4 Random Graphs

Large graphs appear in many contexts such as the World Wide Web, the internet, social networks, journal citations, and other places. What is different about the modern study of large graphs from traditional graph theory and graph algorithms is that here one seeks statistical properties of these very large graphs rather than an exact answer to questions. This is akin to the switch physics made in the late 19^{th} century in going from mechanics to statistical mechanics. Just as the physicists did, one formulates abstract models of graphs that are not completely realistic in every situation, but admit a nice mathematical development that can guide what happens in practical situations. Perhaps the most basic such model is the $G(n,p)$ model of a random graph. In this chapter, we study properties of the $G(n,p)$ model as well as other models.

4.1 The $G(n,p)$ Model

The $G(n,p)$ model, due to Erdös and Rényi, has two parameters, n and p. Here n is the number of vertices of the graph and p is the edge probability. For each pair of distinct vertices, v and w, p is the probability that the edge (v,w) is present. The presence of each edge is statistically independent of all other edges. The graph-valued random variable with these parameters is denoted by $G(n,p)$. When we refer to "the graph $G(n,p)$", we mean one realization of the random variable. In many cases, p will be a function of n such as $p = d/n$ for some constant d. In this case, the expected degree of a vertex of the graph is $\frac{d}{n}(n-1) \approx d$. The interesting thing about the $G(n,p)$ model is that even though edges are chosen independently with no "collusion", certain global properties of the graph emerge from the independent choices. For small p, with $p = d/n$, $d < 1$, each connected component in the graph is small. For $d > 1$, there is a giant component consisting of a constant fraction of the vertices. In addition, as d increases there is a rapid transition in probability of a giant component at the threshold $d = 1$. Below the threshold, the probability of a giant component is very small, and above the threshold, the probability is almost one.

The phase transition at the threshold $d = 1$ from very small $o(n)$ size components to a giant $\Omega(n)$ sized component is illustrated by the following example. Suppose the vertices of the graph represents people and an edge means the two people it connects have met and became friends. Assume that the probability two people meet and become friends is $p = d/n$ and is statistically independent of all other friendships. The value of d can be interpreted as the expected number of friends a person knows. The question arises as to how large are the components in this friendship graph?

If the expected number of friends each person has is more than one, then a giant component will be present consisting of a constant fraction of all the people. On the other hand, if in expectation, each person has less than one friend, the largest component is a vanishingly small fraction of the whole. Furthermore, the transition from the vanishing fraction to a constant fraction of the whole happens abruptly between d slightly less than

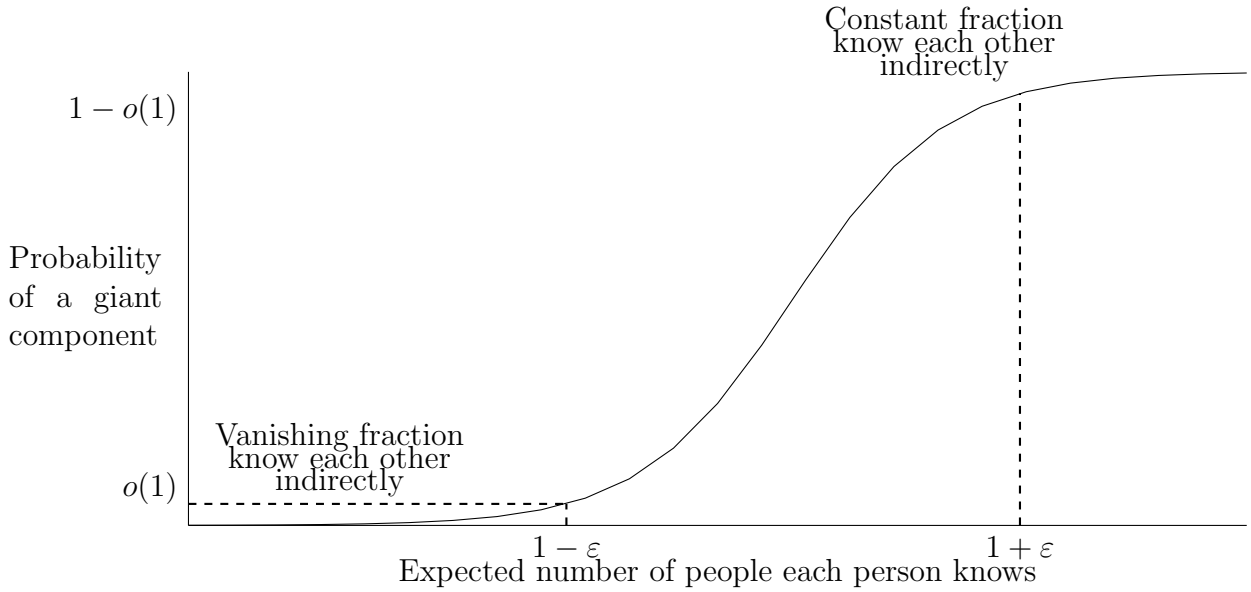


Figure 4.1: Probability of a giant component as a function of the expected number of people each person knows directly.

one to d slightly more than one. See Figure 4.1. Note that there is no global coordination of friendships. Each pair of individuals becomes friends independently.

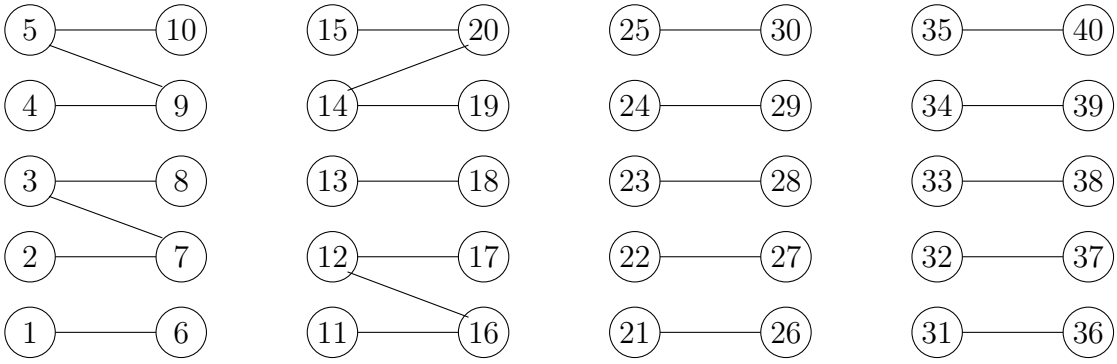
4.1.1 Degree Distribution

One of the simplest quantities to observe in a real graph is the number of vertices of given degree, called the vertex degree distribution. It is also very simple to study these distributions in $G(n,p)$ since the degree of each vertex is the sum of $n-1$ independent random variables, which results in a binomial distribution. Since p is the probability of an edge being present, the expected degree of a vertex is $d \approx pn$. The actual degree distribution is given by

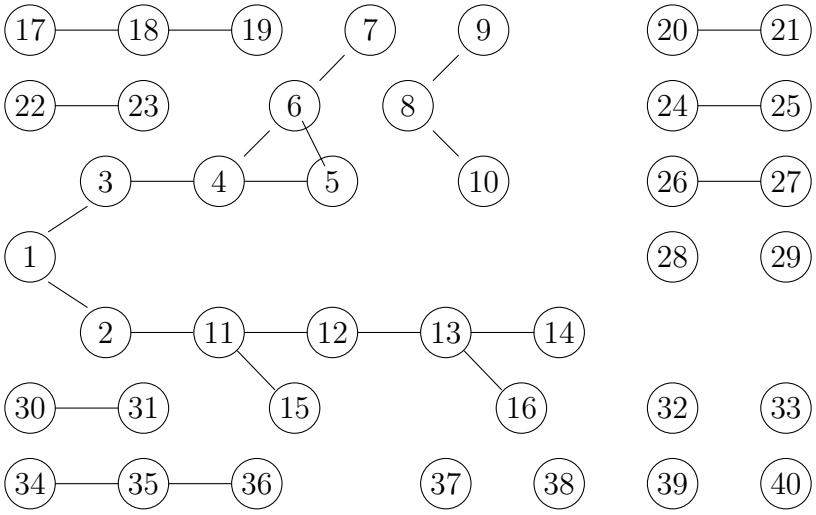
$$\text{Prob(vertex has degree } k) = \binom{n-1}{k} p^k (1-p)^{n-k-1} \approx \binom{n}{k} p^k (1-p)^{n-k}.$$

The quantity $\binom{n-1}{k}$ is the number of ways of choosing k edges, out of the possible $n-1$ edges, and $p^k(1-p)^{n-k-1}$ is the probability that the k selected edges are present and the remaining $n-k-1$ are not. Since n is large, replacing $n-1$ by n does not cause much error.

The binomial distribution falls off exponentially fast as one moves away from the mean. However, the degree distributions of graphs that appear in many applications do not exhibit such sharp drops. Rather, the degree distributions are much broader. This is often referred to as having a "heavy tail". The term tail refers to values of a random variable far away from its mean, usually measured in number of standard deviations. Thus, although the $G(n,p)$ model is important mathematically, more complex models are needed



A graph with 40 vertices and 24 edges



A randomly generated $G(n, p)$ graph with 40 vertices and 24 edges

Figure 4.2: Two graphs, each with 40 vertices and 24 edges. The second graph was randomly generated using the $G(n, p)$ model with $p = 1.2/n$. A graph similar to the top graph is almost surely not going to be randomly generated in the $G(n, p)$ model, whereas a graph similar to the lower graph will almost surely occur. Note that the lower graph consists of a giant component along with a number of small components that are trees.

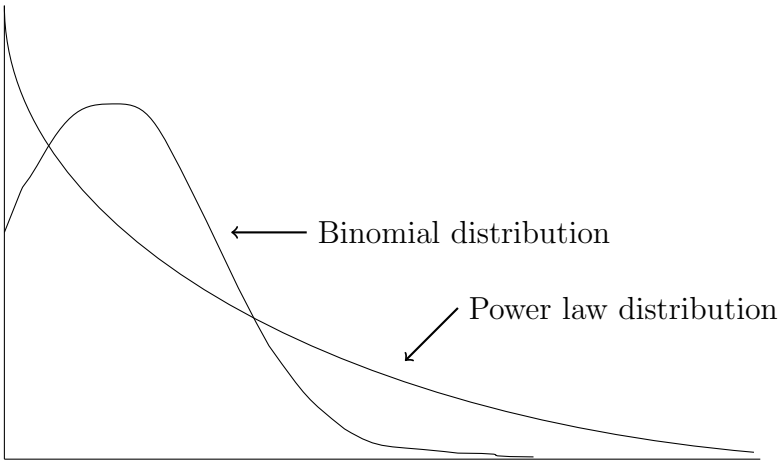


Figure 4.3: Illustration of the binomial and the power law distributions.

to represent real world graphs.

Consider an airline route graph. The graph has a wide range of degrees, from degree one or two for a small city, to degree 100, or more, for a major hub. The degree distribution is not binomial. Many large graphs that arise in various applications appear to have power law degree distributions. A power law degree distribution is one in which the number of vertices having a given degree decreases as a power of the degree, as in

$$\text{Number(degree } k \text{ vertices)} = c\frac{n}{k^r},$$

for some small positive real r, often just slightly less than three. Later, we will consider a random graph model giving rise to such degree distributions.

The following theorem claims that the degree distribution of the random graph $G(n,p)$ is tightly concentrated about its expected value. That is, the probability that the degree of a vertex differs from its expected degree, np, by more than $\lambda\sqrt{np}$, drops off exponentially fast with λ.

Theorem 4.1 *Let v be a vertex of the random graph $G(n,p)$. For $0 < \alpha < \sqrt{np}$*

$$Prob(|np - deg(v)| \geq \alpha\sqrt{np}) \leq 3e^{-\alpha^2/8}.$$

Proof: The degree $\deg(v)$ of vertex v is the sum of $n-1$ independent Bernoulli random variables, $x_1, x_2, \ldots, x_{n-1}$, where x_i is the indicator variable that the i^{th} edge from v is present. The theorem follows from Theorem **??**. ∎

Theorem 4.1 was for one vertex. The following corollary deals with all vertices.

Corollary 4.2 *Suppose ε is a positive constant. If p is $\Omega(\ln n/n\varepsilon^2)$, then, almost surely, every vertex has degree in the range $(1-\varepsilon)np$ to $(1+\varepsilon)np$.*

Proof: Apply Theorem 4.1 with $\alpha = \varepsilon\sqrt{np}$ to get that the probability that an individual vertex has degree outside the range $[(1-\varepsilon)np, (1+\varepsilon)np]$ is at most $3e^{-\varepsilon^2 np/8}$. By the union bound, the probability that some vertex has degree outside this range is at most $3ne^{-\varepsilon^2 np/8}$. For this to be $o(1)$, it suffices for p to be $\Omega(\ln n/n\varepsilon^2)$. Hence the Corollary. ∎

The assumption p is $\Omega(\ln n/n\varepsilon^2)$ is necessary. If $p = d/n$ for d a constant, then some vertices may have degrees outside the range. For $p = \frac{1}{n}$, Corollary 4.1 would claim almost surely that no vertex had a degree greater than a constant independent of n. But shortly we will see that it is highly likely that for $p = \frac{1}{n}$ there is a vertex of degree $\Omega(\log n/\log\log n)$.

When p is a constant, the expected degree of vertices in $G(n,p)$ increases with n. For example, in $G\left(n, \frac{1}{2}\right)$, the expected degree of a vertex is $n/2$. In many real applications, we will be concerned with $G(n,p)$ where $p = d/n$, for d a constant; i.e., graphs whose expected degree is a constant d independent of n. Holding $d = np$ constant as n goes to infinity, the binomial distribution

$$\text{Prob}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

approaches the Poisson distribution

$$\text{Prob}(k) = \frac{(np)^k}{k!} e^{-np} = \frac{d^k}{k!} e^{-d}.$$

move text beginning here to appendix
To see this, assume $k = o(n)$ and use the approximations $n - k \cong n$, $\binom{n}{k} \cong \frac{n^k}{k!}$, and $\left(1 - \frac{1}{n}\right)^{n-k} \cong e^{-1}$ to approximate the binomial distribution by

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{n^k}{k!} \left(\frac{d}{n}\right)^k (1 - \frac{d}{n})^n = \frac{d^k}{k!} e^{-d}.$$

Note that for $p = \frac{d}{n}$, where d is a constant independent of n, the probability of the binomial distribution falls off rapidly for $k > d$, and is essentially zero for all but some finite number of values of k. This justifies the $k = o(n)$ assumption. Thus, the Poisson distribution is a good approximation.
end of material to move

Example: In $G(n, \frac{1}{n})$ many vertices are of degree one, but not all. Some are of degree zero and some are of degree greater than one. In fact, it is highly likely that there is a vertex of degree $\Omega(\log n/\log\log n)$. The probability that a given vertex is of degree k is

$$\text{Prob}(k) = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \approx \frac{e^{-1}}{k!}.$$

If $k = \log n / \log\log n$,

$$\log k^k = k \log k \cong \frac{\log n}{\log\log n} (\log\log n - \log\log\log n) \cong \log n$$

and thus $k^k \cong n$. Since $k! \leq k^k \cong n$, the probability that a vertex has degree $k = \log n / \log\log n$ is at least $\frac{1}{k!} e^{-1} \geq \frac{1}{en}$. If the degrees of vertices were independent random variables, then this would be enough to argue that there would be a vertex of degree $\log n / \log\log n$ with probability at least $1 - \left(1 - \frac{1}{en}\right)^n = 1 - e^{-\frac{1}{e}} \cong 0.31$. But the degrees are not quite independent since when an edge is added to the graph it affects the degree of two vertices. This is a minor technical point, which one can get around. ∎

4.1.2 Existence of Triangles in $G(n, d/n)$

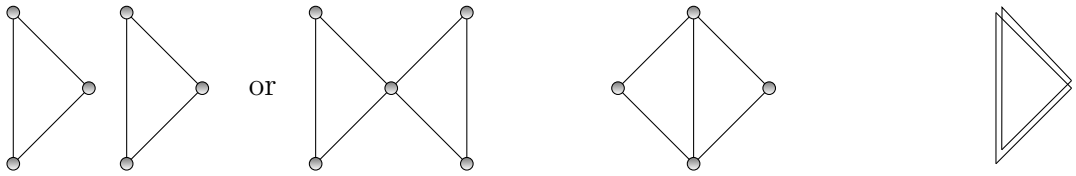
What is the expected number of triangles in $G\left(n, \frac{d}{n}\right)$, when d is a constant? As the number of vertices increases one might expect the number of triangles to increase, but this is not the case. Although the number of triples of vertices grows as n^3, the probability of an edge between two specific vertices decreases linearly with n. Thus, the probability of all three edges between the pairs of vertices in a triple of vertices being present goes down as n^{-3}, exactly canceling the rate of growth of triples.

A random graph with n vertices and edge probability d/n, has an expected number of triangles that is independent of n, namely $d^3/6$. There are $\binom{n}{3}$ triples of vertices. Each triple has probability $\left(\frac{d}{n}\right)^3$ of being a triangle. Let Δ_{ijk} be the indicator variable for the triangle with vertices i, j, and k being present. That is, all three edges (i, j), (j, k), and (i, k) being present. Then the number of triangles is $x = \sum_{ijk} \Delta_{ijk}$. Even though the existence of the triangles are not statistically independent events, by linearity of expectation, which does not assume independence of the variables, the expected value of a sum of random variables is the sum of the expected values. Thus, the expected number of triangles is

$$E(x) = E\left(\sum_{ijk} \Delta_{ijk}\right) = \sum_{ijk} E(\Delta_{ijk}) = \binom{n}{3} \left(\frac{d}{n}\right)^3 \approx \frac{d^3}{6}.$$

Even though on average there are $\frac{d^3}{6}$ triangles per graph, this does not mean that with high probability a graph has a triangle. Maybe half of the graphs have $\frac{d^3}{3}$ triangles and the other half have none for an average of $\frac{d^3}{6}$ triangles. Then, with probability 1/2, a graph selected at random would have no triangle. If $1/n$ of the graphs had $\frac{d^3}{6} n$ triangles and the remaining graphs had no triangles, then as n goes to infinity, the probability that a graph selected at random would have a triangle would go to zero.

We wish to assert that with some nonzero probability there is at least one triangle in $G(n, p)$ when $p = \frac{d}{n}$ for sufficiently large d. If all the triangles were on a small number of



The two triangles of Part 1 are either disjoint or share at most one vertex

The two triangles of Part 2 share an edge

The two triangles in Part 3 are the same triangle

Figure 4.4: The triangles in Part 1, Part 2, and Part 3 of the second moment argument for the existence of triangles in $G(n, \frac{d}{n})$.

graphs, then the number of triangles in those graphs would far exceed the expected value and hence the variance would be high. A second moment argument rules out this scenario where a small fraction of graphs have a large number of triangles and the remaining graphs have none.

Calculate $E(x^2)$ where x is the number of triangles. Write x as $x = \sum_{ijk} \Delta_{ijk}$, where Δ_{ijk} is the indicator variable of the triangle with vertices i, j, and k being present. Expanding the squared term

$$E(x^2) = E\Big(\sum_{i,j,k} \Delta_{ijk}\Big)^2 = E\Big(\sum_{\substack{i,\,j,\,k\\ i',j',k'}} \Delta_{ijk}\Delta_{i'j'k'}\Big).$$

Split the above sum into three parts. In Part 1, let S_1 be the set of i, j, k and i', j', k' which share at most one vertex and hence the two triangles share no edge. In this case, Δ_{ijk} and $\Delta_{i'j'k'}$ are independent and

$$E\Big(\sum_{S_1} \Delta_{ijk}\Delta_{i'j'k'}\Big) = \sum_{S_1} E(\Delta_{ijk})E(\Delta_{i'j'k'}) \le \Big(\sum_{\substack{\text{all}\\ ijk}} E(\Delta_{ijk})\Big)\Big(\sum_{\substack{\text{all}\\ i'j'k'}} E(\Delta_{i'j'k'})\Big) = E^2(x).$$

In the above formula how did we go from S_1 to all ijk?

In Part 2, i, j, k and i', j', k' share two vertices and hence one edge. See Figure 4.4. Four vertices and five edges are involved overall. There are at most $\binom{n}{4} \in O(n^4)$, 4-vertex subsets and $\binom{4}{2}$ ways to partition the four vertices into two triangles with a common edge. The probability of all five edges in the two triangles being present is p^5, so this part sums to $O(n^4p^5) = O(d^5/n)$ and is $o(1)$. There are so few triangles in the graph, the probability of two triangles sharing an edge is extremely unlikely.

In Part 3, i, j, k and i', j', k' are the same sets. The contribution of this part of the summation to $E(x^2)$ is $\binom{n}{3}p^3 = \frac{d^3}{6}$. Thus,

$$E(x^2) \le E^2(x) + \frac{d^3}{6} + o(1),$$

which implies

$$\text{Var}(x) = E(x^2) - E^2(x) \leq \frac{d^3}{6} + o(1).$$

For x to be less than or equal to zero, it must differ from its expected value by at least its expected value. Thus,

$$\text{Prob}(x = 0) \leq \text{Prob}\left(|x - E(x)| \geq E(x)\right).$$

By Chebychev inequality,

$$\text{Prob}(x = 0) \leq \frac{\text{Var}(x)}{E^2(x)} \leq \frac{d^3/6 + o(1)}{d^6/36} \leq \frac{6}{d^3} + o(1). \tag{4.1}$$

Thus, for $d > \sqrt[3]{6} \cong 1.8$, $\text{Prob}(x = 0) < 1$ and $G(n,p)$ has a triangle with nonzero probability. For $d < \sqrt[3]{6}$ and very close to zero, there simply are not enough edges in the graph for there to be a triangle.

4.2 Phase Transitions

Many properties of random graphs undergo structural changes as the edge probability passes some threshold value. This phenomenon is similar to the abrupt phase transitions in physics, as the temperature or pressure increases. Some examples of this are the abrupt appearance of cycles in $G(n,p)$ when p reaches $1/n$ and the disappearance of isolated vertices when p reaches $\frac{\log n}{n}$. The most important of these transitions is the emergence of a giant component, a connected component of size $\Theta(n)$, which happens at $d = 1$. Recall Figure 4.1.

For these and many other properties of random graphs, a threshold exists where an abrupt transition from not having the property to having the property occurs. If there exists a function $p(n)$ such that when $\lim_{n\to\infty} \frac{p_1(n)}{p(n)} = 0$, $G(n, p_1(n))$ almost surely does not have the property, and when $\lim_{n\to\infty} \frac{p_2(n)}{p(n)} = \infty$, $G(n, p_2(n))$ almost surely has the property, then we say that a *phase transition* occurs, and $p(n)$ is the *threshold*. Recall that $G(n,p)$ "almost surely does not have the property" means that the probability that it has the property goes to zero in the limit, as n goes to infinity. We shall soon see that every increasing property has a threshold. This is true not only for increasing properties of $G(n,p)$, but for increasing properties of any combinatorial structure. If for $cp(n)$, $c < 1$, the graph almost surely does not have the property and for $cp(n)$, $c > 1$, the graph almost surely has the property, then $p(n)$ is a *sharp threshold*. The existence of a giant component has a sharp threshold at $1/n$. We will prove this later.

In establishing phase transitions, we often use a variable $x(n)$ to denote the number of occurrences of an item in a random graph. If the expected value of $x(n)$ goes to zero as n goes to infinity, then a graph picked at random almost surely has no occurrence of the

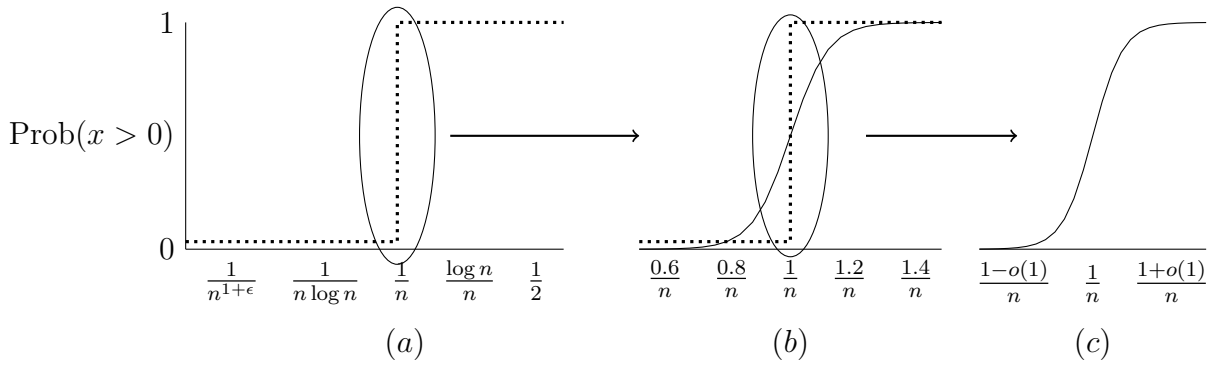


Figure 4.5: Figure 4.5(a) shows a phase transition at $p = \frac{1}{n}$. The dotted line shows an abrupt transition in Prob(x) from 0 to 1. For any function asymptotically less than $\frac{1}{n}$, Prob(x)>0 is zero and for any function asymptotically greater than $\frac{1}{n}$, Prob(x)>0 is one. Figure 4.5(b) expands the scale and shows a less abrupt change in probability unless the phase transition is sharp as illustrated by the dotted line. Figure 4.5(c) is a further expansion and the sharp transition is now more smooth.

item. This follows from Markov's inequality. Since x is a nonnegative random variable Prob($x \geq a$) $\leq \frac{1}{a}E(x)$, which implies that the probability of $x(n) \geq 1$ is at most $E(x(n))$. That is, if the expected number of occurrences of an item in a graph goes to zero, the probability that there are one or more occurrences of the item in a randomly selected graph goes to zero. This is called the *first moment method*.

The previous section showed that the property of having a triangle has a threshold at $p(n) = 1/n$. If the edge probability $p_1(n)$ is $o(1/n)$, then the expected number of triangles goes to zero and by the first moment method, the graph almost surely has no triangle. However, if the edge probability $p_2(n)$ satisfies $np_2(n) \to \infty$, then from (4.1), the probability of having no triangle is at most $6/d^3 + o(1) = 6/(np_2(n))^3 + o(1)$, which goes to zero. This latter case uses what we call the second moment method. The first and second moment methods are broadly used. We describe the second moment method in some generality now.

When the expected value of $x(n)$, the number of occurrences of an item, goes to infinity, we cannot conclude that a graph picked at random will likely have a copy since the items may all appear on a small fraction of the graphs. We resort to a technique called the *second moment method*. It is a simple idea based on Chebyshev's inequality.

Theorem 4.3 (Second Moment method) *Let $x(n)$ be a random variable with $E(x) > 0$. If*

$$Var(x) = o\left(E^2(x)\right),$$

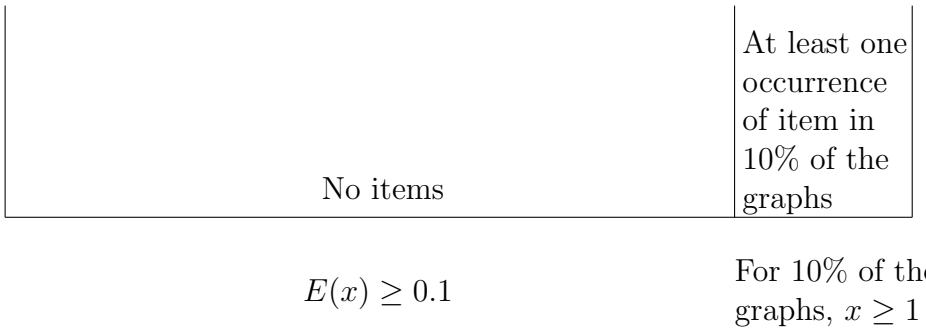


Figure 4.6: If the expected fraction of the number of graphs in which an item occurs did not go to zero, then $E(x)$, the expected number of items per graph, could not be zero. Suppose 10% of the graphs had at least one occurrence of the item. Then the expected number of occurrences per graph must be at least 0.1. Thus, $E(x) = 0$ implies the probability that a graph has an occurrence of the item goes to zero. However, the other direction needs more work. If $E(x)$ were not zero, a second moment argument is needed to conclude that the probability that a graph picked at random had an occurrence of the item was nonzero since there could be a large number of occurrences concentrated on a vanishingly small fraction of all graphs. The second moment argument claims that for a nonnegative random variable x with $E(x) > 0$, if $\mathrm{Var}(x)$ is $o(E^2(x))$ or alternatively if $E(x^2) \leq E^2(x)(1 + o(1))$, then almost surely $x > 0$.

then x is almost surely greater than zero.

Proof: If $E(x) > 0$, then for x to be less than or equal to zero, it must differ from its expected value by at least its expected value. Thus,

$$\mathrm{Prob}(x \leq 0) \leq \mathrm{Prob}\left(|x - E(x)| \geq E(x)\right).$$

By Chebyshev inequality

$$\mathrm{Prob}\left(|x - E(x)| \geq E(x)\right) \leq \frac{\mathrm{Var}(x)}{E^2(x)} \to 0.$$

Thus, $\mathrm{Prob}(x \leq 0)$ goes to zero if $\mathrm{Var}(x)$ is $o(E^2(x))$. ∎

Corollary 4.4 *Let x be a random variable with $E(x) > 0$. If*

$$E(x^2) \leq E^2(x)(1 + o(1)),$$

then x is almost surely greater than zero.

Proof: If $E(x^2) \leq E^2(x)(1 + o(1))$, then

$$Var(x) = E(x^2) - E^2(x) \leq E^2(x)o(1) = o(E^2(x)).$$

∎

Threshold for graph diameter two

We now present the first example of a sharp phase transition for a property. This means that slightly increasing the edge probability p near the threshold takes us from almost surely not having the property to almost surely having it. The property is that of a random graph having diameter less than or equal to two. The diameter of a graph is the maximum length of the shortest path between a pair of nodes.

The following technique for deriving the threshold for a graph having diameter two is a standard method often used to determine the threshold for many other objects. Let x be a random variable for the number of objects such as triangles, isolated vertices, or Hamilton circuits, for which we wish to determine a threshold. Then we determine the value of p, say p_0, where the expected value of x goes from zero to infinity. For $p < p_0$ almost surely a graph selected at random will not have a copy of x. For $p > p_0$, a second moment argument is needed to establish that the items are not concentrated on a vanishingly small fraction of the graphs and that a graph picked at random will almost surely have a copy.

Our first task is to figure out what to count to determine the threshold for a graph having diameter two. A graph has diameter two if and only if for each pair of vertices i and j, either there is an edge between them or there is another vertex k to which both i and j have an edge. The set of neighbors of i and the set of neighbors of j are random subsets of expected cardinality np. For these two sets to intersect requires $np \approx \sqrt{n}$ or $p \approx \frac{1}{\sqrt{n}}$. Such statements often go under the general name of "birthday paradox" though it is not a paradox. In what follows, we will prove a threshold of $O(\sqrt{\ln n}/\sqrt{n})$ for a graph to have diameter two. The extra factor of $\sqrt{\ln n}$ ensures that every one of the $\binom{n}{2}$ pairs of i and j has a common neighbor. When $p = c\sqrt{\frac{\ln n}{n}}$, for $c < \sqrt{2}$, the graph almost surely has diameter greater than two and for $c > \sqrt{2}$, the graph almost surely has diameter less than or equal to two.

Theorem 4.5 *The property that $G(n,p)$ has diameter two has a sharp threshold at $p = \sqrt{2}\sqrt{\frac{\ln n}{n}}$.*

Proof: If G has diameter greater than two, then there exists a pair of nonadjacent vertices i and j such that no other vertex of G is adjacent to both i and j. This motivates calling such a pair *bad*.

Introduce a set of indicator random variables I_{ij}, one for each pair of vertices (i,j) with $i < j$, where I_{ij} is 1 if and only if the pair (i,j) is bad. Let

$$x = \sum_{i<j} I_{ij}$$

be the number of bad pairs of vertices. Putting $i < j$ in the sum ensures each pair (i, j) is counted only once. A graph has diameter at most two if and only if it has no bad pair, i.e., $x = 0$. Thus, if $\lim_{n\to\infty} E(x) = 0$, then for large n, almost surely, a graph has no bad pair and hence has diameter at most two.

The probability that a given vertex is adjacent to both vertices in a pair of vertices (i, j) is p^2. Hence, the probability that the vertex is not adjacent to both vertices is $1 - p^2$. The probability that no vertex is adjacent to the pair (i, j) is $(1 - p^2)^{n-2}$ and the probability that i and j are not adjacent is $1 - p$. Since there are $\binom{n}{2}$ pairs of vertices, the expected number of bad pairs is

$$E(x) = \binom{n}{2}(1 - p)\left(1 - p^2\right)^{n-2}.$$

Setting $p = c\sqrt{\frac{\ln n}{n}}$,

$$\begin{aligned} E(x) &\cong \tfrac{n^2}{2}\left(1 - c\sqrt{\tfrac{\ln n}{n}}\right)\left(1 - c^2\tfrac{\ln n}{n}\right)^n \\ &\cong \tfrac{n^2}{2}e^{-c^2 \ln n} \\ &\cong \tfrac{1}{2}n^{2-c^2}. \end{aligned}$$

For $c > \sqrt{2}$, $\lim_{n\to\infty} E(x) \to 0$. Thus, by the first moment method, for $p = c\sqrt{\frac{\ln n}{n}}$ with $c > \sqrt{2}$, $G(n, p)$ almost surely has no bad pair and hence has diameter at most two.

Next, consider the case $c < \sqrt{2}$ where $\lim_{n\to\infty} E(x) \to \infty$. We appeal to a second moment argument to claim that almost surely a graph has a bad pair and thus has diameter greater than two.

$$E(x^2) = E\left(\sum_{i<j} I_{ij}\right)^2 = E\left(\sum_{i<j} I_{ij} \sum_{k<l} I_{kl}\right) = E\left(\sum_{\substack{i<j \\ k<l}} I_{ij} I_{kl}\right) = \sum_{\substack{i<j \\ k<l}} E(I_{ij} I_{kl}).$$

The summation can be partitioned into three summations depending on the number of distinct indices among i, j, k, and l. Call this number a.

$$E\left(x^2\right) = \sum_{\substack{i<j \\ k<l}} E(I_{ij} I_{kl}) + \sum_{\substack{i<j \\ i<k}} E(I_{ij} I_{ik}) + \sum_{i<j} E\left(I_{ij}^2\right). \tag{4.2}$$

$$a = 4 \qquad a = 3 \qquad a = 2$$

Consider the case $a = 4$ where i, j, k, and l are all distinct. If $I_{ij} I_{kl} = 1$, then both pairs (i, j) and (k, l) are bad and so for each $u \notin \{i, j, k, l\}$, one of the edges (i, u) or (j, u)

is absent and, in addition, one of the edges (k,u) or (l,u) is absent. The probability of this for one u not in $\{i,j,k,l\}$ is $(1-p^2)^2$. As u ranges over all the $n-4$ vertices not in $\{i,j,k,l\}$, these events are all independent. Thus,

$$E(I_{ij}I_{kl}) \leq (1-p^2)^{2(n-4)} \leq (1-c^2\frac{\ln n}{n})^{2n}(1+o(1)) \leq n^{-2c^2}(1+o(1))$$

and the first sum is

$$\sum_{\substack{i<j\\k<l}} E(I_{ij}I_{kl}) \leq n^{4-2c^2}(1+o(1)).$$

For the second summation, observe that if $I_{ij}I_{ik}=1$, then for every vertex u not equal to i, j, or k, either there is no edge between i and u or there is an edge (i,u) and both edges (j,u) and (k,u) are absent. The probability of this event for one u is

$$1-p+p(1-p)^2 = 1-2p^2+p^3 \approx 1-2p^2.$$

Thus, the probability for all such u is $(1-2p^2)^{n-3}$. Substituting $c\sqrt{\frac{\ln n}{n}}$ for p yields

$$\left(1-\tfrac{2c^2\ln n}{n}\right)^{n-3} \cong e^{-2c^2\ln n} = n^{-2c^2},$$

which is an upper bound on $E(I_{ij}I_{kl})$ for one i,j,k, and l with $a=3$. Summing over all distinct triples yields n^{3-2c^2} for the second summation in (4.2).

For the third summation, since the value of I_{ij} is zero or one, $E\left(I_{ij}^2\right) = E\left(I_{ij}\right)$. Thus,

$$\sum_{ij} E\left(I_{ij}^2\right) = E\left(x\right).$$

Hence, $E\left(x^2\right) \leq n^{4-2c^2} + n^{3-2c^2} + n^{2-c^2}$ and $E\left(x\right) \cong n^{2-c^2}$, from which it follows that for $c<\sqrt{2}$, $E\left(x^2\right) \leq E^2\left(x\right)(1+o(1))$. By a second moment argument, Corollary 4.4, a graph almost surely has at least one bad pair of vertices and thus has diameter greater than two. Therefore, the property that the diameter of $G(n,p)$ is less than or equal to two has a sharp threshold at $p=\sqrt{2}\sqrt{\frac{\ln n}{n}}$ ∎

Disappearance of Isolated Vertices

The disappearance of isolated vertices in $G\left(n,p\right)$ has a sharp threshold at $\frac{\ln n}{n}$. At this point the giant component has absorbed all the small components and with the disappearance of isolated vertices, the graph becomes connected.

Theorem 4.6 *The disappearance of isolated vertices in $G\left(n,p\right)$ has a sharp threshold of $\frac{\ln n}{n}$.*

Proof: Let x be the number of isolated vertices in $G(n,p)$. Then,

$$E(x) = n(1-p)^{n-1}.$$

Since we believe the threshold to be $\frac{\ln n}{n}$, consider $p = c\frac{\ln n}{n}$. Then,

$$\lim_{n\to\infty} E(x) = \lim_{n\to\infty} n\left(1-\tfrac{c\ln n}{n}\right)^n = \lim_{n\to\infty} ne^{-c\ln n} = \lim_{n\to\infty} n^{1-c}.$$

If c >1, the expected number of isolated vertices, goes to zero. If $c < 1$, the expected number of isolated vertices goes to infinity. If the expected number of isolated vertices goes to zero, it follows that almost all graphs have no isolated vertices. On the other hand, if the expected number of isolated vertices goes to infinity, a second moment argument is needed to show that almost all graphs have an isolated vertex and that the isolated vertices are not concentrated on some vanishingly small set of graphs with almost all graphs not having isolated vertices.

Assume $c < 1$. Write $x = I_1 + I_2 + \cdots + I_n$ where I_i is the indicator variable indicating whether vertex i is an isolated vertex. Then $E(x^2) = \sum_{i=1}^{n} E(I_i^2) + 2\sum_{i<j} E(I_iI_j)$. Since I_i equals 0 or 1, $I_i^2 = I_i$ and the first sum has value $E(x)$. Since all elements in the second sum are equal

$$\begin{aligned} E\left(x^2\right) &= E(x) + n(n-1)E(I_1I_2) \\ &= E(x) + n(n-1)(1-p)^{2(n-1)-1}. \end{aligned}$$

The minus one in the exponent $2(n-1)-1$ avoids counting the edge from vertex 1 to vertex 2 twice. Now,

$$\begin{aligned} \frac{E(x^2)}{E^2(x)} &= \frac{n(1-p)^{n-1} + n(n-1)(1-p)^{2(n-1)-1}}{n^2(1-p)^{2(n-1)}} \\ &= \frac{1}{n(1-p)^{n-1}} + (1-\frac{1}{n})\frac{1}{1-p}. \end{aligned}$$

For $p = c\frac{\ln n}{n}$ with $c < 1$, $\lim_{n\to\infty} E(x) = \infty$ and

$$\lim_{n\to\infty} \frac{E(x^2)}{E^2(x)} = \lim_{n\to\infty} \left[\frac{1}{n^{1-c}} + (1-\frac{1}{n})\frac{1}{1-c\frac{\ln n}{n}}\right] = 1 + o(1).$$

By the second moment argument, Corollary 4.4, the probability that $x = 0$ goes to zero implying that almost all graphs have an isolated vertex. Thus, $\frac{\ln n}{n}$ is a sharp threshold for the disappearance of isolated vertices. For $p = c\frac{\ln n}{n}$, when $c > 1$ there almost surely are no isolated vertices, and when $c < 1$ there almost surely are isolated vertices. ∎

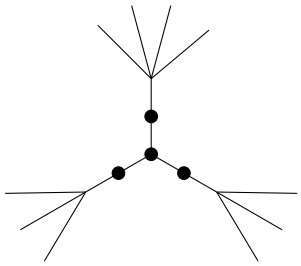


Figure 4.7: A degree three vertex with three adjacent degree two vertices. Graph cannot have a Hamilton circuit.

Hamilton circuits

So far in establishing phase transitions in the $G(n,p)$ model for an item such as the disappearance of isolated vertices, we introduced a random variable x that was the number of occurrences of the item. We then determined the probability p for which the expected value of x went from zero to infinity. For values of p for which $E(x) = 0$, we argued that with probability one, a graph generated at random had no occurrences of x. For values of x for which $E(x) \to \infty$, we used the second moment argument to conclude that with probability one a graph generated at random had occurrences of x. That is, the occurrences that forced $E(x)$ to infinity were not all concentrated on a vanishingly small fraction of the graphs. One might raise the question for the $G(n,p)$ graph model, do there exist items that are so concentrated on a small fraction of the graphs that the value of p where $E(x)$ goes from zero to infinity is not the threshold? An example where this happens is Hamilton circuits.

Let x be the number of Hamilton circuits in $G(n,p)$ and let $p = \frac{d}{n}$ for some constant d. There are $\frac{1}{2}(n-1)!$ potential Hamilton circuits in a graph and each has probability $(\frac{d}{n})^n$ of actually being a Hamilton circuit. Thus,

$$\begin{aligned} E(x) &= \frac{1}{2}(n-1)! \left(\frac{d}{n}\right)^n \\ &\simeq \left(\frac{n}{e}\right)^n \left(\frac{d}{n}\right)^n \\ &= \begin{cases} 0 & d < e \\ \infty & d > e \end{cases}. \end{aligned}$$

This suggests that the threshold for Hamilton circuits occurs when d equals Euler's constant e. This is not possible since the graph still has isolated vertices and is not even connected for $p = \frac{e}{n}$. Thus, the second moment argument is indeed necessary.

The actual threshold for Hamilton circuits is $d = \omega(\log n + \log\log n)$. For any $p(n)$ asymptotically greater than $\frac{1}{n}(\log n + \log\log n)$, $G(n,p)$ will have a Hamilton circuit with

probability one. This is the same threshold as for the disappearance of degree one vertices. Clearly a graph with a degree one vertex cannot have a Hamilton circuit. But it may seem surprising that Hamilton circuits appear as soon as degree one vertices disappear. You may ask why at the moment degree one vertices disappear there cannot be a subgraph consisting of a degree three vertex adjacent to three degree two vertices as shown in Figure 4.7. The reason is that the frequency of degree two and three vertices in the graph is very small and the probability that four such vertices would occur together in such a subgraph is too small for it to happen.

4.3 The Giant Component

Consider $G(n, p)$ as p grows. Starting with $p = 0$, the graph has n vertices and no edges. As p increases and edges are added, a forest of trees emerges. When p is $o(1/n)$ the graph is almost surely a forest of trees, i.e., there are no cycles. When p is d/n, d a constant, cycles appear. For $d < 1$, no connected component has asymptotically more than $\log n$ vertices. The number of components containing a single cycle is a constant independent of n. Thus, the graph consists of a forest of trees plus a few components that have a single cycle with no $\Omega(\log n)$ size components.

At p equal $1/n$, a phase transition occurs in which a giant component emerges. The transition consists of a double jump. At $p = 1/n$, components of $n^{2/3}$ vertices emerge, which are almost surely trees. Then at $p = d/n$, $d > 1$, a true giant component emerges that has a number of vertices proportional to n. This is a seminal result in random graph theory and the main subject of this section. Giant components also arise in many real world graphs; the reader may want to look at large real-world graphs, like portions of the web and find the size of the largest connected component.

When one looks at the connected components of large graphs that appear in various contexts, one observes that often there is one very large component. One example is a graph formed from a data base of protean interactions[1] where vertices correspond to proteins and edges correspond to pairs of proteins that interact. By an interaction, one means two amino acid chains that bind to each other for a function. The graph has 2735 vertices and 3602 edges. At the time we looked at the data base, the associated graph had the number of components of various sizes shown in Table 3.1. There are a number of small components, but only one component of size greater than 16, and that is a giant component of size 1851. As more proteins are added to the data base the giant component will grow even larger and eventually swallow up all the smaller components.

The existence of a giant component is not unique to the graph produced from the protein data set. Take any data set that one can convert to a graph and it is likely that the graph will have a giant component, provided that the ratio of edges to vertices

[1]Science 1999 July 30 Vol. 285 No. 5428 pp751-753.

Size of component	1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$	15	16	$\cdots$	1851
Number of components	48	179	50	25	14	6	4	6	1	1	1	0	0	0	1	0	1

Table 1: Table 3.1 Size of components in the graph implicit in the database of interacting proteins.

is a small number greater than one half. Table 3.2 gives two other examples. This phenomenon, of the existence of a giant component in many real world graphs deserves study.

ftp://ftp.cs.rochester.edu/pub/u/joel/papers.lst
Vertices are papers and edges mean that two papers shared an author.

1	2	3	4	5	6	7	8	14	27488
2712	549	129	51	16	12	8	3	1	1

http://www.gutenberg.org/etext/3202
Vertices represent words and edges connect words that are synonyms of one another.

1	2	3	4	5	14	16	18	48	117	125	128	30242
7	1	1	1	0	1	1	1	1	1	1	1	1

Table 2: Table 3.2 Size of components in two graphs constructed from data sets.

Returning to $G(n,p)$, as p increases beyond d/n, all nonisolated vertices are absorbed into the giant component, and at $p = \frac{1}{2}\frac{\ln n}{n}$, the graph consists only of isolated vertices plus a giant component. At $p = \frac{\ln n}{n}$, the graph becomes completely connected. By $p = 1/2$, the graph is not only connected, but is sufficiently dense that it has a clique of size $(2-\varepsilon)\log n$ for any $\varepsilon > 0$. We prove many of these facts in this chapter.

To compute the size of a connected component of $G(n,p)$, do a breadth first search of a component starting from an arbitrary vertex and generate an edge only when the search process needs to know if the edge exists. Start at an arbitrary vertex and mark it discovered and unexplored. At a general step, select a discovered, but unexplored vertex v, and explore it as follows. For each undiscovered vertex u, independently decide with probability $p = d/n$ whether the edge (v,u) is in and if it is, mark u discovered and unexplored. After this, mark v explored. Discovered but unexplored vertices are called the frontier. The algorithm has found the entire connected component when the frontier becomes empty.

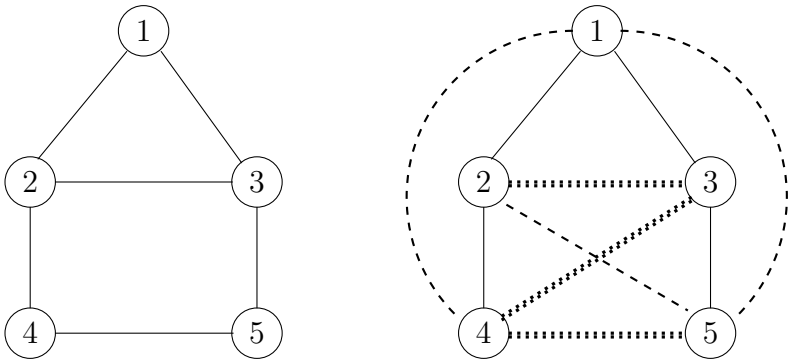


Figure 4.8: A graph (left) and the breadth first search of the graph (right). At vertex 1 the algorithm queried all edges. The solid edges are real edges, the dashed edges are edges that were queried but do not exist. At vertex 2 the algorithm queried all possible edges to vertices not yet discovered. The algorithm does not query whether the edge (2,3) exists since vertex 3 has already been discovered when the algorithm is at vertex 2. Potential edges not queried are illustrated with dotted edges. **IS THIS FIGURE USEFUL?**

For each vertex u, other than the start vertex, the probability that u is undiscovered after the first i steps is precisely $(1-\frac{d}{n})^i$. A step is the full exploration of one vertex. Let z_i be the number of vertices discovered in the first i steps of the search. The distribution of z_i is Binomial$\Big(n-1, 1-\big(1-\frac{d}{n}\big)^i\Big)$.

Consider the case $d > 1$. For small values of i, the probability that a vertex is undiscovered after i steps is

$$\left(1-\frac{d}{n}\right)^i \approx 1-\frac{id}{n}.$$

The probability that a vertex is discovered after i steps is $\frac{id}{n}$. The expected number of discovered vertices grows as id and the expected size of the frontier grows as $(d-1)\,i$. As the fraction of discovered vertices increases, the expected rate of growth of newly discovered vertices decreases since many of the vertices adjacent to the vertex currently being searched have already been discovered. Once $\frac{d-1}{d}n$ vertices have been discovered, the growth of newly discovered vertices slows to one at each step. Eventually for $d >1$, the growth of discovering new vertices drops below one per step and the frontier starts to shrink. For $d <1$, $(d-1)\,i$, the expected size of the frontier is negative. The expected rate of growth is less than one, even at the start.

Now assume $d > 1$. As we saw, the expected size of the frontier grows as $(d-1)i$ for small i. The actual size of the frontier is a random variable. What is the probability that the actual size of the frontier will differ from the expected size of the frontier by a sufficient amount so that the actual size of the frontier is zero? To answer this, we need to understand the distribution of the number of discovered vertices after i steps.

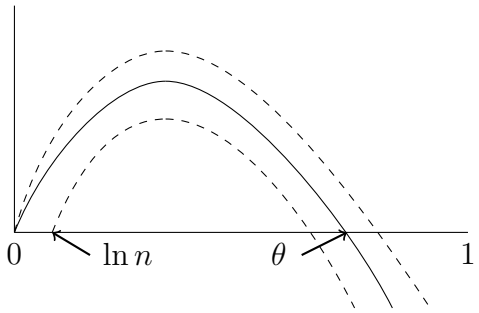


Figure 4.9: The solid curve is the expected size of the frontier. The two dashed curves indicate the range of possible values for the actual size of the frontier.

For small i, the probability that a vertex has been discovered is $1-(1-d/n)^i \approx id/n$ and the binomial distribution for the number of discovered vertices, binomial$(n, \frac{id}{n})$, is well approximated by the Poisson distribution with the same mean id. The probability that a total of k vertices have been discovered in i steps is approximately $e^{-di}\frac{(di)^k}{k!}$. For a connected component to have exactly i vertices, the frontier must drop to zero for the first time at step i. A necessary condition is that exactly i vertices must have been discovered in the first i steps. The probability of this approximately equals

$$e^{-di}\frac{(di)^i}{i!} = e^{-di}\frac{d^i i^i}{i^i}e^i = e^{-(d-1)i}d^i = e^{-(d-1-\ln d)i}.$$

For $d > 1$, $\ln d \leq d-1$ and hence $d-1-\ln d > 0$. This probability drops off exponentially with i. For $i > c\ln n$ and sufficiently large c, the probability that the breadth first search starting from a particular vertex terminates with a component of size i is $o(1/n)$ as long as the Poisson approximation is valid. In the range of this approximation, the probability that a breadth first search started from any vertex terminates with $i > c\ln n$ vertices is $o(1)$. Intuitively, if the component has not stopped growing within $\Omega(\ln n)$ steps, it is likely to continue to grow until it becomes much larger and the expected value of the size of the frontier again becomes small. While the expected value of the frontier is large, the probability that the actual size will differ from the expected size sufficiently for the actual size of the frontier to be zero is vanishingly small.

In Theorem 4.8, we prove that there is one giant component of size $\Omega(n)$ along with a number of components of size $O(\ln n)$. We first prove a technical lemma stating that the probability of a vertex being in a small component is strictly less than one and hence there is a giant component. We refer to a connected component of size $O(\log n)$ as a small component.

Lemma 4.7 *Assume $d > 1$. The probability that $cc(v)$, the connected component containing vertex v, is small (i.e., of size $O(\log n)$) is a constant strictly less than 1.*

Proof: Let p be the probability that $cc(v)$ is small, i.e., the probability that the breadth first search started at v terminates before $c_1 \log n$ vertices are discovered. Slightly modify the breadth first search as follows: If in exploring a vertex u at some point, there are m undiscovered vertices, choose the number k of vertices which will be adjacent to u from Binomial$(m, \frac{d}{n})$ distribution. Having picked k, pick one of the $\binom{m}{k}$ subsets of m undiscovered vertices to be the set of vertices adjacent to u, and make the other $m-k$ vertices not adjacent to u. This process has the same distribution as picking each edge from u independently at random to be present with probability d/n. As the search proceeds, m decreases. If cc(v) is small, m is always greater than $s = n - c_1 \log n$. Modify the process once more picking k from Binomial$(s, \frac{d}{n})$ instead of from Binomial$(m, \frac{d}{n})$. Let p' be the probability that cc(v) is small for the modified process. Clearly, $p' \geq p$, so it suffices to prove that p' is a constant strictly less than one. The mean of the binomial now is $d_1 = sd/n$ which is strictly greater than one. It is clear that the probability that the modified process ends before $c_1 \log n$ vertices are discovered is at least the probability for the original process, since picking from $n - c_1 \log n$ vertices has decreased the number of newly discovered vertices each time. Modifying the process so that the newly discovered vertices are picked from a fixed size set, converts the problem to what is called a branching process..

A branching process is a method for creating a possibly infinite random tree. There is a nonnegative integer-valued random variable y that is the number of children of the node being explored. First, the root v of the tree chooses a value of y according to the distribution of y and spawns that number of children. Each of the children independently chooses a value according to the same distribution of y and spawns that many children. The process terminates when all of the vertices have spawned children. The process may go on forever. If it does terminate with a finite tree, we say that the process has become "extinct". Let Binomial$(s, \frac{d}{n})$ be the distribution of y. Let q be the probability of extinction. Then, $q \geq p'$, since, the breadth first search terminating with at most $c_1 \log n$ vertices is one way of becoming extinct. Let $p_i = \binom{s}{i}(d/n)^i(1-(d/n))^{s-i}$ be the probability that y spawns i children. We have $\sum_{i=0}^{s} p_i = 1$ and $\sum_{i=1}^{s} ip_i = E(y) = ds/n > 1$.

The depth of a tree is at most the number of nodes in the tree. Let a_t be the probability that the branching process terminates at depth at most t. If the root v has no children, then the process terminates with depth one where the root is counted as a depth one node which is at most t. If v has i children, the process from v terminates at depth at most t if and only if the i sub processes, one rooted at each child of v terminate at depth $t-1$ or less. The i processes are independent, so the probability that they all terminate at depth at most $t-1$ is exactly a_{t-1}^i. With this we get:

$$a_t = p_0 + \sum_{i=1}^{s} p_i a_{t-1}^i = \sum_{i=0}^{s} p_i a_{t-1}^i.$$

We have $a_1 = p_0 < 1$. There is a constant $\alpha \in [p_0, 1)$ such that whenever $a_{t-1} \leq \alpha$, the above recursion implies that $a_t \leq \alpha$. This would finish the proof since then $a_1 \leq \alpha$ implies

For a small number i of steps, the probability distribution of the size of the set of discovered vertices at time i is $p(k) = e^{-di}\frac{(di)^k}{k!}$ and has expected value di. Thus, the expected size of the frontier is $(d-1)i$. For the frontier to be empty would require that the size of the set of discovered vertices be smaller than its expected value by $(d-1)i$. That is, the size of the set of discovered vertices would need to be $di-(d-1)i = i$. The probability of this is

$$e^{-di}\frac{(di)^i}{i!} = e^{-di}\frac{d^i i^i}{i^i}e^i = e^{-(d-1)i}d^i = e^{-(d-1-\ln d)i}$$

which drops off exponentially fast with i provided $d > 1$. Since $d-1-\ln d$ is some constant $c > 0$, the probability is e^{-ci} which for $i = \ln n$ is $e^{-c\ln n} = \frac{1}{n^c}$. Thus, with high probability, the largest small component in the graph is of size at most $\ln n$.

Illustration 4.1

$a_2 \leq \alpha$ which implies $a_3 \leq \alpha$ etc. and so $q = \lim_{t\to\infty} a_t \leq \alpha$.

To prove the claim, consider the polynomial

$$h(x) = x - \sum_{i=0}^{s} p_i x^i.$$

We see that $h(1) = 0$ and $h'(1) = 1 - \sum_{i=1}^{s} ip_i \approx 1 - \frac{sd}{n}$, which is at most a strictly negative constant. By continuity of $h(\cdot)$, there is exists some $x_0 < 1$ such that $h(x) \geq 0$ for $x \in [x_0, 1]$. Take $\alpha = \text{Max}(x_0, p_0)$. Now since $\sum_{i=0}^{s} p_i x^i$ has all nonnegative coefficients, it is an increasing function of x and so if a_{t-1} is at least α, then, $\sum_{i=0}^{s} p_i a_{t-1}^i$ is at least $\sum_{i=0}^{s} p_i \alpha^i \geq \alpha$. Now, if $a_{t-1} \leq \alpha$,

$$a_t = \sum_{i=0}^{s} p_i a_{t-1}^i \geq \sum_{i=1}^{s} p_i \alpha^i = \alpha - h(\alpha) \leq \alpha,$$

proving the claim. ∎

We now prove in Theorem 4.8 that in $G(n, \frac{d}{n})$, $d > 1$ there is one giant component containing a fraction of the n vertices and that the remaining vertices are in components of size less than some constant c_1 times $\log n$. There are no components greater than $c_1 \log n$ other than the giant component.

Theorem 4.8 *Let p=d/n with $d > 1$.*

1. *There are constants c_1 and c_2 such that the probability that there is a connected component of size between $c_1 \log n$ and $c_2 n$ is at most $1/n$.*

2. *The number of vertices in components of size $O(\ln n)$ is almost surely at most cn for some $c < 1$. Thus, with probability $1 - o(1)$, there is a connected component of size $\Omega(n)$.*

3. *The probability that there are two or more connected components, each of size more than $n^{2/3}$, is at most $1/n$.*

Proof: In the breadth first search of a component, the probability that a vertex has not been discovered in i steps is $\left(1-\frac{d}{n}\right)^i$. It is easy to see that the approximation $(1-d/n)^i \approx 1-id/n$ is valid as long as $i \leq c_2 n$ for a suitable constant c_2 since the error term in the approximation is $O(i^2d^2/n^2)$, which for $i \leq c_2 n$ is at most a small constant times id/n. This establishes (1).

Next consider (2). For a vertex v, let $\mathrm{cc}(v)$ denote the set of vertices in the connected component containing v. By (1), almost surely, $\mathrm{cc}(v)$ is a small set of size at most $c_1 \log n$ or a large set of size at least $c_2 n$ for every vertex v. The central part of the proof of (2) that the probability of a vertex being in a small component is strictly less than one was established in Lemma 4.7. Let x be the number of vertices in a small connected component. Lemma 4.7 implies that the expectation of the random variable x equals the number of vertices in small connected components is at most some $c_3 n$, for a constant c_3 strictly less than one. But we need to show that for any graph almost surely the actual number x of such vertices is at most some constant strictly less than one times n. For this, we use the second moment method. In this case, the proof that the variance of x is $o(E^2(x))$ is easy. Let x_i be the indicator random variable of the event that $\mathrm{cc}(i)$ is small. Let S and T run over all small sets. Noting that for $i \neq j$, $\mathrm{cc}(i)$ and $\mathrm{cc}(j)$ either are the same or are disjoint,

$$
\begin{aligned}
E(x^2) &= E\left(\left(\sum_{i=1}^{n} x_i\right)^2\right) = \sum_{i,j} E(x_i x_j) = \sum_i E(x_i^2) + \sum_{i \neq j} E(x_i x_j) \\
&= E(x) + \sum_{i \neq j} \sum_S \mathrm{Prob}\left(\mathrm{cc}(i) = \mathrm{cc}(j) = S\right) + \sum_{i \neq j} \sum_{\substack{S,T \\ \text{disjoint}}} \mathrm{Prob}\left(\mathrm{cc}(i) = S;\ \mathrm{cc}(j) = T\right) \\
&= E(x) + \sum_{i \neq j} \sum_S \mathrm{Prob}\left(\mathrm{cc}(i) = \mathrm{cc}(j) = S\right) \\
&\qquad + \sum_{i \neq j} \sum_{\substack{S,T \\ \text{disjoint}}} \mathrm{Prob}\left(\mathrm{cc}(i) = S\right) \mathrm{Prob}\left(\mathrm{cc}(j) = T\right) (1-p)^{-|S||T|} \\
&\leq O(n) + (1-p)^{-|S||T|} \left(\sum_S \mathrm{Prob}\left(\mathrm{cc}(i) = S\right)\right) \left(\sum_T \mathrm{Prob}\left(\mathrm{cc}(j) = T\right)\right) \\
&\leq O(n) + \left(1 + o(1)\right) E(x) E(x).
\end{aligned}
$$

In the next to last line, if S containing i and T containing j are disjoint sets, then the two events, S is a connected component and T is a connected component, depend on disjoint sets of edges except for the $|S||T|$ edges between S vertices and T vertices. Let c_4 be a constant in the interval $(c_3, 1)$. Then, by Chebyshev inequality,

$$
\mathrm{Prob}(x > c_4 n) \leq \frac{\mathrm{Var}(x)}{(c_4 - c_3)^2 n^2} \leq \frac{O(n) + o(1) c_3^2 n^2}{(c_4 - c_3)^2 n^2} = o(1).
$$

For the proof of (3) suppose a pair of vertices u and v belong to two different connected components, each of size at least $n^{2/3}$. With high probability, they should have merged into one component producing a contradiction. First, run the breadth first search process starting at v for $\frac{1}{2}n^{2/3}$ steps. Since v is in a connected component of size $n^{2/3}$, there are $\Omega(n^{2/3})$ frontier vertices. The expected size of the frontier continues to grow until some constant times n and the actual size of the frontier does not differ significantly from the expected size. The size of the component also grows linearly with n. Thus, the frontier is of size $n^{\frac{2}{3}}$. See Exercise 4.27. By the assumption, u does not belong to this connected component. Now, temporarily stop the breadth first search tree of v and begin a breadth first search tree starting at u, again for $\frac{1}{2}n^{2/3}$ steps. It is important to understand that this change of order of building $G(n,p)$ does not change the resulting graph. We can choose edges in any order since the order does not affect independence or conditioning. The breadth first search tree from u also will have $\Omega(n^{2/3})$ frontier vertices with high probability . Now grow the u tree further. The probability that none of the edges between the two frontier sets is encountered is $(1-p)^{\Omega(n^{4/3})} \leq e^{-\Omega(dn^{1/3})}$, which converges to zero. So almost surely, one of the edges is encountered and u and v end up in the same connected component. This argument shows for a particular pair of vertices u and v, the probability that they belong to different large connected components is very small. Now use the union bound to conclude that this does not happen for any of the $\binom{n}{2}$ pairs of vertices. The details are left to the reader. ∎

4.4 Branching Processes

A *branching process* is a method for creating a random tree. Starting with the root node, each node has a probability distribution for the number of its children. The root of the tree denotes a parent and its descendants are the children with their descendants being the grandchildren. The children of the root are the first generation, their children the second generation, and so on. Branching processes have obvious applications in population studies, but also in exploring a connected component in a random graph.

We analyze a simple case of a branching process where the distribution of the number of children at each node in the tree is the same. The basic question asked is what is the probability that the tree is finite, i.e., the probability that the branching process dies out? This is called the *extinction probability*.

Our analysis of the branching process will give the probability of extinction, as well as the expected size of the components conditioned on extinction. Not surprisingly, the expected size of components conditioned on extinction is $O(1)$. This says that in $G(n, \frac{d}{n})$, with $d > 1$, there is one giant component of size $\Omega(n)$, the rest of the components are $O(\ln n)$ in size and the expected size of the small components is $O(1)$.

An important tool in our analysis of branching processes is the generating function. The generating function for a nonnegative integer valued random variable y is $f(x) = \sum_{i=0}^{\infty} p_i x^i$ where p_i is the probability that y equals i. The reader not familiar with generating functions should consult Section **??** of the appendix.

Let the random variable z_j be the number of children in the j^{th} generation and let $f_j(x)$ be the generating function for z_j. Then $f_1(x) = f(x)$ is the generating function for the first generation where $f(x)$ is the generating function for the number of children at a node in the tree. The generating function for the 2^{nd} generation is $f_2(x) = f(f(x))$. In general, the generating function for the $j+1^{st}$ generation is given by $f_{j+1}(x) = f_j(f(x))$. To see this, observe two things.

First, the generating function for the sum of two identically distributed integer valued random variables x_1 and x_2 is the square of their generating function

$$f^2(x) = p_0^2 + (p_0p_1 + p_1p_0)\, x + (p_0p_2 + p_1p_1 + p_2p_0)\, x^2 + \cdots .$$

For $x_1 + x_2$ to have value zero, both x_1 and x_2 must have value zero, for $x_1 + x_2$ to have value one, exactly one of x_1 or x_2 must have value zero and the other have value one, and so on. In general, the generating function for the sum of i independent random variables, each with generating function $f(x)$, is $f^i(x)$.

The second observation is that the coefficient of x^i in $f_j(x)$ is the probability of there being i children in the j^{th} generation. If there are i children in the j^{th} generation, the number of children in the $j+1^{st}$ generation is the sum of i independent random variables each with generating function $f(x)$. Thus, the generating function for the $j+1^{st}$ generation, given i children in the j^{th} generation, is $f^i(x)$. The generating function for the $j+1^{st}$ generation is given by

$$f_{j+1}(x) = \sum_{i=0}^{\infty} \text{Prob}(z_j = i) f^i(x).$$

If $f_j(x) = \sum_{i=0}^{\infty} a_i x^i$, then f_{j+1} is obtained by substituting $f(x)$ for x in $f_j(x)$.

Since $f(x)$ and its iterates, $f_2, f_3, \ldots$, are all polynomials in x with nonnegative coefficients, $f(x)$ and its iterates are all monotonically increasing and convex on the unit interval. Since the probabilities of the number of children of a node sum to one, if $p_0 < 1$, some coefficient of x to a power other than zero in $f(x)$ is nonzero and $f(x)$ is strictly increasing.

Let q be the probability that the branching process dies out. If there are i children in the first generation, then each of the i subtrees must die out and this occurs with probability q^i. Thus, q equals the summation over all values of i of the product of the

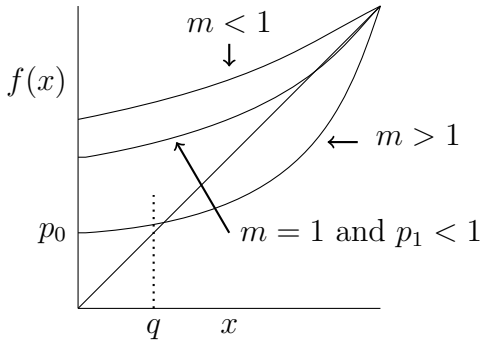


Figure 4.10: Illustration of the root of equation $f(x) = x$ in the interval [0,1).

probability of i children times the probability that i subtrees will die out. This gives $q = \sum_{i=0}^{\infty} p_i q^i$. Thus, q is the root of $x = \sum_{i=0}^{\infty} p_i x^i$, that is $x = f(x)$.

This suggests focusing on roots of the equation $f(x) = x$ in the interval [0,1]. The value $x = 1$ is always a root of the equation $f(x) = x$ since $f(1) = \sum_{i=0}^{\infty} p_i = 1$. When is there a smaller nonnegative root? The derivative of $f(x)$ at $x = 1$ is $f'(1) = p_1 + 2p_2 + 3p_3 + \cdots$. Let $m = f'(1)$. Thus, m is the expected number of children of a node. If $m > 1$, one might expect the tree to grow forever, since each node at time j is expected to have more than one child. But this does not imply that the probability of extinction is zero. In fact, if $p_0 > 0$, then with positive probability, the root will have no children and the process will become extinct right away. Recall that for $G(n, \frac{d}{n})$, the expected number of children is d, so the parameter m plays the role of d.

If $m < 1$, then the slope of $f(x)$ at $x = 1$ is less than one. This fact along with convexity of $f(x)$ implies that $f(x) > x$ for x in $[0, 1)$ and there is no root of $f(x) = x$ in the interval $[0, 1)$.

If $m = 1$ and $p_1 < 1$, then once again convexity implies that $f(x) > x$ for $x \in [0, 1)$ and there is no root of $f(x) = x$ in the interval $[0, 1)$. If $m = 1$ and $p_1 = 1$, then $f(x)$ is the straight line $f(x) = x$.

If $m >1$, then the slope of $f(x)$ is greater than the slope of x at $x = 1$. This fact, along with convexity of $f(x)$, implies $f(x) = x$ has a unique root in [0,1). When $p_0 = 0$, the root is at $x = 0$.

Let q be the smallest nonnegative root of the equation $f(x) = x$. For $m < 1$ and for m=1 and $p_0 < 1$, q equals one and for $m >1$, q is strictly less than one. We shall see that the value of q is the *extinction probability* of the branching process and that $1 - q$ is the *immortality probability*. That is, q is the probability that for some j, the number of

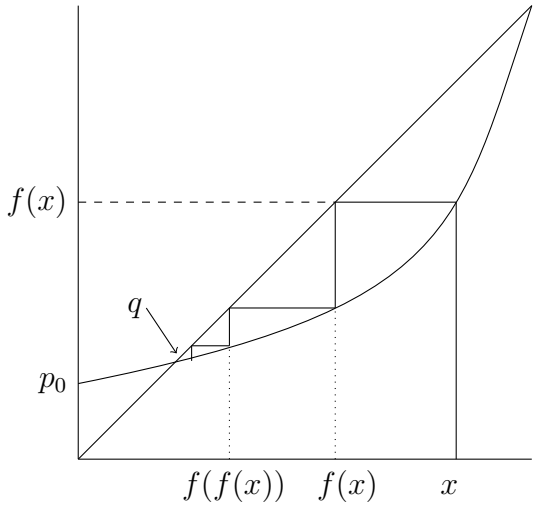


Figure 4.11: Illustration of convergence of the sequence of iterations $f_1(x), f_2(x), \ldots$ to q.

children in the j^{th} generation is zero. To see this, note that for $m > 1$, $\lim_{j\to\infty} f_j(x) = q$ for $0 \leq x < 1$. Figure 4.11 illustrates the proof which is given in Lemma 4.9. Similarly note that when $m < 1$ or $m = 1$ with $p_0 < 1$, $f_j(x)$ approaches one as j approaches infinity.

Lemma 4.9 *Assume $m > 1$. Let q be the unique root of f(x)=x in [0,1). In the limit as j goes to infinity, $f_j(x) = q$ for x in $[0, 1)$.*

Proof: If $0 \leq x \leq q$, then $x < f(x) \leq f(q)$ and iterating this inequality

$$x < f_1(x) < f_2(x) < \cdots < f_j(x) < f(q) = q.$$

Clearly, the sequence converges and it must converge to a fixed point where $f(x) = x$. Similarly, if $q \leq x < 1$, then $f(q) \leq f(x) < x$ and iterating this inequality

$$x > f_1(x) > f_2(x) > \cdots > f_j(x) > f(q) = q.$$

In the limit as j goes to infinity $f_j(x) = q$ for all x, $0 \leq x < 1$. ∎

Recall that $f_j(x)$ is the generating function $\sum_{i=0}^{\infty} \text{Prob}(z_j = i)\, x^i$. The fact that in the limit the generating function equals the constant q, and is not a function of x, says that $\text{Prob}(z_j = 0) = q$ and $\text{Prob}(z_j = i) = 0$ for all finite nonzero values of i. The remaining probability is the probability of a nonfinite component. Thus, when $m >1$, q is the extinction probability and 1-q is the probability that z_j grows without bound, i.e., immortality.

Theorem 4.10 *Consider a tree generated by a branching process. Let $f(x)$ be the generating function for the number of children at each node.*

1. *If the expected number of children at each node is less than or equal to one, then the probability of extinction is one unless the probability of exactly one child is one.*

2. *If the expected number of children of each node is greater than one, then the probability of extinction is the unique solution to* $f(x) = x$ *in* $[0, 1)$.

Proof: Let p_i be the probability of i children at each node. Then $f(x) = p_0 + p_1 x + p_2 x^2 + \cdots$ is the generating function for the number of children at each node and $f'(1) = p_1 + 2p_2 + 3p_3 + \cdots$ is the slope of $f(x)$ at $x = 1$. Observe that $f'(1)$ is the expected number of children at each node.

Since the expected number of children at each node is the slope of $f(x)$ at $x = 1$, if the expected number of children is less than or equal to one, the slope of $f(x)$ at $x = 1$ is less than or equal to one and the unique root of $f(x) = x$ in $(0, 1]$ is at $x = 1$ and the probability of extinction is one unless $f'(1) = 1$ and $p_1 = 1$. If $f'(1) = 1$ and $p_1 = 1$, $f(x) = x$ and the tree is an infinite degree one chain. If the slope of $f(x)$ at $x = 1$ is greater than one, then the probability of extinction is the unique solution to $f(x) = x$ in $[0, 1)$. ∎

A branching process with m <1 or m=1 and $p_1 < 1$ dies out with probability one. If m=1 and $p_1 = 1$, then the branching process consists of an infinite chain with no fan out. If m >1, then the branching process will die out with some probability less than one unless $p_0 = 0$ in which case it cannot die out, since a node always has at least one descendent.

Note that the branching process corresponds to finding the size of a component in an infinite graph. In a finite graph, the probability distribution of descendants is not a constant as more and more vertices of the graph get discovered.

The simple branching process defined here either dies out or goes to infinity. In biological systems there are other factors, since processes often go to stable populations. One possibility is that the probability distribution for the number of descendants of a child depends on the total population of the current generation.

Expected size of extinct families

We now show that the expected size of an extinct family is finite, provided that $m \neq 1$. Note that at extinction, the size must be finite. However, the expected size at extinction could conceivably be infinite, if the probability of dying out did not decay fast enough. To see how the expected value of a random variable that is always finite could be infinite, let x be an integer valued random variable. Let p_i be the probability that $x = i$. If $\sum_{i=1}^{\infty} p_i = 1$, then with probability one, x will be finite. However, the expected value of x may be infinite. That is, $\sum_{i=0}^{\infty} i p_i = \infty$. For example, if for i >0, $p_i = \frac{6}{\pi}\frac{1}{i^2}$, then $\sum_{i=1}^{\infty} p_i = 1$,

but $\sum_{i=1}^{\infty} ip_i = \infty$. The value of the random variable x is always finite, but its expected value is infinite. This does not happen in a branching process, except in the special case where the slope $m = f'(1)$ equals one and $p_1 \neq 1$

Lemma 4.11 *If the slope $m = f'(1)$ does not equal one, then the expected size of an extinct family is finite. If the slope m equals one and $p_1 = 1$, then the tree is an infinite degree one chain and there are no extinct families. If $m=1$ and $p_1 < 1$, then the expected size of the extinct family is infinite.*

Proof: Let z_i be the random variable denoting the size of the i^{th} generation and let q be the probability of extinction. The probability of extinction for a tree with k children in the first generation is q^k since each of the k children has an extinction probability of q. Note that the expected size of z_1, the first generation, over extinct trees will be smaller than the expected size of z_1 over all trees since when the root node has a larger number of children than average, the tree is more likely to be infinite.

By Bayes rule

$$\text{Prob}(z_1 = k|\text{extinction}) = \text{Prob}(z_1 = k) \frac{\text{Prob}(\text{extinction}|z_1 = k)}{\text{Prob}(\text{extinction})} = p_k \frac{q^k}{q} = p_k q^{k-1}.$$

Knowing the probability distribution of z_1 given extinction, allows us to calculate the expected size of z_1 given extinction.

$$E(z_1|\text{extinction}) = \sum_{k=0}^{\infty} k p_k q^{k-1} = f'(q).$$

We now prove, using independence, that the expected size of the i^{th} generation given extinction is

$$E(z_i|\text{extinction}) = \left(f'(q)\right)^i.$$

For $i = 2$, z_2 is the sum of z_1 independent random variables, each independent of the random variable z_1. So, $E(z_2|z_1 = j \text{ and extinction}) = E(\text{ sum of } j \text{ copies of } z_1|\text{extinction}) = jE(z_1|\text{extinction})$. Summing over all values of j

$$\begin{aligned} E(z_2|\text{extinction}) &= \sum_{j=1}^{\infty} E(z_2|z_1 = j \text{ and extinction})\text{Prob}(z_1 = j|\text{extinction}) \\ &= \sum_{j=1}^{\infty} jE(z_1|\text{extinction})\text{Prob}(z_1 = j|\text{extinction}) \\ &= E(z_1|\text{extinction}) \sum_{j=1}^{\infty} j\text{Prob}(z_1 = j|\text{extinction}) = E^2(z_1|\text{extinction}). \end{aligned}$$

Since $E(z_1|\text{extinction}) = f'(q)$, $E(z_2|\text{extinction}) = (f'(q))^2$. Similarly, $E(z_i|\text{extinction}) = (f'(q))^i$. The expected size of the tree is the sum of the expected sizes of each generation. That is,

$$\text{Expected size of tree given extinction} = \sum_{i=0}^{\infty} E(z_i|\text{extinction}) = \sum_{i=0}^{\infty} (f'(q))^i = \frac{1}{1-f'(q)}.$$

Thus, the expected size of an extinct family is finite since $f'(q) < 1$ provided $m \neq 1$.

The fact that $f'(q) < 1$ is illustrated in Figure 4.10. If m <1, then q=1 and $f'(q) = m$ is less than one. If m >1, then $q \in [0, 1)$ and again $f'(q)$ <1 since q is the solution to $f(x) = x$ and $f'(q)$ must be less than one for the curve $f(x)$ to cross the line x. Thus, for m <1 or m >1, $f'(q)$ <1 and the expected tree size of $\frac{1}{1-f'(q)}$ is finite. For m=1 and $p_1 < 1$, one has q=1 and thus $f'(q) = 1$ and the formula for the expected size of the tree diverges. ∎

4.5 Cycles and Full Connectivity

This section considers when cycles form and when the graph becomes fully connected. For both of these problems, we look at each subset of k vertices and see when they form either a cycle or a connected component.

4.5.1 Emergence of Cycles

The emergence of cycles in $G(n, p)$ has a threshold when p equals to $1/n$.

Theorem 4.12 *The threshold for the existence of cycles in $G(n, p)$ is $p = 1/n$.*

Proof: Let x be the number of cycles in $G(n, p)$. To form a cycle of length k, the vertices can be selected in $\binom{n}{k}$ ways. Given the k vertices of the cycle, they can be ordered by arbitrarily selecting a first vertex, then a second vertex in one of k-1 ways, a third in one of $k-2$ ways, etc. Since a cycle and its reversal are the same cycle, divide by 2. Thus, there are $\binom{n}{k}\frac{(k-1)!}{2}$ cycles of length k and

$$E(x) = \sum_{k=3}^{n} \binom{n}{k} \tfrac{(k-1)!}{2} p^k \leq \sum_{k=3}^{n} \tfrac{n^k}{2k} p^k \leq \sum_{k=3}^{n} (np)^k = (np)^3 \, \tfrac{1-(np)^{n-2}}{1-np} \leq 2(np)^3,$$

provided that $np < 1/2$. When p is asymptotically less than $1/n$, then $\lim_{n\to\infty} np = 0$ and $\lim_{n\to\infty} \sum_{k=3}^{n} (np)^k = 0$. So, as n goes to infinity, $E(x)$ goes to zero. Thus, the graph almost surely has no cycles by the first moment method. A second moment argument can be used to show that for $p = d/n$, $d > 1$, a graph will have a cycle with probability tending to one. ∎

The argument above does not yield a sharp threshold since we argued that $E(x) \to 0$ only under the assumption that p is asymptotically less that $\frac{1}{n}$.. A sharp threshold requires $E(x) \to 0$ for $p = d/n$, $d < 1$.

Consider what happens in more detail when $p = d/n$, d a constant.

$$\begin{aligned} E(x) &= \sum_{k=3}^{n} \binom{n}{k} \frac{(k-1)!}{2} p^k \\ &= \frac{1}{2} \sum_{k=3}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!} (k-1)!\, p^k \\ &= \frac{1}{2} \sum_{k=3}^{n} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{d^k}{k}. \end{aligned}$$

$E(x)$ converges if $d < 1$, and diverges if $d \geq 1$. If $d < 1$, $E(x) \leq \frac{1}{2} \sum_{k=3}^{n} \frac{d^k}{k}$ and $\lim_{n\to\infty} E(x)$ equals a constant greater than zero. If $d = 1$, $E(x) = \frac{1}{2} \sum_{k=3}^{n} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{1}{k}$. Consider only the first $\log n$ terms of the sum. Since $\frac{n}{n-i} = 1 + \frac{i}{n-i} \leq e^{i/n-i}$, it follows that $\frac{n(n-1)\cdots(n-k+1)}{n^k} \geq 1/2$. Thus,

$$E(x) \geq \frac{1}{2} \sum_{k=3}^{\log n} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{1}{k} \geq \frac{1}{4} \sum_{k=3}^{\log n} \frac{1}{k}.$$

Then, in the limit as n goes to infinity

$$\lim_{n\to\infty} E(x) \geq \lim_{n\to\infty} \frac{1}{4} \sum_{k=3}^{\log n} \frac{1}{k} \geq \lim_{n\to\infty} (\log\log n) = \infty.$$

For $p = d/n$, $d < 1$, $E(x)$ converges to a nonzero constant and with some nonzero probability, graphs will have a constant number of cycles independent of the size of the graph. For $d > 1$, $E(x)$ converges to infinity and a second moment argument shows that graphs will have an unbounded number of cycles increasing with n.

4.5.2 Full Connectivity

As p increases from $p = 0$, small components form. At $p = 1/n$ a giant component emerges and swallows up smaller components, starting with the larger components and ending up swallowing isolated vertices forming a single connected component at $p = \frac{\ln n}{n}$, at which point the graph becomes connected. We begin our development with a technical lemma.

Lemma 4.13 *The expected number of connected components of size k in $G(n,p)$ is at most*

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{kn-k^2}.$$

Property	Threshold
cycles	$1/n$
giant component	$1/n$
giant component + isolated vertices	$\frac{1}{2}\frac{\ln n}{n}$
connectivity, disappearance of isolated vertices	$\frac{\ln n}{n}$
diameter two	$\sqrt{\frac{2\ln n}{n}}$

Proof: The probability that k vertices form a connected component consists of the product of two probabilities. The first is the probability that the k vertices are connected, and the second is the probability that there are no edges out of the component to the remainder of the graph. The first probability is at most the sum over all spanning trees of the k vertices, that the edges of the spanning tree are present. The "at most" in the lemma statement is because $G(n,p)$ may contain more than one spanning tree on these nodes and, in this case, the union bound is higher than the actual probability. There are k^{k-2} spanning trees on k nodes. See Section **??** in the appendix. The probability of all the $k-1$ edges of one spanning tree being present is p^{k-1} and the probability that there are no edges connecting the k vertices to the remainder of the graph is $(1-p)^{k(n-k)}$. Thus, the probability of one particular set of k vertices forming a connected component is at most $k^{k-2}p^{k-1}(1-p)^{kn-k^2}$. Thus, the expected number of connected components of size k is $\binom{n}{k}k^{k-2}p^{k-1}(1-p)^{kn-k^2}$. ∎

We now prove that for $p=\frac{1}{2}\frac{\ln n}{n}$, the giant component has absorbed all small components except for isolated vertices.

Theorem 4.14 *Let $p=c\frac{\ln n}{n}$. For $c>1/2$, almost surely there are only isolated vertices and a giant component. For $c>1$, almost surely the graph is connected.*

Proof: We prove that almost surely for $c>1/2$, there is no connected component with k vertices for any k, $2\le k\le n/2$. This proves the first statement of the theorem since, if there were two or more components that are not isolated vertices, both of them could not be of size greater than $n/2$. The second statement that for $c>1$ the graph is connected then follows from Theorem 4.6 which states that isolated vertices disappear at $c=1$.

We now show that for $p=c\frac{\ln n}{n}$, the expected number of components of size k, $2\le k\le n/2$, is less than n^{1-2c} and thus for $c>1/2$ there are no components, except for isolated vertices and the giant component. Let x_k be the number of connected components of size k. Substitute $p=c\frac{\ln n}{n}$ into $\binom{n}{k}k^{k-2}p^{k-1}(1-p)^{kn-k^2}$ and simplify using $\binom{n}{k}\le(en/k)^k$, $1-p\le e^{-p}$, $k-1<k$, and $x=e^{\ln x}$ to get

$$E(x_k)\le\exp\left(\ln n+k+k\ln\ln n-2\ln k+k\ln c-ck\ln n+ck^2\frac{\ln n}{n}\right).$$

Keep in mind that the leading terms here for large k are the last two and, in fact, at $k = n$, they cancel each other so that our argument does not prove the fallacious statement for $c \geq 1$ that there is no connected component of size n, since there is. Let

$$f(k) = \ln n + k + k \ln \ln n - 2 \ln k + k \ln c - ck \ln n + ck^2 \frac{\ln n}{n}.$$

Differentiating with respect to k,

$$f'(k) = 1 + \ln \ln n - \frac{2}{k} + \ln c - c \ln n + \frac{2ck \ln n}{n}$$

and

$$f''(k) = \frac{2}{k^2} + \frac{2c \ln n}{n} > 0.$$

Thus, the function $f(k)$ attains its maximum over the range $[2, n/2]$ at one of the extreme points 2 or $n/2$. At $k = 2$, $f(2) \approx (1 - 2c) \ln n$ and at $k = n/2$, $f(n/2) \approx -c\frac{n}{4} \ln n$. So $f(k)$ is maximum at $k = 2$. For $k = 2$, $E(x)_k = e^{f(k)}$ is approximately $e^{(1-2c) \ln n} = n^{1-2c}$ and is geometrically falling as k increases from 2. At some point $E(x_k)$ starts to increase but never gets above $n^{-\frac{c}{4}n}$. Thus, the expected sum of the number of components of size k, for $2 \leq k \leq n/2$ is

$$E\left(\sum_{k=2}^{n/2} x_k\right) = O(n^{1-2c}).$$

This expected number goes to zero for $c > 1/2$ and the first-moment method implies that, almost surely, there are no components of size between 2 and $n/2$. This completes the proof of Theorem 4.14. ∎

4.5.3 Threshold for $O(\ln n)$ Diameter

We now show that within a constant factor of the threshold for graph connectivity, not only is the graph connected, but its diameter is $O(\ln n)$. That is, if p is $\Omega(\ln n/n)$, the diameter of $G(n, p)$ is $O(\ln n)$.

Consider a particular vertex v. Let S_i be the set of vertices at distance i from v. We argue that as i grows, $|S_1| + |S_2| + \cdots + |S_i|$ grows by a constant factor up to a size of $n/1000$. This implies that in $O(\ln n)$ steps, at least $n/1000$ vertices are connected to v. Then, there is a simple argument at the end of the proof of Theorem 4.16 that a pair of $n/1000$ sized subsets, connected to two different vertices v and w, have an edge between them.

Lemma 4.15 *Consider $G(n, p)$ for sufficiently large n with $p = c \ln n/n$ for any $c > 0$. Let S_i be the set of vertices at distance i from some fixed vertex v. If $|S_1|+|S_2|+\cdots+|S_i| \leq n/1000$, then*

$$\text{Prob}\left(|S_{i+1}| < 2(|S_1| + |S_2| + \cdots + |S_i|)\right) \leq e^{-10|S_i|}.$$

Proof: Let $|S_i| = k$. For each vertex u not in $S_1 \cup S_2 \cup \ldots \cup S_i$, the probability that u is not in S_{i+1} is $(1-p)^k$ and these events are independent. So, $|S_{i+1}|$ is the sum of $n-(|S_1|+|S_2|+\cdots+|S_i|)$ independent Bernoulli random variables, each with probability of

$$1-(1-p)^k \geq 1-e^{-ck\ln n/n}$$

of being one. Note that $n-(|S_1|+|S_2|+\cdots+|S_i|) \geq 999n/1000$. So,

$$E(|S_{i+1}|) \geq \frac{999n}{1000}(1-e^{-ck\frac{\ln n}{n}}).$$

Subtracting $200k$ from each side

$$E(|S_{i+1}|) - 200k \geq \frac{n}{2}\left(1-e^{-ck\frac{\ln n}{n}} - 400\frac{k}{n}\right).$$

Let $\alpha = \frac{k}{n}$ and $f(\alpha) = 1-e^{-c\alpha\ln n} - 400\alpha$. By differentiation $f''(\alpha) \leq 0$, so f is concave and the minimum value of f over the interval $[0, 1/1000]$ is attained at one of the end points. It is easy to check that both $f(0)$ and $f(1/1000)$ are greater than or equal to zero for sufficiently large n. Thus, f is nonnegative throughout the interval proving that $E(|S_{i+1}|) \geq 200|S_i|$. The lemma follows from Chernoff bounds. ∎

Theorem 4.16 *For $p \geq c\ln n/n$, where c is a sufficiently large constant, almost surely, $G(n,p)$ has diameter $O(\ln n)$.*

Proof: By Corollary 4.2, almost surely, the degree of every vertex is $\Omega(np) = \Omega(\ln n)$, which is at least $20\ln n$ for c sufficiently large. Assume this holds. So, for a fixed vertex v, S_1 as defined in Lemma 4.15 satisfies $|S_1| \geq 20\ln n$.

Let i_0 be the least i such that $|S_1|+|S_2|+\cdots+|S_i| > n/1000$. From Lemma 4.15 and the union bound, the probability that for some $i, 1 \leq i \leq i_0-1$, $|S_{i+1}| < 2(|S_1|+|S_2|+\cdots+|S_i|)$ is at most $\sum_{k=20\ln n}^{n/1000} e^{-10k} \leq 1/n^4$. So, with probability at least $1-(1/n^4)$, each S_{i+1} is at least double the sum of the previous S_j 's, which implies that in $O(\ln n)$ steps, i_0+1 is reached.

Consider any other vertex w. We wish to find a short $O(\ln n)$ length path between v and w. By the same argument as above, the number of vertices at distance $O(\ln n)$ from w is at least $n/1000$. To complete the argument, either these two sets intersect in which case we have found a path from v to w of length $O(\ln n)$ or they do not intersect. In the latter case, with high probability there is some edge between them. For a pair of disjoint sets of size at least $n/1000$, the probability that none of the possible $n^2/10^6$ or more edges between them is present is at most $(1-p)^{n^2/10^6} = e^{-\Omega(n\ln n)}$. There are at most 2^{2n} pairs of such sets and so the probability that there is some such pair with no edges is $e^{-\Omega(n\ln n)+O(n)} \to 0$. Note that there is no conditioning problem since we are arguing this for every pair of such sets. Think of whether such an argument made for just the n subsets of vertices, which are vertices at distance at most $O(\ln n)$ from a specific vertex, would work. ∎

4.6 Phase Transitions for Increasing Properties

For many graph properties such as connectivity, having no isolated vertices, having a cycle, etc., the probability of a graph having the property increases as edges are added to the graph. Such a property is called an increasing property. Q is an *increasing property* of graphs if when a graph G has the property, any graph obtained by adding edges to G must also have the property. In this section we show that any increasing property, in fact, has a threshold, although not necessarily a sharp one.

The notion of increasing property is defined in terms of adding edges. The following lemma proves that if Q is an increasing property, then increasing p in $G(n,p)$ increases the probability of the property Q.

Lemma 4.17 *If Q is an increasing property of graphs and $0 \leq p \leq q \leq 1$, then the probability that $G(n,q)$ has property Q is greater than or equal to the probability that $G(n,p)$ has property Q.*

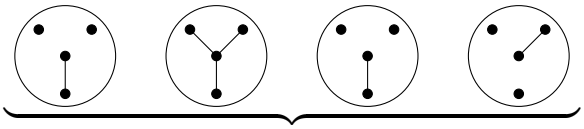
Proof: This proof uses an interesting relationship between $G(n,p)$ and $G(n,q)$. Generate $G(n,q)$ as follows. First generate $G(n,p)$. This means generating a graph on n vertices with edge probabilities p. Then, independently generate another graph $G\left(n, \frac{q-p}{1-p}\right)$ and take the union by putting in an edge if either of the two graphs has the edge. Call the resulting graph H. The graph H has the same distribution as $G(n,q)$. This follows since the probability that an edge is in H is $p + (1-p)\frac{q-p}{1-p} = q$, and, clearly, the edges of H are independent. The lemma follows since whenever $G(n,p)$ has the property Q, H also has the property Q. ∎

We now introduce a notion called *replication*. An m-fold replication of $G(n,p)$ is a random graph obtained as follows. Generate m independent copies of $G(n,p)$. Include an edge in the m-fold replication if the edge is in any one of the m copies of $G(n,p)$. The resulting random graph has the same distribution as $G(n,q)$ where $q = 1-(1-p)^m$ since the probability that a particular edge is not in the m-fold replication is the product of probabilities that it is not in any of the m copies of $G(n,p)$. If the m-fold replication of $G(n,p)$ does not have an increasing property Q, then none of the m copies of $G(n,p)$ has the property. The converse is not true. If no copy has the property, their union may have it. Since Q is an increasing property and $q = 1-(1-p)^m \leq 1-(1-mp) = mp$

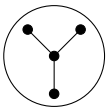
$$\text{Prob}\left(G(n,mp) \text{ has } Q\right) \geq \text{Prob}\left(G(n,q) \text{ has } Q\right) \tag{4.3}$$

We now show that any increasing property Q has a phase transition. The transition occurs at the point at which the probability that $G(n,p)$ has property Q is $\frac{1}{2}$. We will prove that for any function asymptotically less then $p(n)$ that the probability of having property Q goes to zero as n goes to infinity.

Theorem 4.18 *Every increasing property Q of $G(n,p)$ has a phase transition at $p(n)$, where for each n, $p(n)$ is the minimum real number a_n for which the probability that $G(n,a_n)$ has property Q is 1/2.*

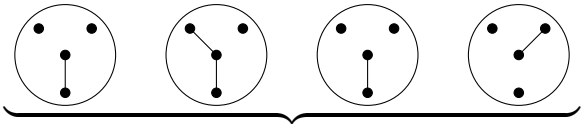


copies of G

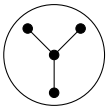


The m-fold replication H

If any graph has three or more edges, then the m-fold replication has three or more edges.



copies of G



The m-fold replication H

Even if no graph has three or more edges, the m-fold replication might have three or more edges.

Figure 4.12: The property that G has three or more edges is an increasing property. Let H be the m-fold replication of G. If any copy of G has three or more edges, H has three or more edges. However, H can have three or more edges even if no copy of G has three or more edges.

Proof: Let $p_0(n)$ be any function such that

$$\lim_{n\to\infty} \frac{p_0(n)}{p(n)} = 0.$$

We assert that almost surely $G(n, p_0)$ does not have the property Q. Suppose for contradiction, that this is not true. That is, the probability that $G(n, p_0)$ has the property Q does not converge to zero. By the definition of a limit, there exists $\varepsilon > 0$ for which the probability that $G(n, p_0)$ has property Q is at least ε on an infinite set I of n. Let $m = \lceil (1/\varepsilon) \rceil$. Let $G(n, q)$ be the m-fold replication of $G(n, p_0)$. The probability that $G(n, q)$ does not have Q is at most $(1 - \varepsilon)^m \leq e^{-1} \leq 1/2$ for all $n \in I$. For these n, by (4.3)

$$\text{Prob}(G(n, mp_0) \text{ has } Q) \geq \text{Prob}(G(n, q) \text{ has } Q) \geq 1/2.$$

Since $p(n)$ is the minimum real number a_n for which the probability that $G(n, a_n)$ has property Q is $1/2$, it must be that $mp_0(n) \geq p(n)$. This implies that $\frac{p_0(n)}{p(n)}$ is at least $1/m$ infinitely often, contradicting the hypothesis that $\lim_{n\to\infty} \frac{p_0(n)}{p(n)} = 0$.

A symmetric argument shows that for any $p_1(n)$ such that $\lim_{n\to\infty} \frac{p(n)}{p_1(n)} = 0$, $G(n, p_1)$ almost surely has property Q. ∎

4.7 Phase Transitions for CNF-sat

Phase transitions occur not only in random graphs, but in other random structures as well. An important example is that of satisfiability for a Boolean formula in conjunctive normal form.

Generate a random CNF formula f with n variables, m clauses, and k literals per clause. Each clause is picked independently with k literals picked uniformly at random from the set of $2n$ possible literals to form the clause. Here, the number of clauses n is going to infinity, m is a function of n, and k is a fixed constant. A reasonable value to think of for k is $k = 3$. A literal is a variable or its negation. Unsatisfiability is an increasing property since adding more clauses preserves unsatisfiability. By arguments similar to the last section, there is a phase transition, i.e., a function $m(n)$ such that if $m_1(n)$ is $o(m(n))$, a random formula with $m_1(n)$ clauses is, almost surely, satisfiable and for $m_2(n)$ with $m_2(n)/m(n) \to \infty$, a random formula with $m_2(n)$ clauses is, almost surely, unsatisfiable. It has been conjectured that there is a constant r_k independent of n such that $r_k n$ is a sharp threshold.

Here we derive upper and lower bounds on r_k. It is relatively easy to get an upper bound on r_k. A fixed truth assignment satisfies a random k clause with probability $1 - \frac{1}{2^k}$. Of the 2^k truth assignments to the k variables in the clause, only one fails to satisfy the clause. Thus, with probability $\frac{1}{2^k}$, the clause is not satisfied, and with probability $1 - \frac{1}{2^k}$, the clause is satisfied. Let $m = cn$. Now, cn independent clauses are all satisfied by the fixed assignment with probability $\left(1 - \frac{1}{2^k}\right)^{cn}$. Since there are 2^n truth assignments, the expected number of satisfying assignments for a formula with cn clauses is $2^n \left(1 - \frac{1}{2^k}\right)^{cn}$. If $c = 2^k \ln 2$, the expected number of satisfying assignments is

$$2^n \left(1 - \tfrac{1}{2^k}\right)^{n 2^k \ln 2}.$$

$\left(1 - \frac{1}{2^k}\right)^{2^k}$ is at most $1/e$ and approaches $1/e$ in the limit. Thus,

$$2^n \left(1 - \tfrac{1}{2^k}\right)^{n 2^k \ln 2} \leq 2^n e^{-n \ln 2} = 2^n 2^{-n} = 1.$$

For $c > 2^k \ln 2$, the expected number of satisfying assignments goes to zero as $n \to \infty$. Here the expectation is over the choice of clauses which is random, not the choice of a truth assignment. From the first moment method, it follows that a random formula with cn clauses is almost surely not satisfiable. Thus, $r_k \leq 2^k \ln 2$.

The other direction, showing a lower bound for r_k, is not that easy. From now on, we focus only on the case $k = 3$. The statements and algorithms given here can be extended to $k \geq 4$, but with different constants. It turns out that the second moment method cannot be directly applied to get a lower bound on r_3 because the variance is too high. A simple algorithm, called the Smallest Clause Heuristic (abbreviated SC), yields a satisfying assignment with probability tending to one if $c < \frac{2}{3}$, proving that $r_3 \geq \frac{2}{3}$. Other more

difficult to analyze algorithms, push the lower bound on r_3 higher.

The Smallest Clause Heuristic repeatedly executes the following. Assign true to a random literal in a random smallest length clause and delete the clause since it is now satisfied. Pick at random a 1-literal clause, if one exists, and set that literal to true. If there is no 1-literal clause, pick a 2-literal clause, select one of its two literals and set the literal to true. Otherwise, pick a 3-literal clause and a literal in it and set the literal to true. If we encounter a 0-length clause, then we have failed to find a satisfying assignment; otherwise, we have found one.

A related heuristic, called the Unit Clause Heuristic, selects a random clause with one literal, if there is one, and sets the literal in it to true. Otherwise, it picks a random as yet unset literal and sets it to true. The "pure literal" heuristic sets a random "pure literal", a literal whose negation does not occur in any clause, to true, if there are any pure literals; otherwise, it sets a random literal to true.

When a literal w is set to true, all clauses containing w are deleted, since they are satisfied, and $\bar{w}$ is deleted from any clause containing $\bar{w}$. If a clause is reduced to length zero (no literals), then the algorithm has failed to find a satisfying assignment to the formula. The formula may, in fact, be satisfiable, but the algorithm has failed.

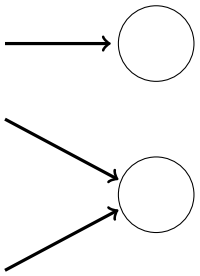
Example: Consider a 3-CNF formula with n variables and cn clauses. With n variables there are $2n$ literals, since a variable and its complement are distinct literals. The expected number of times a literal occurs is calculated as follows. Each clause has three literals. Thus, each of the $2n$ different literals occurs $\frac{(3cn)}{2n} = \frac{3}{2}c$ times on average. Suppose $c = 5$. Then each literal appears 7.5 times on average. If one sets a literal to true, one would expect to satisfy 7.5 clauses. However, this process is not repeatable since after setting a literal to true there is conditioning so that the formula is no longer random. ∎

4.8 Nonuniform and Growth Models of Random Graphs

4.8.1 Nonuniform Models

So far we have considered the random graph $G(n, p)$ in which all vertices have the same expected degree and showed that the degree is concentrated close to its expectation. However, large graphs occurring in the real world tend to have power law degree distributions. For a power law degree distribution, the number $f(d)$ of vertices of degree d plotted as a function of d satisfies $f(d) \leq c/d^{\alpha}$, where α and c are constants.

To generate such graphs, we stipulate that there are $f(d)$ vertices of degree d and choose uniformly at random from the set of graphs with this degree distribution. Clearly, in this model the graph edges are not independent and this makes these random graphs harder to analyze. But the question of when phase transitions occur in random graphs with arbitrary degree distributions is still of interest. In this section, we consider when



Consider a graph in which half of the vertices are degree one and half are degree two. If a vertex is selected at random, it is equally likely to be degree one or degree two. However, if we select an edge at random and walk to its endpoint, the vertex is twice as likely to be degree two as degree one. In many graph algorithms, a vertex is reached by randomly selecting an edge and traversing the edge to reach an endpoint. In this case, the probability of reaching a degree i vertex is proportional to $i\lambda_i$ where λ_i is the fraction of vertices that are degree i.

Figure 4.13: Probability of encountering a degree d vertex when following a path in a graph.

a random graph with a nonuniform degree distribution has a giant component. Our treatment in this section, and subsequent ones, will be more intuitive without providing rigorous proofs.

4.8.2 Giant Component in Random Graphs with Given Degree Distribution

Molloy and Reed address the issue of when a random graph with a nonuniform degree distribution has a giant component. Let λ_i be the fraction of vertices of degree i. There will be a giant component if and only if $\sum_{i=0}^{\infty} i(i-2)\lambda_i > 0$.

To see intuitively that this is the correct formula, consider exploring a component of a graph starting from a given seed vertex. Degree zero vertices do not occur except in the case where the vertex is the seed. If a degree one vertex is encountered, then that terminates the expansion along the edge into the vertex. Thus, we do not want to encounter too many degree one vertices. A degree two vertex is neutral in that the vertex is entered by one edge and left by the other. There is no net increase in the size of the frontier. Vertices of degree i greater than two increase the frontier by $i-2$ vertices. The vertex is entered by one of its edges and thus there are $i-1$ edges to new vertices in the frontier for a net gain of $i-2$. The $i\lambda_i$ in $i\,(i-2)\,\lambda_i$ is proportional to the probability of reaching a degree i vertex and the $i-2$ accounts for the increase or decrease in size of the frontier when a degree i vertex is reached.

Example: Consider applying the Molloy Reed conditions to the $G(n,p)$ model. The summation $\sum_{i=0}^{n} i(i-2)p_i$ gives value zero precisely when $p = 1/n$, the point at which the phase transition occurs. At $p = 1/n$, the average degree of each vertex is one and there are $n/2$ edges. However, the actual degree distribution of the vertices is binomial, where the probability that a vertex is of degree i is given by $p_i = \binom{n}{i} p^i (1-p)^{n-i}$. We now show that $\lim_{n\to\infty} \sum_{i=0}^{n} i(i-2)p_i = 0$ for $p_i = \binom{n}{i} p^i (1-p)^{n-i}$ when $p = 1/n$.

$$\lim_{n\to\infty} \sum_{i=0}^{n} i(i-2)\binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i}$$

$$= \lim_{n\to\infty} \sum_{i=0}^{n} i(i-2)\frac{n(n-1)\cdots(n-i+1)}{i!\,n^i}\left(1-\frac{1}{n}\right)^n\left(1-\frac{1}{n}\right)^{-i}$$

$$= \frac{1}{e}\lim_{n\to\infty} \sum_{i=0}^{n} i(i-2)\frac{n(n-1)\cdots(n-i+1)}{i!\,n^i}\left(\frac{n}{n-1}\right)^i$$

$$\leq \sum_{i=0}^{\infty}\frac{i(i-2)}{i!}.$$

To see that $\sum_{i=0}^{\infty}\frac{i(i-2)}{i!} = 0$, note that

$$\sum_{i=0}^{\infty}\frac{i}{i!} = \sum_{i=1}^{\infty}\frac{i}{i!} = \sum_{i=1}^{\infty}\frac{1}{(i-1)!} = \sum_{i=0}^{\infty}\frac{1}{i!}$$

and

$$\sum_{i=0}^{\infty}\frac{i^2}{i!} = \sum_{i=1}^{\infty}\frac{i}{(i-1)!} = \sum_{i=0}^{\infty}\frac{i+1}{i!} = \sum_{i=0}^{\infty}\frac{i}{i!} + \sum_{i=0}^{\infty}\frac{1}{i!} = 2\sum_{i=0}^{\infty}\frac{1}{i!}.$$

Thus,

$$\sum_{i=0}^{\infty}\frac{i(i-2)}{i!} = \sum_{i=0}^{\infty}\frac{i^2}{i!} - 2\sum_{i=0}^{\infty}\frac{i}{i!} = 0.$$

■

4.9 Growth Models

4.9.1 Growth Model With Preferential Attachment

Consider a growth model with preferential attachment. At each time unit, a vertex is added to the graph. Then with probability δ, an edge is attached to the new vertex and to a vertex selected at random with probability proportional to its degree. This model generates a tree with a power law distribution.

Let $d_i(t)$ be the expected degree of the i^{th} vertex at time t. The sum of the degrees of all vertices at time t is $2\delta t$ and thus the probability that an edge is connected to vertex i at time t is $\frac{d_i(t)}{2\delta t}$. The degree of vertex i is governed by the equation

$$\frac{\partial}{\partial t}d_i(t) = \delta\frac{d_i\,(t)}{2\delta t} = \frac{d_i(t)}{2t}$$

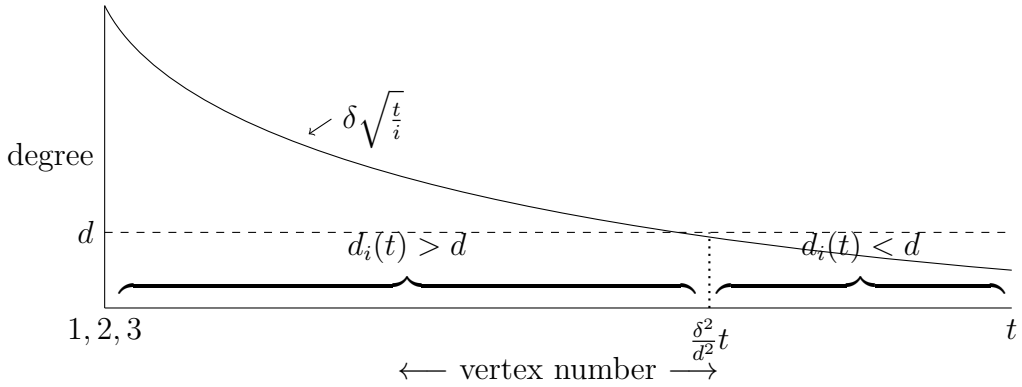


Figure 4.14: Illustration of degree of i^{th} vertex at time t. At time t, vertices numbered 1 to $\frac{\delta^2}{d^2}t$ have degrees greater than d.

where δ is the probability that an edge is added at time t and $\frac{d_i(t)}{2\delta t}$ is the probability that the vertex i is selected for the end point of the edge.

The two in the denominator governs the solution which is of the form $at^{\frac{1}{2}}$. The value of a is determined by the initial condition $d_i(t) = \delta$ at $t = i$. Thus, $\delta = ai^{\frac{1}{2}}$ or $a = \delta i^{-\frac{1}{2}}$. Hence, $d_i(t) = \delta\sqrt{\frac{t}{i}}$.

Next, we determine the probability distribution of vertex degrees. Now, $d_i(t)$ is less than d provided $i > \frac{\delta^2}{d^2}t$. The fraction of the t vertices at time t for which $i > \frac{\delta^2}{d^2}t$ and thus that the degree is less than d is $1 - \frac{\delta^2}{d^2}$. Hence, the probability that a vertex has degree less than d is $1 - \frac{\delta^2}{d^2}$. The probability density $P(d)$ satisfies

$$\int_0^d P(d)\partial d = \text{Prob(degree} < d) = 1 - \frac{\delta^2}{d^2}$$

and can be obtained from the derivative of Prob(degree $< d$).

$$P(d) = \frac{\partial}{\partial d}\left(1 - \frac{\delta^2}{d^2}\right) = 2\frac{\delta^2}{d^3},$$

a power law distribution.

4.10 Small World Graphs

In the 1960's, Stanley Milgram carried out an experiment that indicated that any two individuals in the United States were connected by a short sequence of acquaintances. Milgram would ask a source individual, say in Nebraska, to start a letter on its journey to a target individual in Massachusetts. The Nebraska individual would be given basic

information about the target including his address and occupation and asked to send the letter to someone he knew on a first name basis, who was closer to the target individual, in order to transmit the letter to the target in as few steps as possible. Each person receiving the letter would be given the same instructions. In successful experiments, it would take on average five to six steps for a letter to reach its target. This research generated the phrase "six degrees of separation" along with substantial research in social science on the interconnections between people. Surprisingly, there was no work on how to find the short paths using only local information.

In many situations, phenomena are modeled by graphs whose edges can be partitioned into local and long distance. We adopt a simple model of a directed graph due to Kleinberg, having local and long distance edges. Consider a 2-dimensional $n \times n$ grid where each vertex is connected to its four adjacent vertices. In addition to these local edges, there is one long distance edge out of each vertex. The probability that the long distance edge from vertex u terminates at v, $v \neq u$, is a function of the distance $d(u, v)$ from u to v. Here distance is measured by the shortest path consisting only of local grid edges. The probability is proportional to $1/d^r\,(u, v)$ for some constant r. This gives a one parameter family of random graphs. For r equal zero, $1/d^0\,(u, v) = 1$ for all u and v and thus the end of the long distance edge at u is uniformly distributed over all vertices independent of distance. As r increases the expected length of the long distance edge decreases. As r approaches infinity, there are no long distance edges and thus no paths shorter than that of the lattice path. What is interesting is that for r less than two, there are always short paths, but no local algorithm to find them. A local algorithm is an algorithm that is only allowed to remember the source, the destination, and its current location and can query the graph to find the long-distance edge at the current location. Based on this information, it decides the next vertex on the path.

The difficulty is that for $r < 2$, the end points of the long distance edges tend to be uniformly distributed over the vertices of the grid. Although short paths exist, it is unlikely on a short path to encounter a long distance edge whose end point is close to the destination. When r equals two, there are short paths and the simple algorithm that always selects the edge that ends closest to the destination will find a short path. For r greater than two, again there is no local algorithm to find a short path. Indeed, with high probability, there are no short paths at all.

The probability that the long distance edge from u goes to v is proportional to $d^{-r}(u, v)$. Note that the constant of proportionality will vary with the vertex u depending on where u is relative to the border of the $n \times n$ grid. However, the number of vertices at distance exactly k from u is at most $4k$ and for $k \leq n/2$ is at least k. Let $c_r(u) = \sum_v d^{-r}(u, v)$ be the normalizing constant. It is the inverse of the constant of proportionality.

$r > 2$ The lengths of long distance edges tend to be short so the probability of encountering a sufficiently long, long-distance edge is too low.

$r = 2$ Selecting the edge with end point closest to the destination finds a short path.

$r < 2$ The ends of long distance edges tend to be uniformly distributed. Short paths exist but a polylog length path is unlikely to encounter a long distance edge whose end point is close to the destination.

Figure 4.15: Effects of different values of r on the expected length of long distance edges and the ability to find short paths.

For $r > 2$, $c_r(u)$ is lower bounded by

$$c_r(u) = \sum_v d^{-r}(u,v) \geq \sum_{k=1}^{n/2}(k)k^{-r} = \sum_{k=1}^{n/2} k^{1-r} \geq 1.$$

No matter how large r is the first term of $\sum_{k=1}^{n/2} k^{1-r}$ is at least one.

For $r = 2$ the normalizing constant $c_r(u)$ is upper bounded by

$$c_r(u) = \sum_v d^{-r}(u,v) \leq \sum_{k=1}^{2n}(4k)k^{-2} \leq 4\sum_{k=1}^{2n}\frac{1}{k} = \theta(\ln n).$$

For $r < 2$, the normalizing constant $c_r(u)$ is lower bounded by

$$c_r(u) = \sum_v d^{-r}(u,v) \geq \sum_{k=1}^{n/2}(k)k^{-r} \geq \sum_{k=n/4}^{n/2} k^{1-r}.$$

The summation $\sum_{k=n/4}^{n/2} k^{1-r}$ has $\frac{n}{4}$ terms, the smallest of which is $\left(\frac{n}{4}\right)^{1-r}$ or $\left(\frac{n}{2}\right)^{1-r}$ depending on whether r is greater or less than one. This gives the following lower bound on $c_r(u)$.

$$c_r(u) \geq \frac{n}{4}\omega(n^{1-r}) = \omega(n^{2-r}).$$

No short paths exist for the $r > 2$ case.

For $r > 2$, we first show that for at least one half the pairs of vertices there is no short path between them. We begin by showing that the expected number of edges of length

greater than $n^{\frac{r+2}{2r}}$ goes to zero. The probability of an edge from u to v is $d^{-r}(u,v)/c_r(u)$ where $c_r(u)$ is lower bounded by a constant. Thus, the probability that a long edge is of length greater than or equal to $n^{\frac{r+2}{2r}}$ is upper bounded by some constant c times $\left(n^{\frac{r+2}{2r}}\right)^{-r}$ or $cn^{-(\frac{r+2}{2})}$. Since there are n^2 long edges, the expected number of edges of length at least $n^{\frac{r+2}{2r}}$ is at most $cn^2n^{-\frac{(r+2)}{2}}$ or $cn^{\frac{2-r}{2}}$, which for $r > 2$ goes to zero. Thus, by the first moment method, almost surely, there are no such edges.

For at least one half of the pairs of vertices, the grid distance, measured by grid edges between the vertices, is greater than or equal to $n/4$. Any path between them must have at least $\frac{1}{4}n/n^{\frac{r+2}{2r}} = \frac{1}{4}n^{\frac{r-2}{2r}}$ edges since there are no edges longer than $n^{\frac{r+2}{2r}}$ and so there is no polylog length path.

An algorithm for the $r = 2$ case

For $r = 2$, the local algorithm that selects the edge that ends closest to the destination t finds a path of expected length $O(\ln n)^3$. Suppose the algorithm is at a vertex u which is a at distance k from t. Then within an expected $O(\ln n)^2$ steps, the algorithm reaches a point at distance at most $k/2$. The reason is that there are $\Omega(k^2)$ vertices at distance at most $k/2$ from t. Each of these vertices is at distance at most $k + k/2 = O(k)$ from u. See Figure 4.16. Recall that the normalizing constant c_r is upper bounded by $O(\ln n)$, and hence, the constant of proportionality is lower bounded by some constant times $1/\ln n$. Thus, the probability that the long-distance edge from u goes to one of these vertices is at least

$$\Omega(k^2k^{-r}/\ln n) = \Omega(1/\ln n).$$

Consider $\Omega(\ln n)^2$ steps of the path from u. The long-distance edges from the points visited at these steps are chosen independently and each has probability $\Omega(1/\ln n)$ of reaching within $k/2$ of t. The probability that none of them does is

$$\big(1 - \Omega(1/\ln n)\big)^{c(\ln n)^2} = c_1e^{-\ln n} = \frac{c_1}{n}$$

for a suitable choice of constants. Thus, the distance to t is halved every $O(\ln n)^2$ steps and the algorithm reaches t in an expected $O(\ln n)^3$ steps.

A local algorithm cannot find short paths for the $r < 2$ case

For $r < 2$ no local polylog time algorithm exists for finding a short path. To illustrate the proof, we first give the proof for the special case $r = 0$, and then give the proof for $r < 2$.

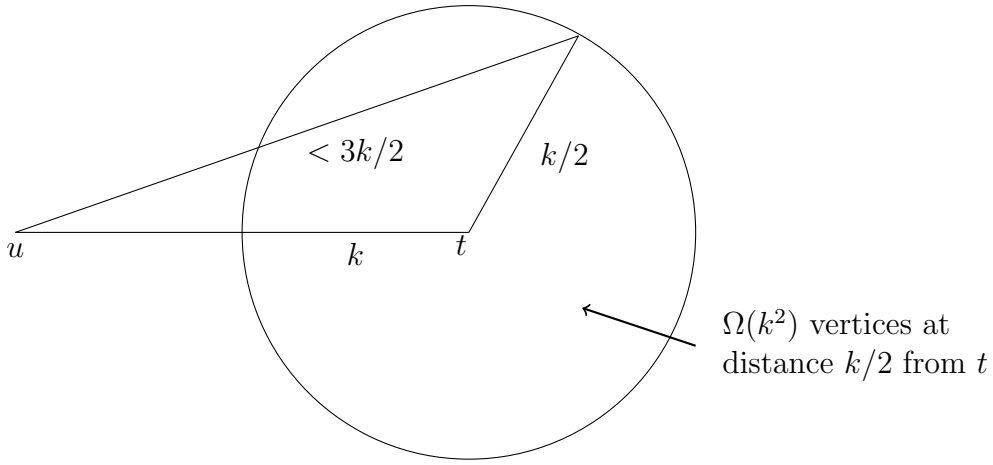


Figure 4.16: Small worlds.

When $r = 0$, all vertices are equally likely to be the end point of a long distance edge. Thus, the probability of a long distance edge hitting one of the n vertices that are within distance $\sqrt{n}$ of the destination is $1/n$. Along a path of length $\sqrt{n}$, the probability that the path does not encounter such an edge is $(1 - 1/n)^{\sqrt{n}}$. Now,

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^{\sqrt{n}} = \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^{n\frac{1}{\sqrt{n}}} = \lim_{n\to\infty} e^{-\frac{1}{\sqrt{n}}} = 1.$$

Since with probability $1/2$ the starting point is at distance at least $n/4$ from the destination and in $\sqrt{n}$ steps, the path will not encounter a long distance edge ending within distance $\sqrt{n}$ of the destination, for at least half of the starting points the path length will be at least $\sqrt{n}$. Thus, the expected time is at least $\frac{1}{2}\sqrt{n}$ and hence not in polylog time.

For the general $r < 2$ case, we show that a local algorithm cannot find paths of length $O(n^{(2-r)/4})$. Let $\delta = (2 - r)/4$ and suppose the algorithm finds a path with at most n^δ edges. There must be a long-distance edge on the path which terminates within distance n^δ of t; otherwise, the path would end in n^δ grid edges and would be too long. There are $O(n^{2\delta})$ vertices within distance n^δ of t and the probability that the long distance edge from one vertex of the path ends at one of these vertices is at most $n^{2\delta}\left(\frac{1}{n^{2-r}}\right) = n^{(r-2)/2}$. To see this, recall that the lower bound on the normalizing constant is $\theta(n^{2-r})$ and hence an upper bound on the probability of a long distance edge hitting v is $\theta\left(\frac{1}{n^{2-r}}\right)$ independent of where v is. Thus, the probability that the long distance edge from one of the n^δ vertices on the path hits any one of the $n^{2\delta}$ vertices within distance n^δ of t is $n^{2\delta}\frac{1}{n^{2-r}} = n^{\frac{r-2}{2}}$. The probability that this happens for any one of the n^δ vertices on the path is at most $n^{\frac{r-2}{2}}n^\delta = n^{\frac{r-2}{2}}n^{\frac{2-r}{4}} = n^{(r-2)/4} = o(1)$ as claimed.

Short paths exist for $r < 2$

Finally we show for $r < 2$ that there are $O(\ln n)$ length paths between s and t. The proof is similar to the proof of Theorem 4.16 showing $O(\ln n)$ diameter for $G(n,p)$ when p is $\Omega(\ln n/n)$, so we do not give all the details here. We give the proof only for the case when $r = 0$.

For a particular vertex v, let S_i denote the set of vertices at distance i from v. Using only local edges, if i is $O(\sqrt{\ln n})$, then $|S_i|$ is $\Omega(\ln n)$. For later i, we argue a constant factor growth in the size of S_i as in Theorem 4.16. As long as $|S_1|+|S_2|+\cdots+|S_i| \leq n^2/2$, for each of the $n^2/2$ or more vertices outside, the probability that the vertex is not in S_{i+1} is $(1-\frac{1}{n^2})^{|S_i|} \leq 1-\frac{|S_i|}{2n^2}$ since the long-distance edge from each vertex of S_i chooses a long-distance neighbor at random. So, the expected size of S_{i+1} is at least $|S_i|/4$ and using Chernoff, we get constant factor growth up to $n^2/2$. Thus, for any two vertices v and w, the number of vertices at distance $O(\ln n)$ from each is at least $n^2/2$. Any two sets of cardinality at least $n^2/2$ must intersect giving us a $O(\ln n)$ length path from v to w.

4.11 Bibliographic Notes

The $G(n,p)$ random graph model is from Erdös Rényi [**?**]. Among the books written on properties of random graphs a reader may wish to consult Palmer [**?**], Jansen, Luczak and Ruciński [**?**],or Bollobás [**?**]. Material on phase transitions can be found in [**?**]. The work on phase transitions for CNF was started by Chao and Franco [**?**]. Further work was done in [**?**], [**?**], [**?**], and others. The proof here that the SC algorithm produces a solution when the number of clauses is cn for $c < \frac{2}{3}$ is from [**?**].

For material on the giant component consult [**?**] or [**?**]. Material on branching process can be found in [**?**]. The phase transition for giant components in random graphs with given degree distributions is from Molloy and Reed [**?**].

There are numerous papers on growth models. The material in this chapter was based primarily on [**?**] and [**?**]. The material on small world is based on Kleinberg, [**?**] which follows earlier work by Watts and Strogatz [**?**].

4.12 Exercises

Exercise 4.1 *Search the World Wide Web to find some real world graphs in machine readable form or data bases that could automatically be converted to graphs.*

1. *Plot the degree distribution of each graph.*
2. *Compute the average degree of each graph.*
3. *Count the number of connected components of each size in each graph.*
4. *Describe what you find.*

Exercise 4.2 *Find a data base in machine readable form that can be viewed as a graph. What is the average vertex degree? If the graph were a $G(n,p)$ graph, what would the value of p be? Find the number of components of various sizes. Check that your work is correct by multiplying the number of components of size s by s and summing over all sizes. Is the sum equal to the total number of vertices? Examine the small components and see if any have cycles.*

Exercise 4.3 *In $G(n,p)$ the probability of a vertex having degree k is $\binom{n}{k}p^k(1-p)^{n-k}$.*

1. *Show by direct calculation that the expected degree is np.*
2. *Compute directly the variance of the distribution.*
3. *Where is the mode of the binomial distribution for a given value of p? The mode is the point at which the probability is maximum.*

Exercise 4.4

1. *Plot the degree distribution for $G(1000, 0.003)$.*
2. *Plot the degree distribution for $G(1000, 0.030)$.*

Exercise 4.5 *In $G\left(n, \frac{1}{n}\right)$, what is the probability that there is a vertex of degree $\log n$? Give an exact formula; also derive simple approximations.*

Exercise 4.6 *The example of Section 4.1.1 showed that if the degrees in $G(n, \frac{1}{n})$ were independent there would almost surely be a vertex of degree $\log n / \log\log n$. However, the degrees are not independent. Show how to overcome this difficulty.*

Exercise 4.7 *Let $f(n)$ be a function that is asymptotically less than n. Some such functions are $1/n$, a constant d, $\log n$ or $n^{\frac{1}{3}}$. Show that*

$$\left(1 + \tfrac{f(n)}{n}\right)^n \simeq e^{f(n)}.$$

for large n. That is

$$\lim_{n\to\infty} \frac{\left(1 + \frac{f(n)}{n}\right)^n}{e^{f(n)}} = 1.$$

Exercise 4.8

1. *In the limit as n goes to infinity, how does $\left(1-\frac{1}{n}\right)^{n \ln n}$ behave.*

2. *What is $\lim_{n\to\infty} \left(\frac{n+1}{n}\right)^n$?*

Exercise 4.9 *Consider a random permutation of the integers 1 to n. The integer i is said to be a fixed point of the permutation if i is the integer in the i^{th} position of the permutation. Use indicator variables to determine the expected number of fixed points in a random permutation.*

Exercise 4.10 *Generate a graph $G\left(n, \frac{d}{n}\right)$ with $n = 1000$ and d=2, 3, and 6. Count the number of triangles in each graph. Try the experiment with n=100.*

Exercise 4.11 *What is the expected number of squares (4-cycles) in $G\left(n, \frac{d}{n}\right)$? What is the expected number of 4-cliques in $G\left(n, \frac{d}{n}\right)$?*

Exercise 4.12 *Carry out an argument, similar to the one used for triangles, to show that $p = \frac{1}{n^{2/3}}$ is a threshold for the existence of a 4-clique. A 4-clique consists of four vertices with all $\binom{4}{2}$ edges present.*

Exercise 4.13 *What is the expected number of paths of length 3, $\log n$, $\sqrt{n}$, and $n-1$ in $G(n, \frac{d}{n})$? The expected number of paths of a given length being infinite does not imply that a graph selected at random has such a path.*

Exercise 4.14 *Consider $G(n, \frac{1}{2})$. Give an algorithm that with high probability will find*

1. *a clique of size $\log n$.*

2. *an independent set of size $\log n$. A set of vertices is an independent set if there is no edge between any pair of vertices in the set.*

3. *a subgraph[2] S in $G(n, \frac{1}{2})$, where S is any specified graph with $\log n$ vertices.*

Exercise 4.15 *Let x be an integer chosen uniformly at random from $\{1, 2, \ldots, n\}$. Count the number of distinct prime factors of n. The exercise is to show that the number of prime factors almost surely is $\Theta(\ln \ln n)$. Let p stand for a prime number between 2 and n.*

1. *For each fixed prime p, let I_p be the indicator function of the event that p divides x. Show that $E(I_p) = \frac{1}{p} + O\left(\frac{1}{n}\right)$. It is known that $\sum_{p \le n} \frac{1}{p} = \ln \ln n$ and you may assume this.*

[2] A subgraph of a graph is a subset of the vertices along with all the edges of the graph that connect pairs of vertices in the subset. Some books refer to this as an induced subgraph.

2. *The random variable of interest, $y = \sum_p I_p$, is the number of prime divisors of x picked at random. Show that the variance of y is $O(\ln \ln n)$. For this, assume the known result that the number of primes up to n is $O(n/\ln n)$. To bound the variance of y, think of what $E(I_p I_q)$ is for $p \neq q$, both primes.*

3. *Use (1) and (2) to prove that the number of prime factors is almost surely $\theta(\ln \ln n)$.*

Exercise 4.16 *Show for $\epsilon > 0$ that with high probability there exists a clique of size $(2-\epsilon)\log n$ in $G(n, \frac{1}{2})$, but no clique of size $2\log n$.*

Exercise 4.17 *Suppose one hides a clique of size k in a random graph $G\left(n, \frac{1}{2}\right)$. I.e., in the random graph, choose some subset S of k vertices and put in the missing edges to make S a clique. Presented with the modified graph, find S. The larger S is, the easier it should be to find. In fact, if k is more than $c\sqrt{n \ln n}$, then the clique leaves a telltale sign identifying S as the k vertices of largest degree. Prove this statement by appealing to Theorem 4.1. It remains a puzzling open problem to do this when k is smaller, say, $O(n^{1/3})$.*

Exercise 4.18 *The clique problem in a graph is to find the maximal size clique. This problem is known to be NP-hard and so a polynomial time algorithm is thought unlikely. We can ask the corresponding question about random graphs. For example, in $G\left(n, \frac{1}{2}\right)$ there almost surely is a clique of size $(2-\varepsilon)\log n$ for any $\varepsilon > 0$. But it is not known how to find one in polynomial time.*

1. *Show that in $G(n, \frac{1}{2})$, there are, almost surely, no cliques of size $2\log_2 n$.*

2. *Use the second moment method to show that in $G(n, \frac{1}{2})$, almost surely there are cliques of size $(2-\varepsilon)\log_2 n$.*

3. *Show that for any $\varepsilon > 0$, a clique of size $(2-\varepsilon)\log n$ can be found in $G\left(n, \frac{1}{2}\right)$ in time $n^{O(\ln n)}$.*

4. *Give an $O(n^2)$ algorithm for finding a clique of size $\Omega(\log n)$ in $G(n, \frac{1}{2})$. Hint: use a greedy algorithm. Apply your algorithm to $G\left(1000, \frac{1}{2}\right)$. What size clique do you find?*

5. *An independent set of vertices in a graph is a set of vertices, no two of which are connected by an edge. Give a polynomial time algorithm for finding an independent set in $G\left(n, \frac{1}{2}\right)$ of size $\Omega(\log n)$.*

Exercise 4.19 *Does there exist a copy of every subgraph with $(2-\epsilon)\log n$ vertices and $\frac{1}{4}\binom{(2-\epsilon)\log n}{2}$ edges in $G(n, \frac{1}{4})$?*

Exercise 4.20 *Given two instances, G_1 and G_2 of $G(n, \frac{1}{2})$, what is the largest subgraph common to both G_1 and G_2?*

Exercise 4.21 *(Birthday problem) What is the number of integers that must be drawn with replacement from a set of n integers so that some integer, almost surely, will be selected twice?*

Exercise 4.22 *Suppose the graph of a social network has 20,000 vertices. You have a program that starting from a random seed produces a community. A community is a set of vertices where each vertex in the set has more edges connecting it toother vertices in the set than to vertices outside of the set. In running the algorithm you find thousands of communities and wonder how many communities there are in the graph. Finally, when you find the* $10,000^{th}$ *community, it is a duplicate. It is the same community as one found earlier.*

1. *Use the birthday problem to derive a lower bound on the number of communities.*
2. *Why do you only get a lower bound and not a good estimate?*

Exercise 4.23 *To better understand the binomial distribution plot* $\binom{n}{k}p^k(1-p)^(n-k)$ *as a function of* k *for* $n = 50$ *and* $k = 0.05, 0.5, 0.95$. *For each value of* p *check the sum over all* k *to ensure that the sum is one.*

Exercise 4.24 *Consider the binomial distribution binomial* $\left(n, 1-\left(1-\frac{d}{n}\right)^i\right)$ *for* $d > 1$. *Here the distribution giving the probability of drawing* i *items is a different distribution for each value of* i.*Prove that as* $n \to \infty$, *the distribution goes to zero for all* i *except for* i *in the two ranges* $[0, c_1 \log n]$ *and* $[\theta n - c_2\sqrt{n}, \theta n + c_2\sqrt{n}]$.

Exercise 4.25 *Let* s *be the expected number of vertices discovered as a function of the number of steps* t *in a breadth first search of* $G\left(n, \frac{d}{n}\right)$. *Write a differential equation using expected values for the size of* s. *Show that the normalized size* $f = \frac{s-t}{n}$ *of the frontier is* $f(x) = 1 - e^{-dx} - x$ *where* $x = \frac{t}{n}$ *is the normalized time.*

Exercise 4.26 *The normalized frontier in a breadth first search of* $G(n, \frac{d}{n})$ *is* $f(x) = 1 - e^{-dx} - x$ *For* $d > 1$ *let* θ *be the unique root in* $(0, 1)$ *of* $1 - e^{-dx} - x = 0$. *Prove that the expected value of the size of the frontier increases varies with* i *for* i *in the neighborhood of* θ.

Exercise 4.27 *For* $f(x) = 1 - e^{-dx} - x$, *what is the value of* $x_{max} = \arg\max f(x)$*? What is the value of* $f(x_{max})$*? Where does the maximum expected value of the frontier of a breadth search in* $G(n, \frac{d}{n})$ *occur as a function of* n?

Exercise 4.28 *If* y *and* z *are independent, nonnegative random variables, then the generating function of the sum* $y + z$ *is the product of the generating function of* y *and* z. *Show that this follows from* $E(x^{y+z}) = E(x^y x^z) = E(x^y)E(x^z)$.

Exercise 4.29 *Let* $f_j(x)$ *be the* j^{th} *iterate of the generating function* $f(x)$ *of a branching process. When* $m > 1$, $lim_{j\to\infty} f_j(x) = q$ *for* $0 < x < 1$. *In the limit this implies* $Prob(z_j = 0) = q$ *and* $Prob(z_j = i) = 0$ *for all nonzero finite values of* i. *Shouldn't the probabilities add up to 1? Why is this not a contradiction?*

Exercise 4.30 *Try to create a probability distribution for a branching process which varies with the current population in which future generations neither die out, nor grow to infinity.*

Exercise 4.31 *Let d be a constant strictly greater than 1. Show that for a branching process with number of children distributed as Binomial$(n - c_1 n^{2/3}, \frac{d}{n})$, the root of the $f(x) = 1$ in $(0, 1)$ is at most a constant strictly less than 1.*

Exercise 4.32 *Randomly generate $G(50, p)$ for several values of p. Start with $p = \frac{1}{50}$.*

1. *For what value of p do cycles first appear?*
2. *For what value of p do isolated vertices disappear and the graphs become connected?*

Exercise 4.33 *Consider $G(n, p)$ with $p = \frac{1}{3n}$. Then, almost surely, there are no cycles of length 10.*

1. *Use the second moment method to show that, almost surely, there is a simple path of length 10.*
2. *What goes wrong if we try to modify the argument that, almost surely, there are no cycles of length 10 to show that there is no path of length 10?*

Exercise 4.34 *Complete the second moment argument of Theorem 4.12 to show that for $p = \frac{d}{n}$, $d > 1$, $G(n, p)$ almost surely has a cycle.*
Hint: If two cycles share one or more edges, then the union of the two cycles is at least one greater than the union of the vertices.

Exercise 4.35 *Let $G(n, p)$ be a random graph and let x be the random variable denoting the number of unordered pairs of nonadjacent vertices (u, v) such that no other vertex of G is adjacent to both u and v. Prove that if $\lim_{n\to\infty} E(x) = 0$, then for large n there are almost no disconnected graphs, i.e. Prob$(x = 0) \to 1$ and hence Prob$(G$ is connected$) \to 1$. Actually, the graph becomes connected long before this condition is true.*

Exercise 4.36 *Draw a tree with 10 vertices and label each vertex with a unique integer from 1 to 10. Construct the Prüfer sequence (Appendix **??**) for the tree. Given the Prüfer sequence, recreate the tree.*

Exercise 4.37 *Construct the tree corresponding to the following Prüfer sequences (Appendix **??**)*

1. *113663 (1,2),(1,3),(1,4),(3,5),(3,6),(6,7), and (6,8)*
2. *552833226.*

Exercise 4.38 *What is the expected number of isolated vertices in $G(n, p)$ for $p = \frac{1}{2}\frac{\ln n}{n}$?*

Exercise 4.39 *Theorem 4.16 shows that for some $c > 0$ and $p = c \ln n/n$, $G(n,p)$ has diameter $O(\ln n)$. Tighten the argument to pin down as low a value as possible for c.*

Exercise 4.40 *Let $f(n)$ be a function that is asymptotically less than n. Some such functions are $1/n$, a constant d, $\log n$ or $n^{\frac{1}{3}}$. Show that*

$$\left(1 + \tfrac{f(n)}{n}\right)^n \simeq e^{f(n)}.$$

for large n. That is

$$\lim_{n\to\infty} \frac{\left(1 + \frac{f(n)}{n}\right)^n}{e^{f(n)}} = 1.$$

Exercise 4.41 *What is diameter of G(n,p) for various values of p?*

Exercise 4.42

1. *List five increasing properties of $G(n,p)$.*
2. *List five non increasing properties .*

Exercise 4.43 *Consider generating the edges of a random graph by flipping two coins, one with probability p_1 of heads and the other with probability p_2 of heads. Add the edge to the graph if either coin comes down heads. What is the value of p for the generated $G(n,p)$ graph?*

Exercise 4.44 *In the proof of Theorem 4.18, we proved for $p_0(n)$ such that $\lim_{n\to\infty} \frac{p_0(n)}{p(n)} = 0$ that $G(n,p_0)$ almost surely did not have property Q. Give the symmetric argument that for any $p_1(n)$ such that $\lim_{n\to\infty} \frac{p(n)}{p_1(n)} = 0$, $G(n,p_1)$ almost surely has property Q.*

Exercise 4.45 *Consider a model of a random subset $N(n,p)$ of integers $\{1,2,\ldots n\}$ where, $N(n,p)$ is the set obtained by independently at random including each of $\{1,2,\ldots n\}$ into the set with probability p. Define what an "increasing property" of $N(n,p)$ means. Prove that every increasing property of $N(n,p)$ has a threshold.*

Exercise 4.46 *$N(n,p)$ is a model of a random subset of integers $\{1,2,\ldots n\}$ where, $N(n,p)$ is the set obtained by independently at random including each of $\{1,2,\ldots n\}$ into the set with probability p. What is the threshold for $N(n,p)$ to contain*

1. *a perfect square,*
2. *a perfect cube,*
3. *an even number,*
4. *three numbers such that $x + y = z$?*

Exercise 4.47 *Explain why the property, that $N(n,p)$ contains the integer 1, has a threshold. What is the threshold?*

Exercise 4.48 *Is there a condition such that any property satisfying the condition has a sharp threshold? For example, is monotonicity such a condition?*

Exercise 4.49 *The Sudoku game consists of a 9×9 array of squares. The array is partitioned into nine 3×3 squares. Each small square should be filled with an integer between 1 and 9 so that each row, each column, and each 3×3 square contains exactly one copy of each integer. Initially the board has some of the small squares filled in in such a way that there is exactly one way to complete the assignments of integers to squares. Some simple rules can be developed to fill in the remaining squares such as if the row and column containing a square already contain a copy of every integer except one, that integer should be placed in the square.*

Start with a 9×9 array of squares with each square containing a number between 1 and 9 such that no row, column, or 3×3 square has two copies of any integer.

1. *How many integers can you randomly erase and there still be only one way to correctly fill in the board?*

2. *Develop a set of simple rules for filling in squares such as if a row does not contain a given integer and if every column except one in which the square in the row is blank contains the integer, then place the integer in the remaining blank entry in the row. How many integers can you randomly erase and your rules will still completely fill in the board?*

Exercise 4.50 *Generalize the Sudoku game for arrays of size $n^2 \times n^2$. Develop a simple set of rules for completing the game. An example of a rule is the following. If the a row does not contain a given integer and if every column except one in which the square in the row is blank contains the integer, then place the integer in the remaining blank entry in the row. Start with a legitimate completed array and erase k entries at random.*

1. *Is there a threshold for the integer k such that if only k entries of the array are erased, your set of rules will find a solution?*

2. *Experimentally determine k for some large value of n.*

Exercise 4.51 *Let $\{x_i | 1 \leq i \leq n\}$, be a set of indicator variables with identical probability distributions. Let $x = \sum_{i=1}^{n} x_i$ and suppose $E(x) \to \infty$. Show that if the x_i are statistically independent, then $Prob(x = 0) \to 0$.*

Exercise 4.52 *In a square $n \times n$ grid, each of the $O(n^2)$ edges is randomly chosen to be present with probability p and absent with probability $1 - p$. Consider the increasing property that there is a path from the bottom left corner to the top right corner which always goes to the right or up. Show that $p = 1/2$ is a threshold for the property. Is it a sharp threshold?*

Exercise 4.53 *The threshold property seems to be related to uniform distributions. What if we considered other distributions? Consider a model where i is selected from the set $\{1, 2, \ldots, n\}$ with probability $\frac{c(n)}{i}$. Is there a threshold for perfect squares? Is there a threshold for arithmetic progressions?*

Exercise 4.54 *Modify the proof that every increasing property of $G(n, p)$ has a threshold to apply to the 3-CNF satisfiability problem.*

Exercise 4.55 *Evaluate $\left(1 - \frac{1}{2^k}\right)^{2^k}$ for k=3, 5, and 7. How close is it to 1/e?*

Exercise 4.56 *Randomly generate clauses for a Boolean formula in 3-CNF. Compute the number of solutions and the number of connected components of the solution set as a function of the number of clauses generated. What happens?*

Exercise 4.57 *Consider a random process for generating a Boolean function f in conjunctive normal form where each of c clauses is generated by placing each of n variables in the clause with probability p and complementing the variable with probability $1/2$. What is the distribution of clause sizes for various p such as $p = 3/n$, $1/2$, other values? Experimentally determine the threshold value of p for f to cease to be satisfied.*

Exercise 4.58 *For a random 3-CNF formula with n variables and cn clauses, what is the expected number of satisfying assignments?*

Exercise 4.59 *Which of the following variants of the SC algorithm admit a theorem like Theorem* **??***?*

1. *Among all clauses of least length, pick the first one in the order in which they appear in the formula.*

2. *Set the literal appearing in most clauses independent of length to 1.*

Exercise 4.60 *Suppose we have a queue of jobs serviced by one server. There is a total of n jobs in the system. At time t, each remaining job independently decides to join the queue to be serviced with probability $p = d/n$, where $d < 1$ is a constant. Each job has a processing time of 1 and at each time the server services one job, if the queue is nonempty. Show that with high probability, no job waits more than $\Omega(\ln n)$ time to be serviced once it joins the queue.*

Exercise 4.61 *Consider $G(n, p)$.*

1. *Where is phase transition for 2-colorability? Hint: For $p = d/n$ with $d < 1$, $G(n, p)$ is acyclic, so it is bipartite and hence 2-colorable. When $pn \to \infty$, the expected number of triangles goes to infinity. Show that, almost surely, there is a triangle? What does this do for 2-colorability?*

2. *What about 3-colorability?*

Exercise 4.62 *A vertex cover of size k for a graph is a set of k vertices such that one end of each edge is in the set. Experimentally play with the following problem. For $G(n, \frac{1}{2})$, for what value of k is there a vertex cover of size k?*

Exercise 4.63 *Consider graph 3-colorability. Randomly generate the edges of a graph and compute the number of solutions and the number of connected components of the solution set as a function of the number of edges generated. What happens?*

Exercise 4.64 *In $G(n,p)$, let x_k be the number of connected components of size k. Using x_k, write down the probability that a randomly chosen vertex is in a connected component of size k. Also write down the expected size of the connected component containing a randomly chosen vertex.*

Exercise 4.65 *For p asymptotically greater than $\frac{1}{n}$, show that*
$\sum_{i=0}^{\infty} i(i-2)\lambda_i > 0.$

Exercise 4.66 *Consider generating a random graph adding one edge at a time. Let n(i,t) be the number of components of size i at time t.*

$$
\begin{aligned}
&n(1,1) = n \\
&n(1,t) = 0 \quad t > 1 \\
&n(i,t) = n(i,t-1) + \sum \frac{j(i-j)}{n^2} n\left(j,t-1\right) n\left(i-j,t-1\right) - \frac{2i}{n} n\left(i\right)
\end{aligned}
$$

Compute n(i,t) for a number of values of i and t. What is the behavior? What is the sum of n(i,t) for fixed t and all i? Can you write a generating function for n(i,t)?

Exercise 4.67 *The global clustering coefficient of a graph is defined as follows. Let d_v be the degree of vertex v and let e_v be the number of edges connecting vertices adjacent to vertex v. The global clustering coefficient c is given by*

$$c = \sum_v \frac{2e_v}{d_v(d_v-1)}.$$

In a social network, for example, it measures what fraction of pairs of friends of each person are themselves friends. If many are, the clustering coefficient is high. What is c for a random graph with $p = \frac{d}{n}$? For a denser graph? Compare this value to that for some social network.

Exercise 4.68 *Consider a structured graph, such as a grid or cycle, and gradually add edges or reroute edges at random. Let L be the average distance between all pairs of vertices in a graph and let C be the ratio of triangles to connected sets of three vertices. Plot L and C as a function of the randomness introduced.*

Exercise 4.69 *Consider an $n \times n$ grid in the plane.*

1. *Prove that for any vertex u, there are at least k vertices at distance k for $1 \leq k \leq n/2$.*

2. *Prove that for any vertex u, there are at most $4k$ vertices at distance k.*

3. *Prove that for one half of the pairs of points, the distance between them is at least $4/4$.*

Exercise 4.70 *Show that in a small-world graph with $r \leq 2$, that there exist short paths with high probability. The proof for $r = 0$ is in the text.*

Exercise 4.71 *Change the small worlds graph as follows. Start with a $n \times n$ grid where each vertex has one long-distance edge to a vertex chosen uniformly at random. These are exactly like the long-distance edges for $r = 0$. But the grid edges are not present. Instead, we have some other graph with the property that for each vertex, there are $\Theta(t^2)$ vertices at distance t from the vertex for $t \leq n$. Show that, almost surely, the diameter is $O(\ln n)$.*

Exercise 4.72 *Given an n node directed graph with two random out edges from each node. For two vertices s and t chosen at random, prove that there exists a path of length at most $O(\ln n)$ from s to t with high probability.*

Exercise 4.73 *How does the diameter of a graph consisting of a cycle change as one adds a few random long distance edges? This question explores how much randomness is needed to get a small world.*

Exercise 4.74 *Ideas and diseases spread rapidly in small world graphs. What about spread of social contagion? A disease needs only one contact and with some probability transfers. Social contagion needs several contacts. How many vertices must one start with to spread social contagion, if the spread of contagion requires two adjacent vertices?*

Exercise 4.75 *How many edges are needed to disconnect a small world graph? By disconnect we mean at least two pieces each of reasonable size. Is this connected to the emergence of a giant component?*

Exercise 4.76 *In the small world model, would it help if the algorithm could look at edges at any node at a cost of one for each node looked at?*

Exercise 4.77 *Consider the $n \times n$ grid in the section on small world graphs. If the probability of an edge from vertex u to vertex v is proportional to $d^{-r}(u, v)$, show that the constant of proportionality $c_r(u)$ is*

$$\begin{array}{ll} \theta(n^{2-r}) & \text{for } r > 2 \\ \theta(\ln n) & \text{for } r = 2 \\ \theta(1) & \text{for } r < 2 \end{array}$$

Exercise 4.78 *In the $n \times n$ grid prove that for at least half of the pairs of vertices, the distance between the vertices is greater than or equal to $n/4$*

Exercise 4.79 *Show that for $r < 2$ in the small world graph model that short paths exist but a polylog length path is unlikely to encounter a long distance edge whose end point is close to the destination.*

Exercise 4.80 *Make a list of the ten most interesting things you learned about random graphs.*