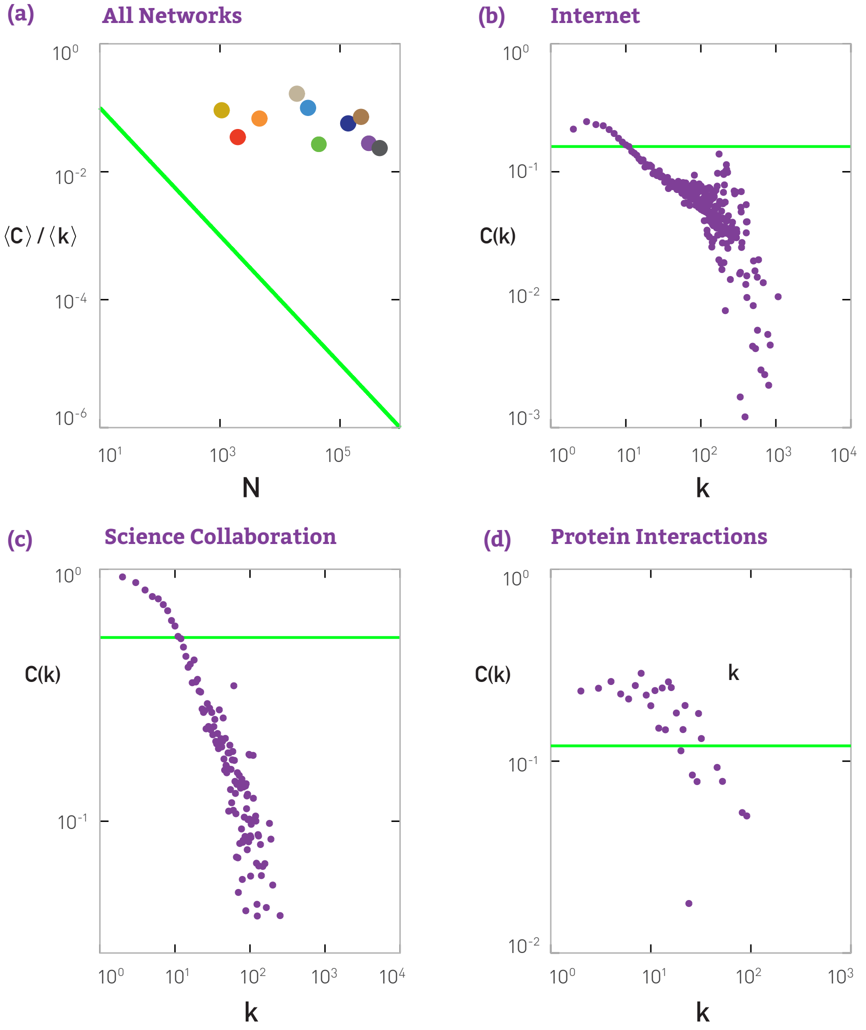


Figure 3.13

Clustering in Real Networks

(a) Comparing the average clustering coefficient of real networks with the prediction (3.21) for random networks. The circles and their colors correspond to the networks of Table 3.2. Directed networks were made undirected to calculate $<C>$ and $<k>$. The green line corresponds to (3.21), predicting that for random networks the average clustering coefficient decreases as N^{-1}. In contrast, for real networks $<C>$ appears to be independent of N.

(b)-(d) The dependence of the local clustering coefficient, C(k), on the node's degree for (b) the Internet, (c) science collaboration network and (d) protein interaction network. $C(k)$ is measured by averaging the local clustering coefficient of all nodes with the same degree k. The green horizontal line corresponds to $<C>$.

BOX 3.9

WATTS-STROGATZ MODEL

Duncan Watts and Steven Strogatz proposed an extension of the random network model (Figure 3.14) motivated by two observations [29]:

(a) Small World Property

In real networks the average distance between two nodes depends logarithmically on N (3.18), rather than following a polynomial expected for regular lattices (Figure 3.11).

(b) High Clustering

The average clustering coefficient of real networks is much higher than expected for a random network of similar N and L (Figure 3.13a).

The *Watts-Strogatz model* (also called the *small-world model*) interpolates between a *regular lattice*, which has high clustering but lacks the small-world phenomenon, and a *random network*, which has low clustering, but displays the small-world property (Figure 3.14a-c). Numerical simulations indicate that for a range of rewiring parameters the model's average path length is low but the clustering coefficient is high, hence reproducing the coexistence of high clustering and small-world phenomena (Figure 3.14d).

Being an extension of the random network model, the Watts-Strogatz model predicts a Poisson-like bounded degree distribution. Consequently high degree nodes, like those seen in Figure 3.6, are absent from it. Furthermore it predicts a k-independent $C(k)$, being unable to recover the k-dependence observed in Figures 3.13b-d. As we show in the next chapters, understanding the coexistence of the small world property with high clustering must start from the network's correct degree distribution.

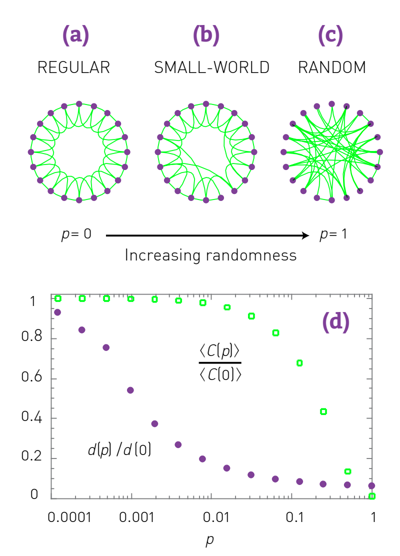


Figure 3.14

The Watts-Strogatz Model

(a) We start from a ring of nodes, each node being connected to their immediate and next neighbors. Hence initially each node has $<C> = 3/4$ ($p = 0$).

(b) With probability p each link is rewired to a randomly chosen node. For small p the network maintains high clustering but the random long-range links can drastically decrease the distances between the nodes.

(c) For $p = 1$ all links have been rewired, so the network turns into a random network.

(d) The dependence of the average path length $d(p)$ and clustering coefficient $<C(p)>$ on the rewiring parameter p. Note that $d(p)$ and $<C(p)>$ have been normalized by $d(0)$ and $<C(0)>$ obtained for a regular lattice (i.e. for $p=0$ in (a)). The rapid drop in $d(p)$ signals the onset of the small-world phenomenon. During this drop, $<C(p)>$ remains high. Hence in the range $0.001<p<0.1$ short path lengths and high clustering coexist in the network. All graphs have N=1000 and $<k>$=10. After [29].

ADVANCED TOPICS 3.A
DERIVING THE POISSON DISTRIBUTION

To derive the Poisson form of the degree distribution we start from the exact binomial distribution (3.7)

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \qquad (3.22)$$

that characterizes a random graph. We rewrite the first term on the r.h.s. as

$$\binom{N-1}{k} = \frac{(N-1)(N-1-1)(N-1-2)...(N-1-k+1)}{k!} \approx \frac{(N-1)^k}{k!}, \qquad (3.23)$$

where in the last term we used that k « N. The last term of (3.22) can be simplified as

$$\ln[(1-p)^{(N-1)-k}] = (N-1-k)\ln(1-\frac{\langle k\rangle}{N-1})$$

and using the series expansion

$$\ln(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - ..., \forall \mid x \mid \le 1$$

we obtain

$$\ln[(1-p)^{N-1-k}] \approx (N-1-k)\frac{\langle k\rangle}{N-1} = -\langle k\rangle(1-\frac{k}{N-1}) \approx -\langle k\rangle,$$

which is valid if N » k. This represents the *small degree approximation* at the heart of this derivation. Therefore the last term of (3.22) becomes

$$(1-p)^{N-1-k} = e^{-\langle k\rangle}. \qquad (3.24)$$

Combining (3.22), (3.23), and (3.24) we obtain the Poisson form of the degree distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} = \frac{(N-1)^k}{k!} p^k e^{-\langle k \rangle}$$

$$= \frac{(N-1)^k}{k!} \left(\frac{\langle k \rangle}{N-1} \right)^k e^{-\langle k \rangle},$$

or

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \qquad (3.25)$$

ADVANCED TOPICS 3.B
MAXIMUM AND MINIMUM DEGREES

To determine the expected degree of the largest node in a random network, called the network's *upper natural cutoff*, we define the degree $k_{\max}$ such that in a network of N nodes we have at most one node with degree higher than $k_{\max}$. Mathematically this means that the area behind the Poisson distribution p_k for $k \geq k_{\max}$ should be approximately one (Figure 3.17). Since the area is given by 1-$P(k_{\max})$, where $P(k)$ is the cumulative degree distribution of p_k, the network's largest node satisfies:

$$N\left[1-P(k_{max})\right] \approx 1. \tag{3.26}$$

We write ≈ instead of =, because $k_{\max}$ is an integer, so in general the exact equation does not have a solution. For a Poisson distribution

$$1-P(k_{max}) = 1-e^{-\langle k\rangle}\sum_{k=0}^{k_{max}}\frac{\langle k\rangle^k}{k!} = e^{-\langle k\rangle}\sum_{k=k_{max}+1}^{\infty}\frac{\langle k\rangle^k}{k!} \approx e^{-\langle k\rangle}\frac{\langle k\rangle^{k_{max}+1}}{(k_{max}+1)!}, \tag{3.27}$$

where in the last term we approximate the sum with its largest term.

For $N = 10^9$ and $<k> = 1,000$, roughly the size and the average degree of the globe's social network, (3.26) and (3.27) predict $k_{\max}$ = 1,185, indicating that a random network lacks extremely popular individuals, or hubs.

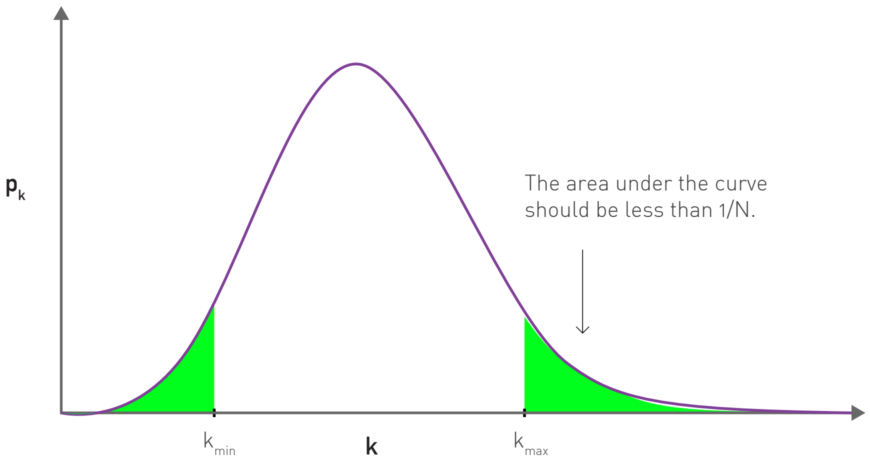
We can use a similar argument to calculate the expected degree of the smallest node, $k_{\min}$. By requiring that there should be at most one node with degree smaller than $k_{\min}$ we can write

$$NP(k_{min}-1) \approx 1. \tag{3.28}$$

For the Erdős-Rényi network we have

$$P(k_{min} - 1) = e^{-\langle k \rangle} \sum_{k=0}^{k_{min}-1} \frac{\langle k \rangle^k}{k!} \,. \qquad (3.29)$$

Solving (3.28) with $N = 10^9$ and $<k> = 1{,}000$ we obtain $k_{\text{min}} = 816$.



The area under the curve should be less than 1/N.

Figure 3.17

Minimum and Maximum Degree

The estimated maximum degree of a network, k_{max}, is chosen so that there is at most one node whose degree is higher than k_{max}. This is often called the *natural upper cutoff* of a degree distribution. To calculate it, we need to set k_{max} such that the area under the degree distribution p_k for $k > k_{\text{max}}$ equals $1/N$, hence the total number of nodes expected in this region is exactly one. We follow a similar argument to determine the expected smallest degree, k_{min}.

ADVANCED TOPICS 3.C
GIANT COMPONENT

In this section we introduce the argument, proposed independently by Solomonoff and Rapoport [11], and by Erdős and Rényi [2], for the emergence of giant component at $<k>= 1$ [33].

Let us denote with $u = 1 - N_G/N$ the fraction of nodes that are not in the giant component (*GC*), whose size we take to be N_G. If node i is part of the *GC*, it must link to another node j, which must also be part of the *GC*. Hence if i is *not* part of the *GC*, that could happen for two reasons:

- There is no link between i and j (probability for this is $1- p$).
- There is a link between i and j, but j is not part of the *GC* (probability for this is pu).

Therefore the total probability that i is not part of the *GC* via node j is $1 - p + pu$. The probability that i is not linked to the *GC* via any other node is therefore $(1 - p + pu)^{N - 1}$, as there are $N - 1$ nodes that could serve as potential links to the *GC* for node i. As u is the fraction of nodes that do not belong to the *GC*, for any p and N the solution of the equation

$$u = (1 - p + pu)^{N-1} \tag{3.30}$$

provides the size of the giant component via $N_G = N(1 - u)$. Using $p = <k> / (N - 1)$ and taking the logarithm of both sides, for $<k> \ll N$ we obtain

$$\ln u = (N-1)\ln\left[1-\frac{\langle k\rangle}{N-1}(1-u)\right] \approx (N-1)\left[-\frac{\langle k\rangle}{N-1}(1-u)\right] = -\langle k\rangle(1-u), \tag{3.31}$$

where we used the series expansion for $\ln(1+x)$.

Taking an exponential of both sides leads to $u = exp[- <k>(1 - u)]$. If we denote with S the fraction of nodes in the giant component, $S = N_G / N$, then $S = 1 - u$ and (3.31) results in

$$S = 1 - e^{-\langle k \rangle S}. \qquad (3.32)$$

This equation provides the size of the giant component *S* in function of <*k*> (Figure 3.18). While (3.32) looks simple, it does not have a closed solution. We can solve it graphically by plotting the right hand side of (3.32) as a function of *S* for various values of <*k*>. To have a nonzero solution, the obtained curve must intersect with the dotted diagonal, representing the left hand side of (3.32). For small <*k*> the two curves intersect each other only at *S* = 0, indicating that for small <*k*> the size of the giant component is zero. Only when <*k*> exceeds a threshold value, does a non-zero solution emerge.

To determine the value of <*k*> at which we start having a nonzero solution we take a derivative of (3.32), as the phase transition point is when the r.h.s. of (3.32) has the same derivative as the l.h.s. of (3.32), i.e. when

$$\frac{d}{dS}\left(1 - e^{-\langle k \rangle S}\right) = 1,$$

$$(3.33)$$

$$\langle k \rangle e^{-\langle k \rangle S} = 1$$

Setting *S* = 0, we obtain that the phase transition point is at <*k*> = 1 (see also ADVANCED TOPICS 3.F).

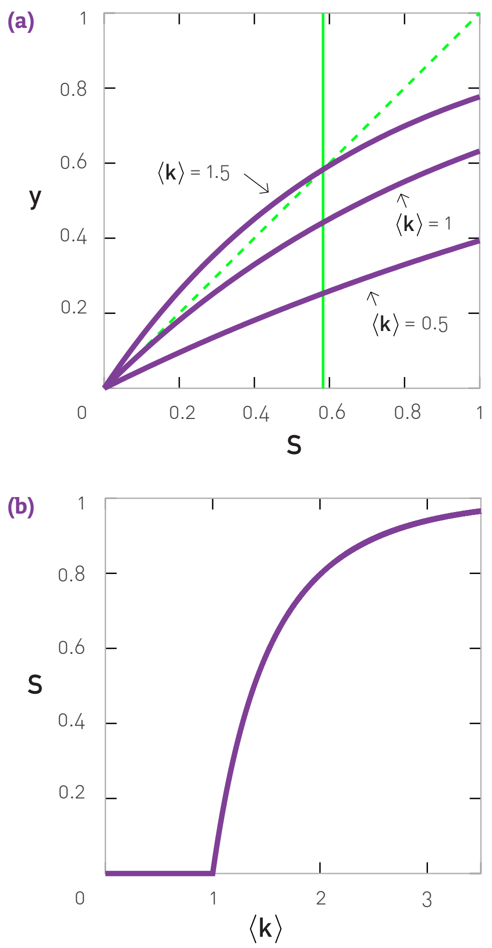


Figure 3.18
Graphical Solution

(a) The three purple curves correspond to y = 1-*exp*[-<*k*> *S*] for <*k*>=*0.5, 1, 1.5*. The green dashed diagonal corresponds y = *S*, and the intersection of the dashed and purple curves provides the solution to (3.32). For <*k*>=*0.5* there is only one intersection at *S* = 0, indicating the absence of a giant component. The <*k*>=*1.5* curve has a solution at *S* = 0.583 (green vertical line). The <*k*>=*1* curve is precisely at the critical point, representing the separation between the regime where a nonzero solution for *S* exists and the regime where there is only the solution at *S* = 0.

(b) The size of the giant component in function of <*k*> as predicted by (3.32). After [33].

ADVANCED TOPICS 3.D
COMPONENT SIZES

In Figure 3.7 we explored the size of the giant component, leaving an important question open: How many components do we expect for a given <*k*>? What is their size distribution? The aim of this section is to discuss these topics.

Component Size Distribution

For a random network the probability that a randomly chosen node belongs to a component of size *s* (which is different from the giant component *G*) is [33]

$$p_s \sim \frac{(s\langle k\rangle)^{s-1}}{s!} e^{-\langle k\rangle s}. \tag{3.34}$$

Replacing <*k*>$^{s-1}$ with *exp*[(*s*-1) ln<*k*>] and using the Stirling-formula

$s! \approx \sqrt{2\pi s}\left(\frac{s}{e}\right)^s$ for large *s* we obtain

$$p_s \sim s^{-3/2} e^{-(\langle k\rangle-1)s+(s-1)\ln\langle k\rangle}. \tag{3.35}$$

Therefore the component size distribution has two contributions: a slowly decreasing power law term $s^{-3/2}$ and a rapidly decreasing exponential term $e^{-(<k>-1)s+(s-1)\ln<k>}$. Given that the exponential term dominates for large *s*, (3.35) predicts that large components are prohibited. At the *critical point*, <*k*> = 1, all terms in the exponential cancel, hence p_s follows the power law

$$p_s \sim s^{-3/2}. \tag{3.36}$$

As a power law decreases relatively slowly, at the critical point we expect to observe clusters of widely different sizes, a property consistent with the behavior of a system during a phase transition (ADVANCED TOPICS 3.F). These predictions are supported by the numerical simulations shown in Figure 3.19.

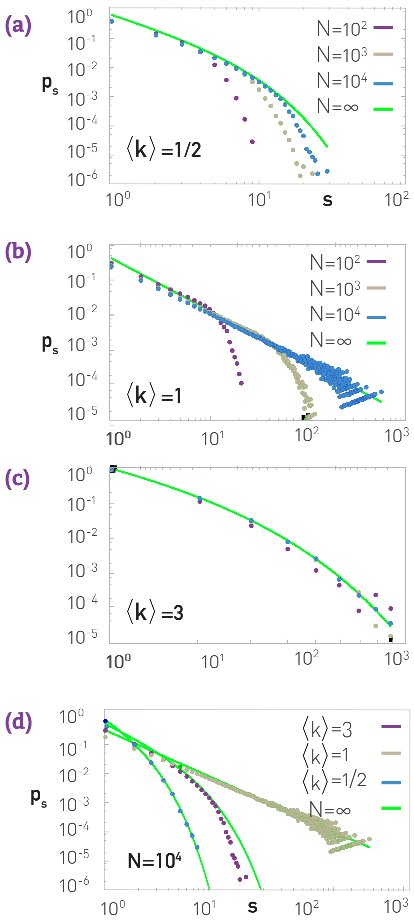


Figure 3.19
Component Size Distribution

Component size distribution p_s in a random network, excluding the giant component.

(a)-(c) p_s for different <*k*> values and *N*, indicating that p_s converges for large *N* to the prediction (3.34).

(d) p_s for *N* = 10^4, shown for different <*k*>. While for <*k*> < 1 and <*k*> > 1 the p_s distribution has an exponential form, right at the critical point <*k*> = 1 the distribution follows the power law (3.36). The continuous green lines correspond to (3.35). The first numerical study of the component size distribution in random networks was carried out in 1998 [34], preceding the exploding interest in complex networks.

Average Component Size

The calculations also indicate that the average component size (once again, excluding the giant component) follows [33]

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle + \langle k \rangle N_G / N} \tag{3.37}$$

For <*k*> < 1 we lack a giant component ($N_G = 0$), hence (3.37) becomes

$$\langle s \rangle = \frac{1}{1 - \langle k \rangle}\ , \tag{3.38}$$

which diverges when the average degree approaches the critical point <*k*> = 1. Therefore as we approach the critical point, the size of the clusters increases, signaling the emergence of the giant component at <*k*> = 1. Numerical simulations support these predictions for large *N* (Figure 3.20).

To determine the average component size for <*k*> > 1 using (3.37), we need to first calculate the size of the giant component. This can be done in a self-consistent manner, obtaining that the average cluster size decreases for <*k*> > 1, as most clusters are gradually absorbed by the giant component.

Note that (3.37) predicts the size of the component to which a randomly chosen node belongs. This is a biased measure, as the chance of belonging to a larger cluster is higher than the chance of belonging to a smaller one. The bias is linear in the cluster size *s*. If we correct for this bias, we obtain the average size of the small components that we would get if we were to inspect each cluster one by one and then measure their average size [33]

$$\langle s' \rangle = \frac{2}{2 - \langle k \rangle + \langle k \rangle N_G / N}\ . \tag{3.39}$$

Figure 3.20 offers numerical support for (3.39).

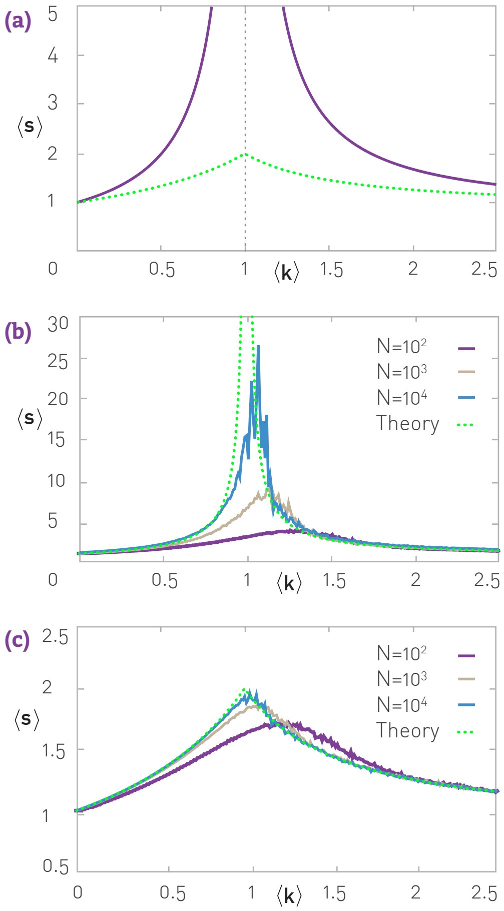


Figure 3.20

Average Component Size

(a) The average size <*s*> of a component to which a randomly chosen node belongs to as predicted by (3.39) (purple). The green curve shows the overall average size <*s*'> of a component as predicted by (3.37). (After [33]).

(b) The average cluster size in a random network. We choose a node and determined the size of the cluster it belongs to. This measure is biased, as each component of size *s* will be counted *s* times. The larger *N* becomes, the more closely the numerical data follows the prediction (3.37). As predicted, <*s*> diverges at the <*k*>=1 critical point, supporting the existence of a phase transition (ADVANCED TOPICS 3.F).

(c) The average cluster size in a random network, where we corrected for the bias in (b) by selecting each component only once.The larger *N* becomes, the more closely the numerical data follows the prediction (3.39).

ADVANCED TOPICS 3.E
FULLY CONNECTED REGIME

To determine the value of <*k*> at which most nodes became part of the giant component, we calculate the probability that a randomly selected node does not have a link to the giant component, which is $(1-p)^{N_G} \approx (1-p)^N$, as in this regime $N_G \simeq N$. The expected number of such isolated nodes is

$$I_N = N(1-p)^N = N\left(1-\frac{N\cdot p}{N}\right)^N \approx Ne^{-Np}, \qquad (3.40)$$

where we used $(1-\frac{x}{n})^n \approx e^{-x}$, an approximation valid for large n. If we make p sufficiently large, we arrive to the point where only one node is disconnected from the giant component. At this point $I_N = 1$, hence according to (3.40) p needs to satisfy $Ne^{-Np} = 1$. Consequently, the value of p at which we are about to enter the fully connected regime is

$$p = \frac{\ln N}{N}, \qquad (3.41)$$

which leads to (3.14) in terms of <*k*>.

ADVANCED TOPICS 3.F
PHASE TRANSITIONS

The emergence of the giant component at <*k*>=1 in the random network model is reminiscent of a *phase transition*, a much studied phenomenon in physics and chemistry [35]. Consider two examples:

i. **Water-Ice Transition** (Figure 3.21a): At high temperatures the H_2O molecules engage in a diffusive motion, forming small groups and then breaking apart to group up with other water molecules. If cooled, at 0˚C the molecules suddenly stop this diffusive dance, forming an ordered rigid ice crystal.

ii. **Magnetism** (Figure 3.21b): In ferromagnetic metals like iron at high temperatures the spins point in randomly chosen directions. Under some critical temperature T_c all atoms orient their spins in the same direction and the metal turns into a magnet.

The freezing of a liquid and the emergence of magnetization are examples of phase transitions, representing *transitions from disorder to order*. Indeed, relative to the perfect order of the crystalline ice, liquid water is rather disordered. Similarly, the randomly oriented spins in a ferromagnet take up the highly ordered common orientation under T_c.

Many properties of a system undergoing a phase transition are *universal*. This means that the same quantitative patterns are observed in a wide range of systems, from magma freezing into rock to a ceramic material turning into a superconductor. Furthermore, near the phase transition point, called the critical point, many quantities of interest follow power-laws.

The phenomena observed near the critical point <*k*> = 1 in a random network in many ways is similar to a phase transition:

- The similarity between Figure 3.7a and the magnetization diagram of Figure 3.21b is not accidental: they both show a transition from disorder to order. In random networks this corresponds to the emergence of a giant component when <*k*> exceeds <*k*> = 1.

- As we approach the freezing point, ice crystals of widely different sizes are observed, and so are domains of atoms with spins pointing in the same direction. The size distribution of the ice crystals or magnetic domains follows a power law. Similarly, while for <*k*> < 1 and <*k*> > 1 the cluster sizes follow an exponential distribution, right at the phase transition point p_s follows the power law (3.36), indicating the coexistence of components of widely different sizes.

- At the critical point the average size of the ice crystals or of the magnetic domains diverges, assuring that the whole system turns into a single frozen ice crystal or that all spins point in the same direction. Similarly in a random network the average cluster size <*s*> diverges as we approach <*k*> = 1 (Figure 3.20).

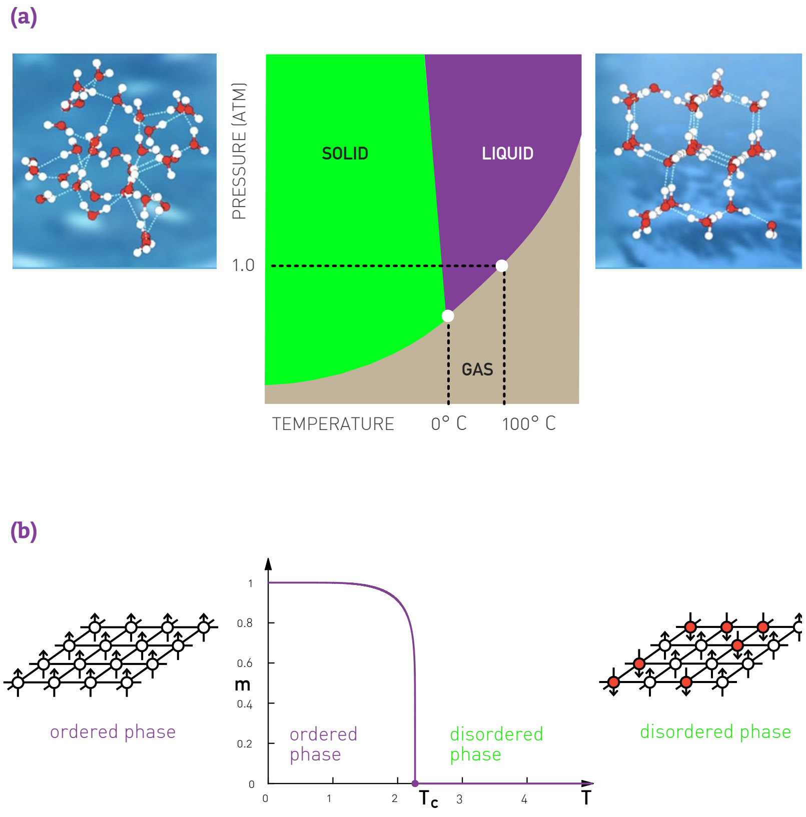


Figure 3.21

Phase Transitions

(a) Water-Ice Phase Transition

The hydrogen bonds that hold the water molecules together (dotted lines) are weak, constantly breaking up and re-forming, maintaining partially ordered local structures (left panel). The temperature-pressure phase diagram indicates (center panel) that by lowering the temperature, the water undergoes a phase transition, moving from a liquid (purple) to a frozen solid (green) phase. In the solid phase each water molecule binds rigidly to four other molecules, forming an ice lattice (right panel). After *http://www.lbl.gov/Science-Articles/Archive/sabl/2005/February/ water-solid.html*.

(b) Magnetic Phase Transition

In ferromagnetic materials the magnetic moments of the individual atoms (spins) can point in two different directions. At high temperatures they choose randomly their direction (right panel). In this *disordered state* the system's total magnetization ($m = \Delta M/N$, where ΔM is the number of up spins minus the number of down spins) is zero. The phase diagram (middle panel) indicates that by lowering the temperature *T*, the system undergoes a phase transition at $T = T_c$, when a nonzero magnetization emerges. Lowering *T* further allows *m* to converge to one. In this *ordered phase* all spins point in the same direction (left panel).

ADVANCED TOPICS 3.G
SMALL WORLD CORRECTIONS

Equation (3.18) offers only an approximation to the network diameter, valid for very large N and small d. Indeed, as soon as $<k>^d$ approaches the system size N the $<k>^d$ scaling must break down, as we do not have enough nodes to continue the $<k>^d$ expansion. Such finite size effects result in corrections to (3.18). For a random network with average degree $<k>$, the network diameter is better approximated by [36]

$$d_{\max} = \frac{\ln N}{\ln\langle k\rangle} + \frac{2\ln N}{\ln[-W(\langle k\rangle \exp -\langle k\rangle)]}, \tag{3.42}$$

where the Lambert W-function $W(z)$ is the principal inverse of $f(z) = z\ exp(z)$. The first term on the r.h.s is (3.18), while the second is the correction that depends on the average degree. The correction increases the diameter, accounting for the fact that when we approach the network's diameter the number of nodes must grow slower than $<k>$. The magnitude of the correction becomes more obvious if we consider the various limits of (3.42).

In the $<k> \to 1$ limit we can calculate the Lambert W-function, finding for the diameter [36]

$$d_{\max} = 3\frac{\ln N}{\ln\langle k\rangle}. \tag{3.43}$$

Hence in the moment when the giant component emerges the network diameter is three times our prediction (3.18). This is due to the fact that at the critical point $<k> = 1$ the network has a tree-like structure, consisting of long chains with hardly any loops, a configuration that increases $d_{\max}$.

In the $<k> \to \infty$ limit, corresponding to a very dense network, (3.42) becomes

$$d_{\max} = \frac{\ln N}{\ln\langle k\rangle} + \frac{2\ln N}{\langle k\rangle} + \ln N\left(\frac{\ln\langle k\rangle}{\langle k\rangle^2}\right). \tag{3.44}$$

Hence if $<k>$ increases, the second and the third terms vanish and the solution (3.42) converges to the result (3.18).