Small Worlds

The Dynamics of Networks between Order and Randomness

DUNCAN J. WATTS

PRINCETON UNIVERSITY PRESS

PRINCETON, NEW JERSEY

2

An Overview of the Small-World Phenomenon

2.1 SOCIAL NETWORKS AND THE SMALL WORLD

The "small-world phenomenon" has long been an object of fascination and anecdotal report. The experience of meeting a complete stranger with whom we have apparently little in common and finding unexpectedly that we share a mutual acquaintance is one with which most of us are familiar. More generally, most people have at least heard of the idea that any two people, selected randomly from almost anywhere on the planet, are "connected" via a chain of only a few intermediate acquaintances. Ouisa, a character in John Guare's play *Six Degrees of Separation* (1990), famously claims that

> 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.

"Six degrees" is now firmly embedded in folklore, embracing everyone from Kevin Bacon to Monica Lewinsky (Kirby and Sahre 1998). As a result, the idea can be hard to take seriously. And yet, whatever the precise number, it seems that human social systems really are constructed in a fashion quite unlike that of physical systems, in that they seem to violate what is known as *transitivity* of distances. In physical systems (which we generally visualise in no more than a three-dimensional space) all lengths between points, objects, or subsystems are related to each other by the triangle inequality which states that if three points (a, b, and c) are anywhere in the same space, then they can be connected via the three sides of a triangle, and the lengths of those sides must obey the inequality $d(a, c) \leq d(a, b) + d(b, c)$. It seems that this need not be

true of social systems, because it is quite possible for person A to be well acquainted with both person B and person C, yet for B and C to be not even remotely familiar with each other. This is a normal part of life, as each of us belongs not to a single group of acquaintances but to many, within each of which everyone pretty much knows everyone else but between which little interaction occurs. Common as it is, this feature of interpersonal relationships has much to do with everyone in the world's being somehow "close" to everyone else, no matter how far away intuition would suggest they are.

2.1.1 A Brief History of the Small World

Theoretical Work

Research specific to the small-world phenomenon did not commence until the 1960s with the formulation and initial mathematical investigation of the problem by Manfred Kochen and Ithiel de Sola Pool (Pool and Kochen 1978). These authors made substantive progress on the problem, estimating both the average number of acquaintances that people possess and the probability of two randomly selected members of a society being connected by a chain of acquaintances consisting of one or two intermediaries. They developed these approximations under a variety of assumptions about the level of social structure and stratification present in the population and concluded (speculatively) that even quite structured populations would have acquaintance chains whose characteristic path lengths are not much longer than those of completely unstructured populations (where the probability of A knowing C, given that A knows B, is independent of whether or not B knows C). For a population about that of the United States and an estimated average number of acquaintances per person of about a thousand, Pool and Kochen estimated that any member of the population could be connected to any other with a chain of associates consisting of at most two intermediaries (hence three degrees of separation).[1]

The study of distances in social networks, however, had begun over twenty-five years before the publication of Pool and Kochen's work, with Anatol Rapoport and his colleagues at the University of Chicago. In a series of papers in the 1950s and 1960s, published in the *Bulletin of Mathematical Biophysics*, Rapoport and colleagues established the theory of *random-biased nets*, which describes the statistics of a disease spreading through populations with varying degrees of structure. Ini-

tially, Solomonoff and Rapoport (1951) developed the idea of dispersion in a randomly connected network, in which every element was assumed to have the same number of connections (k). Based on these assumptions—the *independence* of the connections and the *regularity* of the nodes—they derived approximate formulae for the expected fraction (η) of the population to be reached eventually from a small, initial starting set. If a fraction $P(0) \ll 1$ of the population is infected initially, then $(1 - P(0))$ is initially *uninfected*. A consequence of the independence condition is that the disease spreads exponentially, infecting $P(t)$ new members at each time step, where

$$P(t) = \left(1 - \sum_{i=0}^{t-1} P(i)\right)\left(1 - e^{-kP(t-1)}\right). \tag{2.1}$$

In the limit of large t, Rapoport determined the following expression for the total infected fraction η:

$$\eta = 1 - (1 - P(0))e^{-k\eta}. \tag{2.2}$$

Several advances in realism have been made since upon this approximation, most notable of which are

1. The restriction to a *finite subpopulation* from which the k acquaintances can be chosen and a corresponding *strong overlap of friendship circles* (Rapport 1953a, 1953b).
2. The introduction of *structural biases*, specifically, *homophily* (the tendency to associate with people "like" yourself), *symmetry* of edges (which implies undirected instead of directed edges), and *triad closure* (the tendency of one's acquaintances to also be acquainted with each other; Foster et al. 1963).
3. *Social differentiation* of a population into heterogeneous subgroups (Skvoretz 1989).

In later work Rapoport (1957) accounted for the fact that, in a real network, each infected person may have *contact* with k others, but some of these will already have been infected, so that the *effective* number of connections per member t steps away from the origin of the disease is actually $\kappa(t) \leq k$, for all $t > 0$, and is no longer a constant. Fararo and Sunshine (1964) and Skvoretz (1985) argued for a constant $\kappa(t) = \kappa$, but one that accounts for structural biases, yielding the expression (from Skvoretz 1989)

$$P(t) = \left(1 - \sum_{i=0}^{t-1} P(i)\right)\left(1 - e^{-\kappa P(t-1)}\right), \tag{2.3}$$

where, in the special case of undirected ties,

$$\kappa = \begin{cases} \frac{(k-1-\zeta(k-1))(1-(1-\zeta^2)^k)}{k\zeta^2}, & \zeta > 0 \\ k-1, & \zeta = 0 \end{cases}$$

in which

$\zeta = \zeta_s(1 - qS)$ is the "triad-closure bias"
ζ_s = the triad-closure bias for "strong ties"
ζ_w = the triad-closure bias for "weak ties"
$S = (1 - (\zeta_w/\zeta_s))$ is a measure of the "strength of weak ties"
q = the probability of a connection being weak.

Note that the above expression requires a distinction between "strong" and "weak" ties, where the *strength* of a tie is determined not by some inherent feature of the tie itself, but by the structure of the surrounding network. Specifically, Granovetter (who introduced the idea) defined strength as follows:

> Consider, now, any two arbitrarily selected individuals—call them A and B—and the set, $S = C, D, E, \ldots$, of all persons with ties to either *or* both of them. The hypothesis which enables us to relate dyadic ties to larger structures is: the stronger the tie between A and B, the larger the proportion of individuals in S to whom they will *both* be tied, that is, connected by a weak or strong tie. This overlap in their friendship circles is predicted to be least when their tie is absent, most when it is strong and intermediate when it is weak. (1973, p. 1362)

In a later article (1983), Granovetter stresses that weak ties are, in fact, more significant in a social network than their strong counterparts:

> The argument asserts that our acquaintances ("weak ties") are less likely to be socially involved with one another than are our close friends ("strong ties")....
>
> The overall social structural picture suggested by this argument can be seen by considering the situation of some arbitrarily selected individual—call him or her "Ego." Ego will have a collection of close-knit friends, most of whom are in touch with one another—a densely knit "clump" of social structure. In addition, Ego will have a collection of acquaintances, few of whom know one another. Each of these acquaintances, however, is likely to have close friends in his or her own right and therefore to be enmeshed in a closely knit

> clump of social structure, but one different from Ego's. The weak tie between Ego and his or her acquaintance, therefore, becomes not merely a trivial acquaintance tie, but rather a crucial bridge between the two densely knit clumps of close friends. To the extent that the assertion of the previous paragraph is correct, these clumps would not, in fact, be connected to one another at all were it not for the existence of weak ties. (p. 203)

The probability (q) that a tie is weak corresponds to the likelihood that two connected vertices will have weakly overlapping friendship circles, and S quantifies how much less likely a weak tie is to complete a triad than a strong tie. Hence S is a measure of how significant a role the weak ties will play. The importance of the "strength of weak ties" idea was reinforced by Skvoretz and Fararo (Skvoretz 1989), who showed that "the stronger the weak ties in a population (in two senses, weak tie triads being less likely to close and weak ties being proportionately more frequent), the closer is a randomly chosen starter to all others."

Exactly how this works, and how important it is, is a major concern in this book. Closely related to the strength of ties and triadic closure is the idea of clustering, which has also been an issue of concern to researchers in social networks and has significant connections to the small-world problem. The idea of networks being divisible into cooperative subgroups that do not cooperate with each other was first formulated by Davis (1967), but the idea that neighbourhoods could be more or less densely connected was not quantified until slightly later by Barnes (1969), who defined *density* at a network element v as the proportion of all possible connections in v's immediate neighbourhood (defined by v and those elements to which v is connected directly) that actually exist. A very similar notion of density, termed *clustering*, is defined later in this chapter[2]. Barnes also discussed some qualitatively different networks for the parameters $n = 100, k = 10$, observing that they have different local densities and that the typical separation of network elements appears to increase with increasing density. Whilst this analysis touched on the idea that the local properties of a network (like density) can determine its global properties, the first systematic attempt to relate the two scales appears to have been the development of the concept of *structural eqivalance* and the technique of *block modelling*. According to Lorrain and White (1971), "a is structurally equivalent to b if a relates to every object x of [a category] C in exactly the same way as b does. From the point of view of the logic of the structure, then a and b are absolutely equivalent, they are substitutable."

Hence clustering and structural equivalence capture much of the same information about "who knows whom" at a local level, at least in the restricted case in which only one kind of social relation is considered. Block modelling (White et al. 1976) then considers networks as composed of *blocks* of structurally equivalent elements and represents the graph in terms of the relationships between these blocks. This is analogous to the amalgamation of Barnes's clusters connected by Granovetter's weak ties as a means of constructing a global view of the network whilst retaining some knowledge of the local structure. It is different from the work of the biased-net theorists because, instead of considering the characteristics of *pathways* though networks, it looks directly at the *knittedness* of networks at different scales.

A final area of research in social networks that relates closely to the work presented here is that concerning the dimension and geometry of the space in which social networks are presumed to exist. Many approaches to this question have been devised, but almost all of them fall under the rubric of *multidimensional scaling*. This term refers to a loose bundle of techniques developed by many researchers across several disciplines and decades, but all are based on more or less the same idea, which is basically the following.

A population is assumed to exist in some finite (but possibly large) dimensional "social space," where the coordinates $(x_1, x_2, \ldots, x_m)$ of each member represent quantitative measures of a set of characteristics, which are presumably sufficient to identify each member of the population uniquely. These coordinates, however, are unknown to the observer, as is the dimension of the space. What the observer *does* know is the *distances* $(\delta_{i,j})$ between each pair (i, j) of members, where distance is defined in some manner that is problem-specific but is often related to frequency of interaction or some assessment of similarity, generated by either the observer or the members themselves. The problem then is to reconstruct the space by choosing coordinates in a self-consistent manner such that the known distances are related to the coordinates by a particular choice of metric. Basically, it is at this point that the methods start to differ.

The group of methods generally known as *metric methods* (Chapter 4 of Davidson 1983) utilise a standard Euclidean metric, hence

$$\delta_{i,j} = \left[\sum_d (x_{id} - x_{jd})^2\right]^{\frac{1}{2}}, \tag{2.4}$$

whereas an alternate group of methods known as *nonmetric methods* (Chapter 5 of Davidson 1983) utilise a variant of Equation 2.4:

$$\delta_{i,j} = f\left(\left[\sum_d (x_{id} - x_{jd})^2\right]^{\frac{1}{2}}\right), \tag{2.5}$$

where f is some monotone function (that is, $d_{ij} < d_{i'j'} \Rightarrow f(d_{ij}) < f(d_{i'j'})$). In fact, "nonmetric" is something of a misnomer, because Equation 2.5 is every bit as much a metric as Equation 2.4; it is just not the Euclidean metric. It turns out that this confusion between nonmetric and non-Euclidean is widespread in this particular part of the study of social networks. For instance, Barnett (1989) claims that non-Euclidean geometry is necessary to describe social networks precisely. Pool and Kochen also believed this, asserting that the transitivity of a Euclidean space is violated in social networks where Person A may be very close to both B and C and therefore likely to know them both, but B and C may be very far from each other (Pool and Kochen 1978). The conclusion is that if this basic tenet of a Euclidean space is violated, then, necessarily, the space in which social networks exist must be non-Euclidean. This is actually a misunderstanding about which more will be said in Section 2.1.2.

For the moment, it is important to realise only that multidimensional scaling is simply a process designed to reconstruct the space in which the system is presumed to exist, thus both generating a set of meaningful coordinates, with which to distinguish population members, and providing a visual representation of the data that enables the observer to gain more insight into the relationships between members than would be possible by staring at a large matrix of numbers. Therefore, in applying to the data whichever particular algorithm is chosen, it is to be hoped that only a few dimensions will be sufficient to embed the data within an acceptable degree of error. Obviously, if $\Delta_{i,j}$ is an $n \times n$ matrix containing the intermember distance information, it is always possible to embed the network in an n-dimensional space. However, for $n > 3$, this isn't going to be much help in visualising the data relationships expressed in the matrix, and from another perspective, there is almost no point in doing it at all (for any n) because the resulting relationships between the coordinates will be no less impenetrable than the original distance matrix. Hence there is a substantial trade-off between the goodness of fit of the embedding and its dimension, which should almost always be kept to less than four. This then raises another thicket of issues for the data analyst, but these are not relevant here. See Davidson (1983) and R. N.

Shepard and Nerlove (1972) for an overview of these obstacles and the methods attempted to surmount them.

The theory of social networks, then, has proceeded along four basically distinct, but interrelated, strands:

1. The statistical analysis of pathways through networks with varying degrees of local structure.
2. The qualitative description of the structure of networks in terms of local (e.g., clustering) and nonlocal (e.g., weak ties) features.
3. The renormalisation of networks, viewed as meta-networks of highly clustered or equivalent subnetworks.
4. The embedding of networks into (hopefully low-dimensional) spaces where the coordinates are readily interpretable and relationships between members can be more easily visualised.

In tandem with (and frequently driving) this theoretical development has been the development of empirical techniques that attempt to probe the structure of real social networks directly. Once again, the small world did not start turning in this field of endeavour until the late 1960s.

Empirical Work

The first empirical work was conducted at about the same time as Pool and Kochen were developing their theoretical ideas, by the psychologist Stanley Milgram. Although principally renowned for his remarkable and disturbing work on the apparent submission of human ethical values to authority (Milgram 1969), Milgram also conducted a highly innovative test of the small-world hypothesis (Milgram 1967). In this experiment, Milgram sent a number of packets to agreeable "sources" in Nebraska and Kansas, with instructions to deliver these packets to one of two specific "target" persons in Massachusetts. The targets were named and described in terms of approximate location, profession, and demography, but the sources were only allowed to send the packets directly to someone they knew by first name. The object was to get the packets from source to target with as few of these "first-name-basis links" as possible. Hence each link in the chain was required to think hard about which of their acquaintances would be most likely to know the target person or at least be "closer" to them: demographically, geographically, personally, or professionally. Also, each link was supposed to record, in the packet, details about themselves corresponding to those provided about the target, enabling the experimenters to track the progress of the packet and the demographic nature of chain along which it passed.

The upshot of all this was that Milgram determined that a median of about five intermediaries was all that was required to get such a letter across the intervening expanse of geography and society. Whether this number is, in reality, too low or too high is a matter of debate. On the one hand it would seem unlikely that, at every step, the sender would pick the optimal person to send it to (and they could only pick one), and that this effect would tend to make the chains longer than they needed to be. On the other hand, many of the chains were never completed because of apathy on the part of the participants, and, as longer chains are more likely to terminate than shorter ones, the result might well have been systematically biased in favour of lower numbers. White (1970) proposed a model to account for this effect, which yielded a revised estimate of about seven intermediaries. In any event, Milgram seemed to have demonstrated that whatever the precise number was, it wasn't very big, compared with the overall magnitude of the system (on the order of the population of the United States, which was about 200 million in 1967). A second study by Milgram (Korte and Milgram 1970) used essentially the same method to examine the length of acquaintance chains between whites in Los Angeles and a mixed white-black target population in New York and found similar statistics.

Of course, the study of social networks and their use as a tool for examining the structure of societies already had a considerable history by the time Milgram did his initial experiment (see Mitchell 1969 for a review of the field at the end of the 1960s), but none of this work had looked at the question of *path length* in the same light as had Milgram. It also seems that very little work of this nature and scale has occurred since, even though Milgram's results did (and still do) spur considerable interest. Perhaps the work of Rapoport (Foster et al. 1963) is closest to this, in that he measures the average fraction of a population of students in a junior high school that is reached as a function of number of intermediaries. Even here, though, the system involved is much smaller, and the emphasis is upon justifying parameters for a network model rather than direct experimental verification of the small-world phenomenon.

In fact, it seems that more empirical effort has been devoted to the lower-level question (originally posed by Pool and Kochen 1978) of the number of acquaintances that the typical person possesses. Efforts in this department have been made by Freeman and Thompson (1989), who use a variant of Pool's original "telephone book" method,[3] and Bernard et al. (1989), who use the 1985 Mexico City earthquake victims as a

sample subpopulation to determine the acquaintance volume of residents of Mexico City. This turns out to be a difficult exercise, and it seems unlikely that even if such a number and its variance could be convincingly determined for any given definition of acquaintance, that it would play nearly so as important a role in the understanding of networks as a comparable advance in the understanding of network structure.

2.1.2 Difficulties with the Real World

Theoretical

Although the researchers surveyed in Section 2.1.1 did make significant gains on the issue of the effective size of social networks, their progress was hampered by a number of difficulties that arose from both the questions they chose to ask and the methodologies they used to seek answers. The results of Pool and Kochen are highly suggestive of the small-world property's holding true in real societies. But although their results are not highly sensitive to the estimation of the average number of acquaintances,[4] they *are* highly sensitive both to the assumptions about conditional probability of acquaintanceship and to the large-scale structure of the population, which may dictate different rules of conditional probability in different parts of the population. A more recent article by Kochen (1989b) reports little progress on this essential theoretical difficulty. It turns out that this is a problem faced by all theoreticians who find themselves exploring systems that operate in the intermediate regime between order and randomness. The problem arises in many fields, notably fluid mechanics and the dynamics of coupled, nonlinear oscillators (see Chapter 9), but in terms of social networks, the only networks whose statistical properties are analytically tractable are those that are either (1) completely ordered (for instance, a d-dimensional, hypercubic lattice), or (2) completely random (such as Rapoport's random webs).

Although these cases are at opposite extremes of the structural spectrum, they both share the essential characteristic that their local structure mirrors (either exactly or statistically) their global structure, and hence analysis based on strictly local knowledge is sufficient to capture the statistics of the entire network. That is, in an important sense, they "look" the same everywhere.

Unfortunately, real social systems appear to be firmly in between these extremes, and, to make matters worse, it is not even known where on

the spectrum they lie. What does seem to be true is that, if any theoretical explanation is to capture the important features of social networks, then it must find some way of encapsulating elements of both order and randomness, thus accounting for the appearance of structure at different scales. Much of the work surveyed above has grappled with this problem in creative and insightful ways, but three central issues appear to remain open:

1. Social networks exhibit structural characteristics that are inherently *nonlocal* (Granovetter's "bridges"), and so no purely local analysis can predict their global statistical features.
2. Analytical difficulties increase with the size of the network, and almost none of the work has been tested for large population size (n) with sparse connectivity under any but the most restrictive conditions.
3. It is unknown where on the structural spectrum real social networks lie, but no treatment has been given to the properties of *continuous families of networks*, whose structural properties vary all the way from one extreme to the other, with the intention of determining the location and nature of any *transitions* that occur in between.

Adding to the confusion is the difficulty of determining which kind of space a network exists in and the appropriate metric with which to measure lengths. The root of this difficulty appears to be that networks are frequently defined in the sociological literature on the basis of (at least) two relations: (1) how "far" each pair of vertices is from each other in the (unknown) metric of the (unknown) "social space," and (2) whether or not they are connected and (perhaps) how strongly.

The first relation turns out to be the problematic one because if one takes any single measure of "social distance," such as frequency of interaction, overlap of interests, or common characteristics, ambiguities inevitably arise, and the resulting "distances" will appear to violate the triangle inequality. It is false, however, to declare the corresponding space non-Euclidean. In fact, the violation of the triangle inequality (if it isn't just due to faulty data) is symptomatic of a far more general breakdown in the geometry of a space, because it violates one of the fundamental notions not just of Euclidean distance, but of distance itself. The reason for this is that the triangle inequality is one of four basic properties of a class of spaces known as *metric spaces* (Munkres 1975). This is an extremely general class of topological space that formalizes the idea of distance (that is, a *metric*) and that includes *any* sensible notion of

distance. Hence if the measured "distances" in some network are not consistent with the triangle inequality, then either (1) the criteria used to measure distance are mistaken (the data are somehow incomplete or in error), or (2) the space is not a metric space, and so the concept of distance is meaningless in the first place.

In either case the result is that not much can be done to interpret the measured distances as meaningful without performing some arbitrary (and probably meaningless) transformation on them (such as adding a large constant to all distances or taking logarithms) until they *do* satisfy the requirements of a legitimate metric. Such manipulations, however, do not so much help the data as reveal their inherent flaws and suggest either a new method of measuring social distance or a different approach to the problem altogether.

The most likely source of the problem is the difficulty inherent in measuring how close or far apart people are, not in a network, but in the more general sense implied by the idea of a social space. As soon as one tries to grapple with this issue, it immediately reveals itself to be slippery both theoretically (what does "social distance" mean and what are its most important contributors?) and empirically (even if you knew what it was, how would you go about measuring it?). Even in the best-case scenario, it seems that whatever metric of length is chosen, it will almost certainly not capture all the features that are relevant to relationships between people. It is just as likely, however, that such a measure could even be *multivalued*, as social distance is at least partly a matter of perspective.

In the case of a network, the issue becomes still murkier, because distance can also be defined in terms of the network connections themselves, which may be a function of the underlying space but almost certainly not one that is known. If the network distance and the metric distance do not agree, then the analyst is once again faced with a choice between arbitrary manipulations of the data and ignoring them outright. Because the methodological basis of measuring distances in the network sense, solely in terms of who is connected to whom, rests on much firmer ground, both theoretically and empirically, network distance will be treated here as the sole measure of distance, at which point all talk of either non-Euclidean or nonmetric spaces instantly disappears. A network does not necessarily exist in any particular space at all, but as all network distances must certainly conform to the triangle inequality, then (if one insists on thinking in terms of Euclidean spaces) an embedding is guaranteed by an algorithm that is described briefly in Section 2.2.3.

Empirical

On the empirical side of the same problem, the principal stumbling block seems to have been the practical difficulties associated with obtaining and representing sufficiently detailed relationship data for large, sparse networks. Milgram's methodology (sometimes referred to as "the small-world method") was imaginative and original and did serve to illuminate some interesting characteristics of what might be termed a "random-biased walk" in a network: random, because senders did not have sufficient information to know which of their acquaintances was optimal, and biased, because some attempt at optimality was made. Unfortunately it is difficult to generalise results such as these beyond the scope of the specific study, and so it is hard to tell much about the overall qualitative structure of the network.

Conversely, attempts to reconstruct network connectivity in a broader sense have been forced to concentrate on small systems in which it is practicable to map every single connection (see, for example, Doreian 1974), at which point there is little of the intermediate ground between the local and the global scales in which the interesting small-world phenomena occur.[5]

Finally, the methods used to assess actual network parameters, such as the average number of friends per person, have revealed that deep problems exist with any attempt to estimate this kind of data:

1. Most people seem to be quite poor at estimating their number of friends reliably.
2. Methodological tricks to circumvent this difficulty (such as requiring subjects to keep written records of all interpersonal encounters over an extended period of time) are time and labour intensive.
3. The number (however it is estimated) changes over time.
4. The number is highly sensitive to the definition of a "meaningful" or "substantive" contact or relationship.

Of all these points, perhaps the last is the most damning because it threatens the validity of any result that does not perform at least several identical surveys of acquaintanceship volume using different definitions of what constitutes acquaintanceship in each (for instance, first-name basis versus propensity to lend money). This is a similar objection to that raised earlier concerning the definition of "social distance": acquaintanceship, like distance, can vary widely depending on any one or all of the following: (1) the biases of the observer, (2) the question being posed, and (3) the members of the network in question.

2.1.3 Reframing the Question to Consider *All* Worlds

Given the difficulties inherent to empirical investigations of the small-world phenomenon (and the structure of social networks in general), theoretical investigation seems attractive, if only as a means of focusing future questions for empiricists to answer. Theoretical approaches, however, also seem to have some serious limitations in the regime of interest, at least if one insists upon analytical solutions. What is needed, then, is a new theoretical approach that attempts to exploit the generality of theory without falling prey either to overly restrictive idealisation of network structure ("throwing the baby out with the bathwater") or to impenetrable thickets of numerical solutions in a forest of arbitrary parameters. The motivation behind the work presented here is to chart just such a course by ignoring much of the sociological detail inherent in previous models and considering a much more general statement of the problem:

> Assuming that a network can be represented by nothing more than the connections existing between its members and treating all such connections as equal and symmetric, a broad class of networks can be defined, ranging from highly ordered to highly random. The question then is *Does the Small-World Phenomenon arise at some point in the transition from order to disorder, and if so, what is responsible for it?* In other words, *What is the most general set of characteristics that guarantee that a system will exhibit the small-world phenomenon, and can those characteristics be specified in a manner independent of the model used to create the network*?

As yet, the small-world phenomenon has not been defined precisely in terms of which specific properties a network must possess in order to exhibit it. This will be deferred until after some exploration of different network topologies yields the kind of intuition that will be needed to motivate the appropriate definition. Even so, it should be apparent that if the goal stated above can be achieved, a great deal can be said about the existence of the small-world phenomenon, in what systems it is likely to arise, and in what sort of applications it might be useful. Before continuing, however, some basic terminology and results are required from the theory of graphs that will help to describe and define the networks and properties of interest.

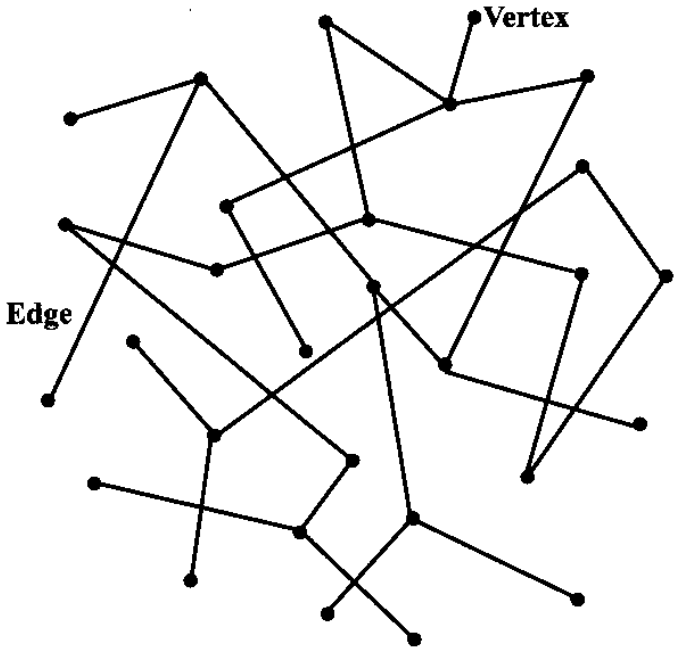

Figure 2.1 A general graph.

2.2 BACKGROUND ON THE THEORY OF GRAPHS

2.2.1 Basic Definitions

A graph, in its most basic sense, is nothing more than a set of points connected in some fashion by a set of lines (see Fig. 2.1). The following definition of a graph is taken from Wilson and Watkins (1990).

Definition 2.2.1. A *graph* G consists of a nonempty set of elements, called *vertices,* and a list of unordered pairs of these elements, called *edges*. The set of vertices of the graph G is called the *vertex set* of G, denoted by $V(G)$, and the list of edges is called the *edge list* of G, denoted by $E(G)$. If v and w are vertices of G, then an edge of the form vw is said to *join* or *connect* v and w.

The number of vertices in $V(G)$ is termed the *order* of the graph (n), and the number of edges in $E(G)$ is termed its *size* (M). Graphs can be used to represent all kinds of networks, where the vertices represent network elements (such as people, animals, computer terminals, organisations, cities, countries, production facilities) and the edges represent some predefined relationship between connected elements (such as friendship, prey-predator relationship, Ethernet connection, business

alliance, highway, diplomatic relationship, product flow). Clearly both elements and connections may embody any number of characteristics, but the theory of graphs generally deals only with the number of elements in the network and their relationships with respect to each other in terms of the characteristics of the edge set.

This very broad definition is capable of representing systems in bewildering detail. The only graphs that will be considered here, however, are those that conform to the following restrictions:

1. *Undirected*. Edges exhibit no inherent direction, implying that any relationship so represented is symmetric.
2. *Unweighted*. Edges are not assigned any a priori strengths. Hence any importance that specific edges may later assume derives solely from their relationship with other edges.
3. *Simple*. Multiple edges between the same pair of vertices or edges connecting a vertex to itself are forbidden.
4. *Sparse*. For an undirected graph, the maximal size (M) of $E(G) = \binom{n}{2} = n(n-1)/2$, corresponding to a "fully connected" or *complete* graph. Sparseness implies $M \ll n(n-1)/2$.
5. *Connected*. Any vertex can be reached from any other vertex by traversing a path consisting of only a finite number of edges.

These assumptions obviously compromise the ability of the resulting models to form realistic representations of many networks. Relationships are often directed (child-parent, teacher-student, and so on), and some are clearly more important than others; many real networks are not connected, and often multiple types of relationships exist between the same set of elements (like business and friendship ties). However, aside from simplifying the resulting analysis a great deal, these assumptions do form a natural starting point for modelling networks in that they introduce a minimum amount of arbitrary structure whilst still allowing meaningful questions to be asked of the network as a whole.

Although graphs can be represented pictorially, most computations of graph properties are accomplished by way of either an *adjacency matrix* or *adjacency list*. The *adjacency matrix* $\mathbf{M}(G)$ is the $n \times n$ matrix in which $M_{i,j}$ is the number of edges joining the vertices i and j. In the unweighted case, all entries would be either 0 or 1. The *adjacency list* simply lists all vertices of the graph and, next to each, the vertices with which it is adjacent. The number of edges incident with a given vertex v (that is, the size of v's adjacency list) is called the *degree* of v, denoted k_v. One statistic that will be referred to frequently is the *average degree* of

the graph, k. Hence, for undirected graphs, k quantifies the relationship between n and M ($M = (n \cdot k)/2$). The corresponding effect on k of the sparseness condition above is that all graphs must have $k \ll n$. A graph in which all vertices have precisely the same degree k is called *k-regular* or just *regular*.

2.2.2 Length and Length Scaling

One of the most important statistics of graphs to be considered here is the *characteristic path length* ($L(G)$), that is, the typical distance $d(i, j)$ between every vertex and every other vertex. "Distance" here refers not to any separately defined metric space in which the graph has been embedded, but to a distinct graph metric—simply the minimum number of edges (in the edge set) that must be traversed in order to reach vertex j from vertex i, or in other words the *shortest path length* between i and j. Investigations of this graph invariant have a long history, spanning several subject areas and utilising a number of approaches. As long ago as 1947, Wiener (1947) investigated the sum of all distances between all pairs of vertices in a graph (sometimes called the "Wiener index") in connection with the boiling point of paraffin, where the vertices of the graph were to represent atoms and the edges, intramolecular bonds. Since then both the sum of all distances in a graph as well as the average distance across all pairs of vertices have appeared as parameters relevant to social status in a hierarchy (Harary 1959), architectural floor plans (March and Steadman 1971), the performance of computer networks (Frank and Chou 1972) and telecommunication networks (Lin 1982; Pippenger 1982; Chung 1986), and the physical properties of as yet unsynthesised hydrocarbons (Rouvray 1986).

Throughout all this the problem of finding a closed-form expression for the characteristic path length of a general, connected graph[6] has remained impregnable, and researchers have had to satisfy themselves with the explication of upper and lower bounds upon the quantity for various classes of graphs. Cerf et al. (1974) determined a lower bound on the average $d(i, j)$ in a k-regular graph by assuming a perfectly expanding graph. That is, starting from any vertex, k vertices can be reached at distance 1, then from each of these vertices another $(k - 1)$ new vertices can be reached at distance 2, and so on, without any redundancies, until the entire graph has been reached. This type of graph, known as a Moore Graph, is the most *efficient* possible k-regular graph, in the sense that every vertex "reaches" k new vertices, but it has since been proven

unrealisable except in a handful of special cases in which it is possible to close the graph upon itself with no redundancies (Chung 1986). An important result of studying Moore Graphs as a theoretical lower bound (even if unattainable) is that, for $k > 2$, the characteristic path length in any *regular* graph must grow at least logarithmically with n. We will see later that random graphs are good approximations to this lower bound.

Entringer et al. (1976) showed that the sum of all distances in any graph must lie between that of a complete graph and that of a one-dimensional *chain,* where each vertex has $k = 2$. Doyle and Graver (1977) later showed that a cycle—a chain with its ends connected—has the maximal characteristic path length of any graph with periodic boundaries. Whilst this result does not necessarily extend to higher k (where vertices are connected to nearest neighbours, next-nearest neighbours, and so on), it suggests that cycles with larger k have close to the maximum possible characteristic path length for a given n and k. It also suggests that the cycle is a particularly interesting object because it is, at once, the most and least efficient 2-regular structure. In fact, it is the *only* 2-regular structure and has the additional property that it is also 2-connected, which is to say that the deletion of any two edges will disconnect the graph. Hence it is the only minimally connected, regular graph topology—a fact that will be useful in Chapter 3.

Following this work, tighter bounds on the average distance or sum of all distances have been determined for specific classes of graphs (Buckley and Superville 1981), digraphs (Plesnik 1984), trees (Winkler 1990; Entringer et al. 1994), and random graphs (Schneck et al. (1997). The greatest problem with these attempts is that they either impose very loose bounds on the quantity of interest or else require strong constraints on the class of graphs to which the bounds apply. In either case, the results do little to assist the task of actually determining a characteristic path length for an arbitrary graph.

Recently some new approaches have been developed by (amongst others) Chung (1988, 1989, 1994) and Mohar (1991), which place bounds on the characteristic path length without restricting the variety of eligible graphs. Unfortunately these bounds are necessarily expressed in terms of other graph invariants that are virtually as inaccessible as the characteristic path length itself.[7] Interesting though such relationships between graph invariants are, they do not really help much if the primary aim is specifically that of computing or estimating length. They also suggest that analytical formulae for the length characteristics of graphs are, in

general, hard to come by. Hence a heavy reliance on numerical results seems appropriate.

At this point, it might seem that the obvious and natural definition of characteristic path length would be $d(i, j)$ averaged over all $\binom{n}{2}$ pairs of vertices and that this is best computed numerically for a known graph. Unfortunately, for large n, this becomes impractical to compute exactly, so a random sampling technique is needed to estimate the length to within a prescribed accuracy. Using such a sampling technique, it turns out that it is significantly easier to estimate the *median* shortest path length than it is the mean. As the mean and the median are practically identical for any reasonably symmetric distribution, then the sampling efficiency of the median seems to mark it as the most appropriate measure of length in a graph. However, the median suffers from a different drawback, which is that it is integer-valued. As the *scaling properties* of length with respect to increasing n are also of interest, and as the characteristic path length of some graphs remain on the same order of magnitude over several orders of magnitude in n, then an integer-valued length cannot provide sufficiently detailed information. A reasonable compromise, which incorporates most of the sampling convenience of the median, with the real-valued advantage of the mean is the following.

Definition 2.2.2. The *characteristic path length* (L) of a graph is the *median* of the *means* of the *shortest path lengths* connecting each vertex $v \in V(G)$ to all other vertices. That is, calculate $d(v, j)\ \forall j \in V(G)$ and find $\bar{d}_v$ for each v. Then define L as the median of $\{\bar{d}_v\}$.

As mentioned above, for large n a random sampling technique is used that is due to Huber (1996). According to this method, $\bar{d}_v$ is calculated for a randomly selected subset of s vertices, where s is determined as follows:

> Finding an approximate median through sampling is relatively straightforward. First, take s samples, then find the median of the samples. More generally, call $M_{(q)}$ a *q-median* if at least qn of the numbers in the set (n) are less than or equal to $M_{(q)}$ and at least $(1 - q)n$ of the numbers are greater than or equal to $M_{(q)}$.
>
> Call $L_{(q,\delta)}$ a (q, δ)-*median* if at least $qn(1 - \delta)$ numbers in the set are less than or equal to $L_{(q,\delta)}$ and at least $(1 - q)n(1 - \delta)$ of the numbers are greater than $L_{(q,\delta)}$. Equivalently, $L_{(q,\delta)} = M_{(q')}$ for some q' that satisfies $(1 - \delta)q \leq q' \leq (1 + \delta)q$.
>
> Finding such a value $L_{(q,\delta)}$ that is correct with high probability is much faster than finding M_q which takes linear time. To find a value

for $L_{(q,\delta)}$, take s samples and look at the q-median of the sample (p. 2).

Theorem 2.2.1. *The above algorithm yields a correct value for* $L_{(q,\delta)}$ *with probability* $1-\varepsilon$ *if* s *samples are taken, where* $s = (2/q^2)\ln(2/\varepsilon)1/(\delta')^2$ *and* $\delta' = 1/(1-\delta) - 1 = \delta/(1-\delta)$. *Note that when* δ *is small,* $\delta \simeq \delta'$ (Huber 1996).

The computation required by Definition 2.2.2 is less efficient than that actually proposed by Huber, which samples only s pairs of vertices instead of s complete search trees. However, the difference in computational time is only a constant factor and so is a reasonable sacrifice to make for the utility of a real-valued measure of length.

Having established either an exact (convenient in practice only for $n \lesssim 1{,}000$) or approximate value of the characteristic path length, the question arises: how does L *scale* with respect to changes in n and k? This question is important because the *scaling* of L is more indicative of the *qualitative structure* (or *topology*) of a graph than the specific value of L itself. Precisely what is meant by maintaining the qualitative structure of a graph whilst changing n and k will become more apparent in Chapter 3 in terms of one-parameter families of graphs (parameterised by some parameter p) that interpolate between order and randomness. The point is that different values of the parameter p represent different qualitative structures and that graphs of different n and k, but with the same p value, are qualitatively the same. This leads to the following definitions.

Definition 2.2.3. For a fixed p, the *length scaling with respect to* n *of* $G(p)$ is

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \left(\frac{L(G(p;n_1,k))}{L(G(p;n_2,k))} \right)$$

for $n_1 > n_2$ and $1 \ll k \ll n_1, n_2$. T is said to exhibit *d-scaling with respect to* n if

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \left(\frac{L(G(p;n_1,k))}{L(G(p;n_2,k))} \right) = \frac{n_1^{\frac{1}{d}}}{n_2^{\frac{1}{d}}}.$$

Definition 2.2.4. For a fixed p, the *length scaling with respect to* k *of* $G(p)$ is

$$\lim_{n\to\infty} \left(\frac{L(G(p;n,k_1))}{L(G(p;n,k_2))} \right)$$

for $k_1 > k_2$ and $1 \ll k_1, k_2 \ll n$.

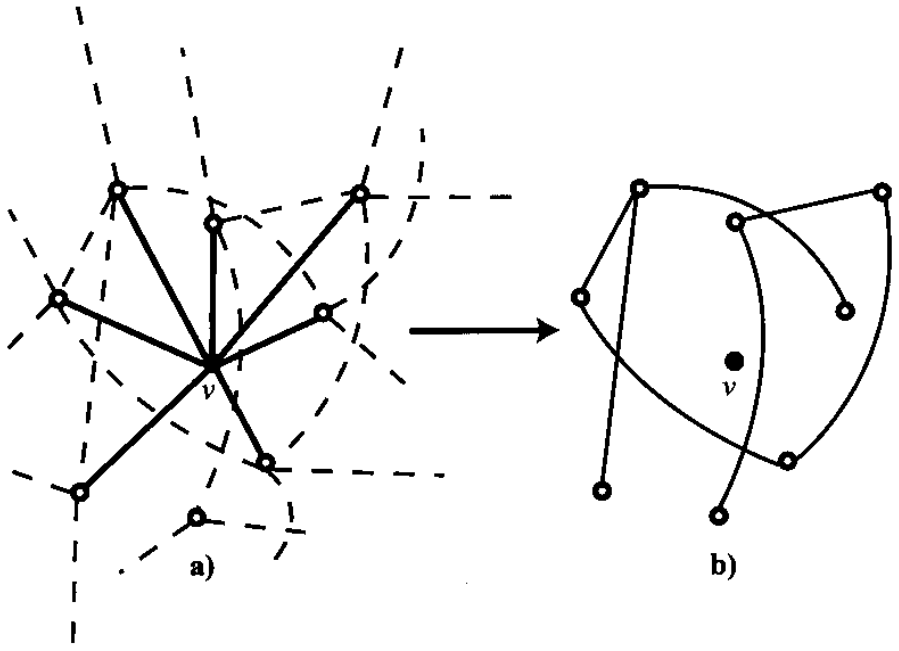

Figure 2.2 The neighbourhood of a vertex v (a) with vertex v included and (b) showing only the edges between vertices in $\Gamma(v)$.

Hence the parameter p defines an infinite set of graphs, each of which exhibits certain structural characteristics common to the set. The characteristic path length of a member of such a set can vary over $1 \leq L \leq \infty$, but the scaling property is invariant across the entire set. This is why the idea of scaling is so useful in characterising the qualities of a given graph: one can compute exactly the properties of small graphs with a given p and then, knowing their scaling properties, also obtain knowledge of their much larger cousins whose properties cannot be computed directly.

2.2.3 Neighbourhoods and Distribution Sequences

One recurrent theme throughout this book is the metaphor of information "spreading" from a single vertex throughout the graph. In connected graphs, there is no issue of whether or not the entire graph will be reached, but only the number of "steps" required to achieve this. The notion of a step is captured in terms of the *neighbourhood* either of a vertex (see Fig. 2.2) or a connected subgraph.

Definition 2.2.5. The *neighbourhood* $\Gamma(v)$ of a vertex v is the subgraph that consists of the vertices adjacent to v (not including v itself).

TABLE 2.1
Distribution Sequence for Kevin Bacon in the "Kevin Bacon Graph" (April 1997)

j (Bacon number)	$\lvert\Gamma_j\rvert$	Λ_j
0	1	1
1	1,181	1,182
2	71,397	72,579
3	124,975	197,554
4	25,665	223,219
5	1,787	225,006
6	196	225,202
7	22	225,224
8	2	225,226

Definition 2.2.6. The neighbourhood $\Gamma(S)$ of a connected subgraph S is the subgraph that consists of all vertices adjacent to any of the vertices in S, but not including the vertices of S.

Definition 2.2.7. In the special case where $S = \Gamma(v)$, $\Gamma(S) = \Gamma(\Gamma(v)) = \Gamma^2(v)$. More generally $\Gamma(\Gamma^{i-1}(v)) = \Gamma^i(v)$, the ith *neighbourhood of* v. Hence $\Gamma^0(v) = \{v\}$.

Definition 2.2.8. The sequence $\Lambda_j(v) = \sum_{i=0}^{j} |\Gamma^i(v)|$ for $0 \leq j \leq j_{\max}$ is the *distribution sequence for* v, where $\Lambda_{j_{\max}}(v) = |G|$.

Definition 2.2.9. $\Lambda_j = \overline{\Lambda_j(v)}$ over all $v \in V(G)$ is the *distribution sequence* for G.

It follows immediately from these definitions that $\max_v(j_{\max}(v)) = D$, the *diameter* of the graph. The functional form of Λ_j is indicative of the rate at which information "spreads" throughout a graph (think of a signal spreading from vertex to vertex along the edges, where all edges take equal "time") and hence the structure of the graph itself. Table 2.1 gives a real example of a distribution sequence for none other than the illustrious Kevin Bacon, where j is the Bacon Number, and Λ_j is the number of actors and actresses who have a Bacon Number of j or less.

2.2.4 Clustering

The idea of a neighbourhood is also useful in quantifying another statistic that will be of interest in this work, namely, the *clustering coefficient* of a graph.

Definition 2.2.10. The clustering coefficient γ_v of Γ_v characterises the extent to which vertices adjacent to any vertex v are adjacent to each other. More precisely,

$$\gamma_v = \frac{|E(\Gamma_v)|}{\binom{k_v}{2}},$$

where $|E(\Gamma_v)|$ is the number of edges in the neighbourhood of v and $\binom{k_v}{2}$ is the total number of *possible* edges in Γ_v.

That is, given k_v vertices in the subgraph Γ_v, at most $\binom{k_v}{2}$ edges can be constructed in that subgraph. Hence γ_v is simply the net fraction of those possible edges that actually occur in the real Γ_v. In terms of a social-network analogy, γ_v is the degree to which a person's acquaintances are acquainted with each other and so measures the *cliquishness* of v's friendship network. Equivalently, γ_v is the probability that two vertices in $\Gamma(v)$ will be connected. A measure of clustering over the entire graph is then in following.

Definition 2.2.11. The *clustering coefficient* of G is $\gamma = \gamma_v$ averaged over all $v \in V(G)$. Hence $\gamma = 1$ would imply that the corresponding graph consisted of $n/(k+1)$ disconnected, but individually complete, subgraphs (cliques), and $\gamma = 0$ would imply that no neighbour of *any* vertex v is adjacent with any other neighbour of v.

2.2.5 "Lattice Graphs" and Random Graphs

There are many other graph statistics that could (and probably should) be measured. But already it is possible to make a crude examination of graph structure, starting with some special classes of graphs that will be useful points of reference in Chapters 3 and 4, namely *lattice graphs* (or *d-lattices*) and *random graphs*.

Properties of d-Lattices

Definition 2.2.12. A *d-lattice* is a labelled, unweighted, undirected simple graph that is similar to a Euclidean cubic lattice of dimension d in that any vertex v is joined to its lattice neighbours, u_i and w_i, as specified by

$$u_i = \left[(v - i^{d'}) + n\right] \pmod{n},$$

$$w_i = (v + i^{d'}) \pmod{n},$$

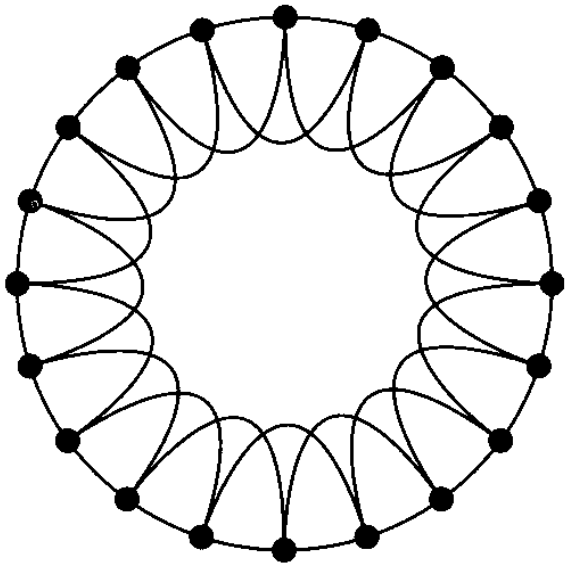

Figure 2.3 Example of a 1-lattice with $k = 4$.

where $1 \leq i \leq k/2$, $1 \leq d' \leq d$, and it is generally assumed that $k \geq 2d$.

Hence a 1-lattice with $k = 2$ is a ring, a 2-lattice with $k = 4$ is a two-dimensional square grid, an so on (see Figs. 2.3 and 2.4 for examples). In principle, k can be any number (although it makes sense to require $k \geq 2d$), and so we could have a 1-lattice with $k = 10$, in which case nearest neighbours, next-nearest neighbours, and so on would be connected (see Fig. 2.5 for another example). These structures are particularly convenient because their characteristic path lengths and clustering coefficients can be calculated explicitly. For a 1-lattice with even $k \geq 2$, simple enumeration shows

$$L = \frac{n(n+k-2)}{2k(n-1)}$$

and

$$\gamma = \frac{3\,(k-2)}{4\,(k-1)}.$$

It is obvious from these statements that L for a 1-lattice scales linearly with respect to n (for large n) and inversely with respect to k. The same length-scaling property can be inferred by considering the distribution sequence of a 1-lattice. Again, simple enumeration leads to the conclusion that $|\Gamma^i(v)| = k$ for all v and i. Hence $\Lambda_j = jk$, which is linear in j—the number of "degrees of separation." Necessarily, a linearly increasing distribution sequence corresponds to linear length scaling with n. Unlike L,

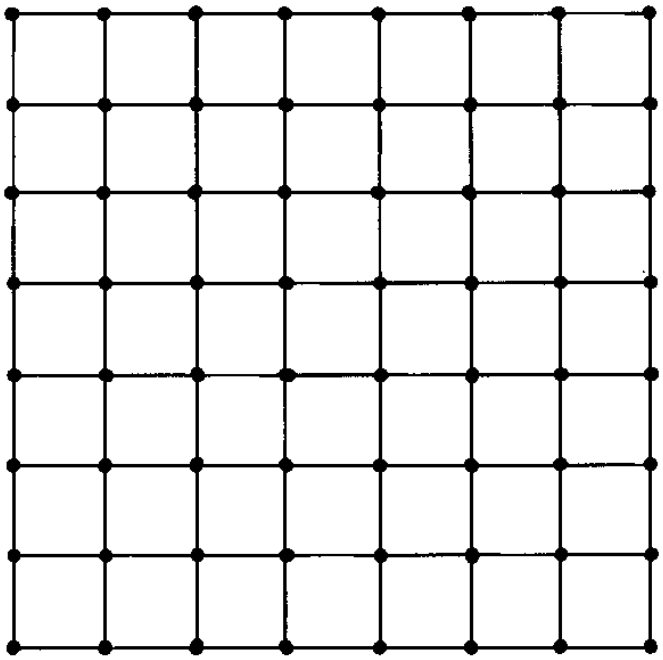

Figure 2.4 Example of a 2-lattice with $k = 4$ (boundaries are periodic).

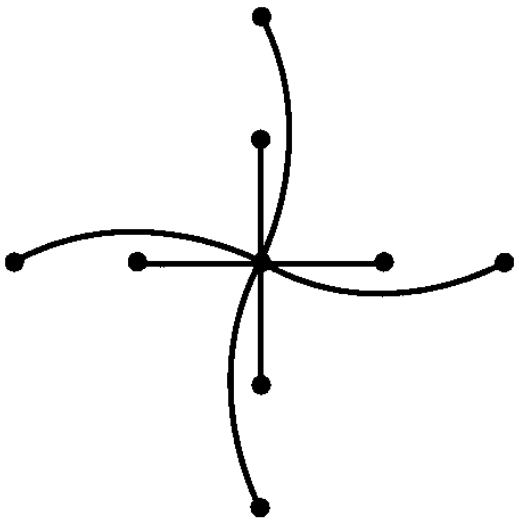

Figure 2.5 A single vertex in a 2-lattice with $k = 8$.

γ for a 1-lattice is independent of n and approaches 3/4 for sufficiently large k, at which point it is effectively independent of k also. *Hence any 1-lattice can be characterised by its length-scaling and clustering properties.* Similar statements hold for lattices in higher dimensions, which exhibit d-scaling as defined above.

Properties of Random Graphs

Strictly speaking, the kinds of random graphs discussed in this section do not appear very often in this book. However, they do constitute an important limiting case and, along with d-lattices, are frequently used as standard yardsticks in later chapters. They are also historically the root from which a number of serious investigations of the structure of social networks have sprung (the classic example being Harary 1959). Hence it is appropriate to run though at least the major classes of random graphs, some relevant terminology and definitions, and a few significant results that help in understanding and contextualising this work.

In its broadest sense, a random graph of order n is nothing more than a vertex set, consisting of n vertices, and an edge set that is generated in some random fashion. The set of all such graphs is called $\mathcal{G}^n$. Almost all of random graph theory, however, concerns itself with the analysis of one of two *models* of random graphs, referred to as $G(n, M)$ and $G(n, p)$, respectively, and the relationship between them. Many of the fundamental properties of these two models, along with the techniques used to analyse them, were developed in the late 1950s and early 1960s in a series of papers by Erdös and Rényi (1959, 1960, 1961a, 1961b), but all the material included here is referenced to the standard text on random graphs by Bollobás (1985).

Definition 2.2.13. $G(n, M)$ is a labelled graph with vertex set $V(G) = \{1, 2, \ldots, n\}$, having M randomly chosen edges (where M usually depends on n). $G(n, M)$ is frequently abbreviated as G_M.

Definition 2.2.14. $G(n, p)$ is a labelled graph with vertex set $V(G) = \{1, 2, \ldots, n\}$, in which every one of the possible $\binom{n}{2}$ edges exists with probability $0 < p < 1$, independent of any other edges. $G(n, p)$ is frequently abbreviated as G_p.

Random graph theory basically defines the conditions under which graphs belonging to either G_M or G_p possess a given *property* Q (for example, that it is connected), usually in the limit $n \to \infty$. Roughly speaking, broadly similar graphs with the same number of vertices share the same properties, and random graph theorists are interested in finding the conditions upon either M or p under which a particular property emerges in *almost all* graphs of the relevant model, in the infinite limit of n.

It turns out that for most purposes G_M and G_p are practically interchangeable, provided that $M \simeq pN$. G_p is easier to prove theorems with

because the edges are independent, whereas in G_M (because the total number of edges is fixed) there is necessarily some dependence of an edge being chosen, based on previous choices. This dependence is small, however, and does not affect any of the important results, so hereafter both models will be referred to simply as random graphs.

One of the most striking results of random graph theory is that most monotone properties[8] appear suddenly. That is, there exists a *threshold* function $M^*(n)$ that determines whether or not a graph is either very unlikely or very likely to have the property Q. This threshold can be defined in a number of ways, but perhaps the most intuitive is to think of $\mathcal{G}^n$ as a *graph process*. That is, starting from a vertex set with no edges, edges are added one-by-one in a random fashion, where each addition is regarded as a unit of time. The threshold function $M^*(n)$ is then regarded as a critical time, before which the property is unlikely to exist, and after which it is very likely. There are certain technical issues surrounding the uniqueness of these functions and how they differ between models, but these are not of concern here. The important thing to understand is that if we imagine random graphs as dynamic "organisms," growing in time, then the appearance of practically any property of interest will occur on a timescale that is very short compared with the timescale of the whole process. Similarly, if we imagine the development of random graphs as a trip through parameter space, then all the action happens in a very narrow region of that space.

This *threshold function* is strongly reminiscent of second-order phase transitions that have been well studied in statistical physics (see, for example, Stauffer and Aharony 1992) and that appear in the dynamical systems in Part II. Furthermore, even though the graphs considered in this book are not random graphs in the strict sense, and even though the number of edges $M(n)$ is preserved for all parameter values, precisely this kind of rapid transition occurs in terms of their large-scale, statistical properties. This interesting similarity and its connection with random graph theory receive more attention in Chapter 4.

Of all properties Q, none seems to have received more attention than *connectedness*. At which point in a graph process do the graphs become connected? What is their structure before they become connected? How do they make this transition? And once they are connected, how connected are they? That is, how many edges could be removed before they would once again be disconnected, and what is the expected distribution of completely connected subgraphs (*cliques*)? These are some of the major issues that have been addressed by random graph theorists

over the last forty-odd years. However, because the primary statistic of interest here is the characteristic path length (L), and because disconnected graphs have infinite L, only connected graphs will be considered. Admittedly this approach glosses over some important and interesting questions that are relevant to a completely general treatment of length in graphs. However, the resulting simplicity is useful for a first pass at the problem, and there is still much of interest to be learned. As far as random graphs are concerned, a famous theorem by Erdös and Rényi (1959) guarantees that "almost any" random graph with more than $n/2\ln(n)$ edges (equivalent to $k \gtrsim \ln(n)$) will be connected.[9] In practice, for finite n, it is sufficient simply to set $k \gg 1$ and check that no disconnected graphs are generated.

A final issue of interest is that of the *diameter* D of a random graph. This is important because, in random graph theory, the diameter is the principal measure of the characteristic path length of a graph, presumably because it is easier to prove theorems about than measures like the mean or median shortest path length. As we will see in Chapter 4, however, the mean path length of a random graph is dominated by the diameter, and so all the important results about diameter apply, more or less, to the notion of characteristic path length. There are two results concerning the diameter that are of interest here: the diameter of a random graph with a maximal degree of two, and the diameter of a random graph of arbitrary degree. The reason for the distinction is that one of the models in Chapter 3, for reasons to do with connectivity, is based on a regular connected substrate of degree two. Hence the length-scaling characteristics of such substrates are important to know. For instance, a cycle (or topological ring) with n vertices and $k = 2$ has $D \simeq n/2$ and thus exhibits linear length scaling. It turns out that this is true for any graph with a maximal degree of two, which is really equivalent to saying that *any* regular graph with $k = 2$ is a ring and so will exhibit a characteristic path length that scales linearly with respect to n. As random graph theory is almost always concerned with the properties of random graphs as $n \to \infty$, little is said about D as a function of n for arbitrary k. Still, two things seem clear from Bollobás's treatment of the problem:

1. Almost all random graphs with the same n; have the same D, for sufficiently large p (or k).
2. Random graphs are likely to be "spreading"; that is, the jth neighbourhood $\Gamma^j(v)$ includes mostly "new" vertices, vertices that have not been included at smaller j. Hence the number of vertices within

a distance j of any vertex v is never much less than $k(k-1)^{j-1}$. This represents an exponentially growing distribution sequence, which implies that $j_{\max} \sim \ln(n)/\ln(k)$.

Both these statements will be useful in the later numerical experiments: the first because it implies that the details of the construction algorithms of Chapter 3 should not be important in the random limit, and the second because it means that random graphs must have close to the smallest possible L for any fixed n and k.

2.2.6 Dimension and Embedding of Graphs

Although graphs are not usually defined in terms of any underlying Euclidean space, and most problems in graph theory do not require graphs to exist in any such space, it will still be useful to think about what dimension would be required if they did. More specifically, any given graph can be thought of as a set of points embedded in a Euclidean space, where the Euclidean distance between any two points is just the shortest path length between the corresponding vertices, to within some distortion. The operative question then is, For a given distortion, what is the minimum dimension Euclidean space required to embed a given graph? Of course, there is nothing unique about the Euclidean metric, and we could just as well ask the same question of any metric space. But Euclidean spaces are a natural choice because they are familiar, and also because the limiting cases for some of the models in later chapters is a d-lattice, which embeds precisely into $\mathbb{R}^d$. This dimension is the *embedding dimension*, defined as follows.

Definition 2.2.15. For a graph G and distortion $c \geq 1$, the *embedding dimension* $\dim_c(G)$ is the least dimension d such that there is an embedding ϕ of G into $\mathbb{R}^d$ where every two vertices $i, j \in G$ satisfy

$$d(i,j) \geq \|\phi(i) - \phi(j)\| \geq \frac{1}{c} d(i,j).$$

A theorem by Linial et al. (1995) guarantees that any graph with n vertices can be embedded in $\mathbb{R}^{\dim_c}$ with distortion $c \leq (1+\epsilon)c^*$, where $\dim_c = O(\log n)$ and $c^* = O(\log n)$. Linial's theorem is rather more general than this, but this restricted version is sufficient.

At this point, a few observations seem appropriate. First, this theorem (which Linial et al. actually present as an embedding algorithm) is closely related to the techniques of *multidimensional scaling* described in Section 2.1.1.

Second, for a fixed n, graphs with different topologies will, in general, have different embedding dimensions. For instance, it is obvious (even without Linial's theorem) that a d-lattice of any size will always have an embedding dimension of d, whereas for a random graph we are only guaranteed that $\dim_c(G) = O(\log n)$. The implication here is that, at least in the limit $n \to \infty$, random graphs "live" in $\mathbb{R}^\infty$, whereas in the same limit, d-lattices "live" in $\mathbb{R}^d$. This raises the conceptual challenge of what would happen to the dimension of a graph if it were to have its edges switched around one-by-one, causing it to move from a d-lattice to a random graph. This is a significant issue that will not be resolved here and that seems to be an interesting open research question.

Finally, if a graph has an embedding dimension d, then we might expect that its distribution sequence would grow like $\Lambda_j \propto j^d$. Certainly this seems plausible, if only by analogy to the distribution sequence of a d-lattice, which must necessarily grow in this fashion. The corresponding result for random graphs would be a distribution sequence that grows exponentially. The flip side of this observation is that one might expect any graph whose characteristic path length $L(n)$ scales logarithmically with respect to n to have a distribution sequence that grows exponentially with distance, and so it can be embedded only in a $\ln(n)$-dimensional space. One of the main results of the next chapter is that such graphs appear to be much more common than one might think.

Notes

CHAPTER 1
KEVEN BACON, THE SMALL WORLD, AND WHY IT ALL MATTERS

1. The "Oracle of Bacon" website, created by Brett Tjaden and Glenn Wasson, is located at www.cs.virginia.edu/bct7m/bacon.html.
2. http://www.us.imdb.com.
3. This fact, along with all others to do with the Kevin Bacon Game cited in this book, was correct as of April 1997. Since then the database has been extensively updated, but nothing essential has changed.

CHAPTER 2
AN OVERVIEW OF THE SMALL-WORLD PHENOMENON

1. This is essentially because, in a randomly connected system, the total number of members "reached" grows exponentially with increasing degree of separation, and $1{,}000 \times 1{,}000 \times 1{,}000$ is greater than the population of the United States.
2. The only difference between Barnes's density and clustering is that here v is not included as a member of its own neighbourhood—an attribute of some convenience because it allows clustering of zero.
3. In the telephone book method, subjects are asked to name acquaintances who have the same last names as those that appear on an imaginary "page" of a telephone directory, which is then treated as a representative subset of the entire population.
4. This calculation works because, in a population where acquaintance probabilities are independently distributed, the number of members linked to A by an acquaintance chain of less than a specified "length" grows exponentially with the length. Hence one thousand "friends" each is many more than is required to encompass the entire U.S. population within six handshakes.
5. In other words, if a network consists of only a few elements, or if at least one element is connected to a significant fraction of the total population, then it is not surprising that it should be "small," in which case there is nothing to explain.
6. Most researchers have confined themselves to the study of connected graphs because of the obvious problems associated with the apparent infinite lengths of disconnected graphs.

7. One method relates the average distance to the independence number of the graph, that is, the maximum size of a subgraph such that every pair of vertices in the subgraph are nonadjacent. Another utilises the eigenvalues of the adjacency matrix or the closely related Laplacian matrix, which treats the graph as a system of masses coupled by linear springs in place of the edges. The eigenvalues then characterise the modes of oscillation of the resulting coupled system. See Fiedler (1973) and Cvetković et al. (1979) for an explanation of the Laplacian matrix and the properties of its eigenvalues.

8. Monotone means just that if a particular random graph G possesses Q, then any graph H that includes G as a subgraph will also have Q.

9. There are a number of technical subleties to this result. See Chapter 10 of Alon and Spencer (1992) for a reasonably accessible description of the appearance (at $k \approx 1$) of the *giant component*, which proceeds to swallow up all remaining vertices, including the last few isolates when $k \gtrsim \ln(n)$.

CHAPTER 3
BIG WORLDS AND SMALL WORLDS: MODELS OF GRAPHS

1. This rapid rise as *fraction of mutual friends* $\rightarrow 1$ is present mostly for the purpose of enforcing continuity, but it could be justified, in modelling terms, by the argument that, even in such a random world, if two people have *all* their current friends in common, then they can't really avoid knowing each other.

2. There are many reasonable choices for p, but, whilst the specific choice can affect the results quantitatively, it appears to make little *qualitative* difference so long as it is sufficiently small (that is, $p \ll \binom{n}{2}^{-1}$). Here p is set at 10^{-10} for all numerical experiments.

3. Perhaps this should not be surprising, given Bollobás's observation of (1985, p. 41) that almost all random graphs of the same order (n) and size (M) are the same (in the sense that for any property Q, almost all graphs have Q or almost none of them do), regardless of the model ($G(n, M)$ or $G(n, p)$) used to create them.

4. The data points for $L(\alpha)$ and also $\gamma(\alpha)$ were averaged over one hundred random realisations of the construction algorithm (which, recall, makes connections at random but biased by the presence of mutually adjacent vertices) to reduce statistical fluctuations. In general, these fluctuations do not affect the qualitative nature of the data. Hence for convenience (where not otherwise noted), only a single realisation of the construction algorithm is used to generate results.

5. This is simply the probability that, when a vertex creates a new edge to another vertex with uniform random probability over the entire graph, that vertex will be in the same neighbourhood.

6. Of course, for manageable values of n, $n^{1/d}$, and $\ln(n)$ can be impossible to distinguish in practice, for any but small d. Hence only two-dimensional lattice substrates will be considered in detail.

7. The main difference between a β-graph with $\beta = 1$ and a true random graph is that all vertices in the β-graph are guaranteed to have degree at least $k/2$. Vertices, however, are still *connected to* at random, so a nonzero variance in