8 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 on specific graphs. 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 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.

8.1 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. For example, if $p = d/n$ then the expected degree of a vertex of the graph is $(n-1)\frac{d}{n} \approx d$. In order to simplify calculations in this chapter, we will often use the approximation that $\frac{n-1}{n} \approx 1$. In fact, conceptually it is helpful to think of n as both the total number of vertices and as the number of potential neighbors of any given node, even though the latter is really $n-1$; for all our calculations, when n is large, the correction is just a low-order term.

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, there is a rapid transition 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 represent people and an edge means the two people it connects know each other. Given a chain of connections, such as A knows B, B knows C, C knows D, ..., and Y knows Z, we say that A indirectly knows Z. Thus, all people belonging to a connected component of the graph indirectly know each other. Suppose each pair of people, independent of other pairs, tosses a coin that comes up heads with probability $p = d/n$. If it is heads, they

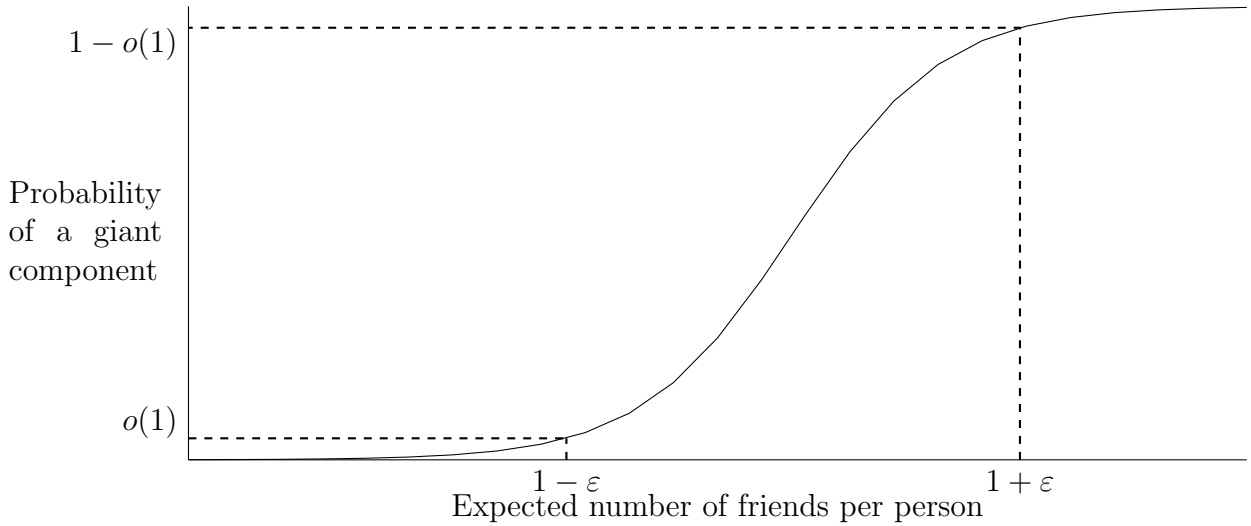


Figure 8.1: Probability of a giant component as a function of the expected number of people each person knows directly.

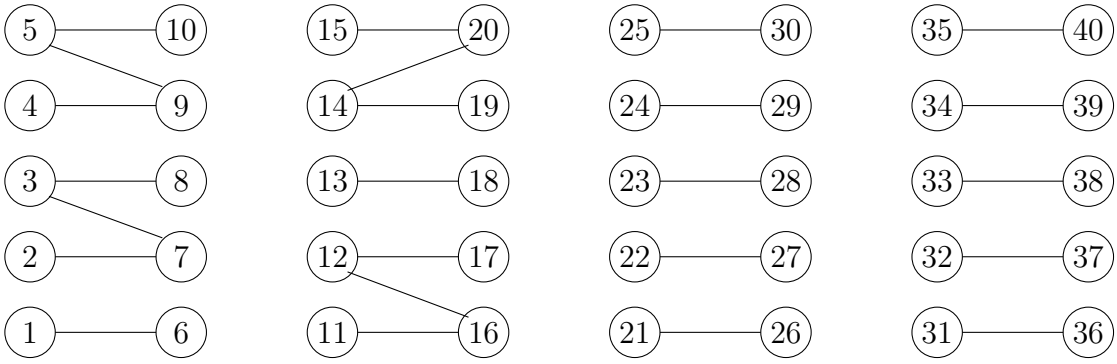
know each other; if it comes up tails, they don't. The value of d can be interpreted as the expected number of people a single person directly knows. The question arises as to how large are sets of people who indirectly know each other?

If the expected number of people each person knows is more than one, then a giant component of people, all of whom indirectly know each other, will be present consisting of a constant fraction of all the people. On the other hand, if in expectation, each person knows less than one person, the largest set of people who know each other indirectly 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 one to d slightly more than one. See Figure 8.1. Note that there is no global coordination of who knows whom. Each pair of individuals decides independently. Indeed, many large real-world graphs, with constant average degree, have a giant component. This is perhaps the most important global property of the $G(n, p)$ model.

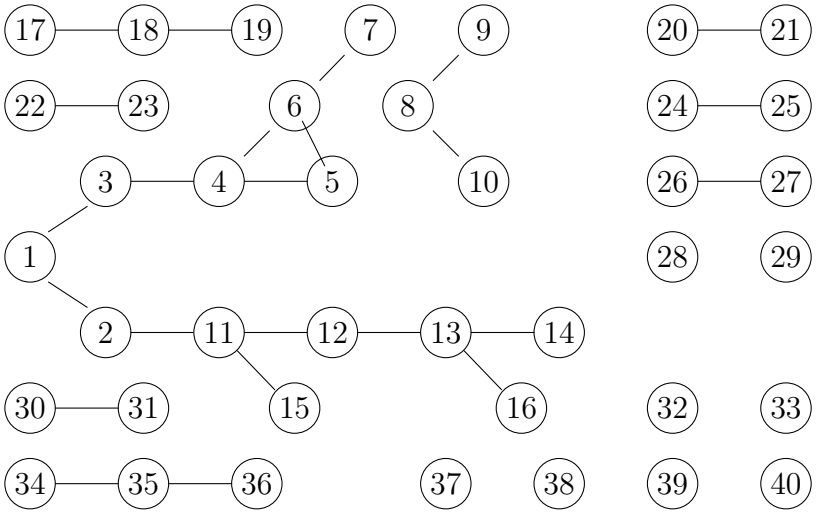
8.1.1 Degree Distribution

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 independent random variables, which results in a binomial distribution.

Example: In $G(n, \frac{1}{2})$, each vertex is of degree close to $n/2$. In fact, for any $\varepsilon > 0$, the degree of each vertex almost surely is within $1 \pm \varepsilon$ times $n/2$. To see this, note that the degree of a vertex is the sum of $n - 1 \approx n$ indicator variables that take on value one or



A graph with 40 vertices and 24 edges



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

Figure 8.2: 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.

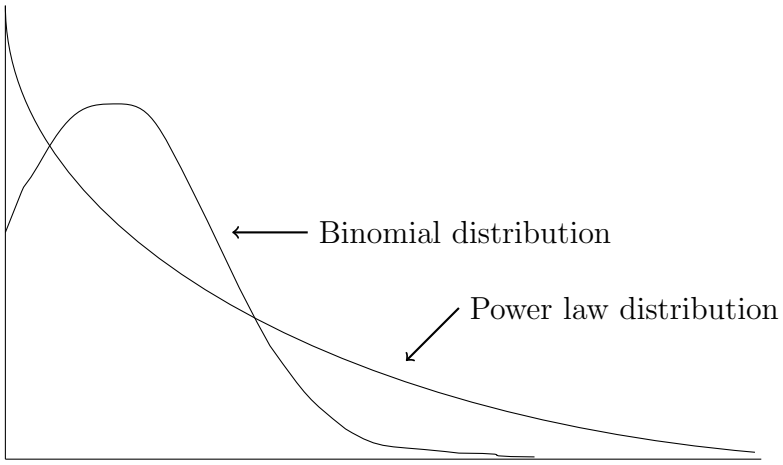


Figure 8.3: Illustration of the binomial and the power law distributions.

zero depending on whether the edge is present or not, each of mean $\frac{1}{2}$ and variance $\frac{1}{4}$. The expected value of the sum is the sum of the expected values and the variance of the sum is the sum of the variances, and hence the degree has mean $\approx \frac{n}{2}$ and variance $\approx \frac{n}{4}$. Thus, the probability mass is within an additive term of $\pm c\sqrt{n}$ of the mean for some constant c and thus within a multiplicative factor of $1 \pm \epsilon$ of $\frac{n}{2}$ for sufficiently large n. ∎

The degree distribution of $G(n,p)$ for general p is also binomial. Since p is the probability of an edge being present, the expected degree of a vertex is $p(n-1) \approx pn$. The 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 $\binom{n}{k}$ is the number of ways of choosing k edges, out of the possible n edges, and $p^k(1-p)^{n-k}$ is the probability that the k selected edges are present and the remaining $n-k$ are not.

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 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 states 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 by more than $\lambda\sqrt{np}$, drops off exponentially fast with λ.

Theorem 8.1 *Let v be a vertex of the random graph $G(n,p)$. Let α be a real number in $(0, \sqrt{np})$.*

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

Proof: The degree $\deg(v)$ 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. So, approximating $n-1$ with n, the theorem follows from Theorem 12.6 in the appendix. ∎

Although the probability that the degree of a single vertex differs significantly from its expected value drops exponentially, the statement that the degree of every vertex is close to its expected value requires that p is $\Omega(\frac{\log n}{n})$. That is, the expected degree grows at least logarithmically with the number of vertices.

Corollary 8.2 *Suppose ε is a positive constant. If $p \geq \frac{9\ln n}{n\varepsilon^2}$, then almost surely every vertex has degree in the range $(1-\varepsilon)np$ to $(1+\varepsilon)np$.*

Proof: Apply Theorem 8.1 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 $p \geq \frac{9\ln n}{n\varepsilon^2}$. ∎

Note that the assumption p is $\Omega(\frac{\log n}{n})$ is necessary. If $p = d/n$ for d a constant, then some vertices may well have degrees outside the range $[(1-\varepsilon)d, (1+\varepsilon)d]$. Indeed, shortly we will see that it is highly likely that for $p = \frac{1}{n}$ there is a vertex of degree $\Omega(\log n/\log\log n)$. Moreover, for $p = \frac{1}{n}$ it is easy to see that with high probability there will be at least one vertex of degree zero.

When p is a constant, the expected degree of vertices in $G(n,p)$ increases with n. In $G\left(n, \frac{1}{2}\right)$ the expected degree of a vertex is approximately $n/2$. 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. As n goes to infinity, the binomial distribution with $p = \frac{d}{n}$

$$\text{Prob}(k) = \binom{n}{k}\left(\frac{d}{n}\right)^k\left(1-\frac{d}{n}\right)^{n-k}$$

approaches the Poisson distribution

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

To see this, assume $k = o(n)$ and use the approximations $\binom{n}{k} \approx \frac{n^k}{k!}$, $n - k \approx n$, and $\left(1 - \frac{d}{n}\right)^{n-k} \approx \left(1 - \frac{d}{n}\right)^n \approx e^{-d}$. Then

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

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 once $k!$ dominates d^k. This justifies the $k = o(n)$ assumption. Thus, the Poisson distribution is a good approximation.

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

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

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

$$\log k^k = k\log k = \frac{\log n}{\log\log n}\left(\log\log n - \log\log\log n\right) \leq \log n$$

and thus $k^k \leq n$. Since $k! \leq k^k \leq n$, 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. ∎

8.1.2 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.

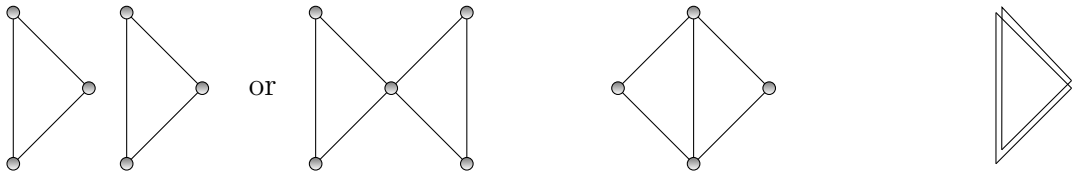
We wish to assert that with some nonzero probability there is at least one triangle in $G(n,p)$ when $p = \frac{d}{n}$. If all the triangles were on a small number of graphs, then the number of triangles in those graphs would far exceed the expected value and hence the variance would be high. A second moment argument rules out this scenario where a small fraction of graphs have a large number of triangles and the remaining graphs have none.

Let's 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

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

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

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



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

In Part 2, i, j, k and i', j', k' share two vertices and hence one edge. See Figure 8.4. 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 $O(n^4 p^5) = O(d^5/n)$ 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, putting all three parts together, we have:

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

which implies

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

For x to be 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,

$$\text{Prob}(x = 0) \leq \frac{\text{Var}(x)}{E^2(x)} \leq \frac{d^3/6 + o(1)}{d^6/36} \leq \frac{6}{d^3} + o(1). \tag{8.1}$$

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 $d < \sqrt[3]{6}$, $E(x) = \frac{d^3}{6} < 1$ and there simply are not enough edges in the graph for there to be a triangle.

8.2 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 $\frac{\ln n}{n}$. 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 8.1.

Probability	Transition
$p = o(\frac{1}{n})$	Forest of trees, no component of size greater than $O(\log n)$
$p = \frac{d}{n}, d < 1$	Cycles appear, no component of size greater than $O(\log n)$
$p = \frac{d}{n}, d = 1$	Components of size $O(n^{\frac{2}{3}})$
$p = \frac{d}{n}, d > 1$	Giant component plus $O(\log n)$ components
$p = \frac{1}{2}\frac{\ln n}{n}$	Giant component plus isolated vertices
$p = \sqrt{\frac{2\ln n}{n}}$	Diameter two
$p = \frac{\ln n}{n}$	Disappearance of isolated vertices Appearance of Hamilton circuit Diameter $O(\log n)$
$p = \frac{1}{2}$	Clique of size $(2-\epsilon)\ln n$

Table 1: Phase transitions

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

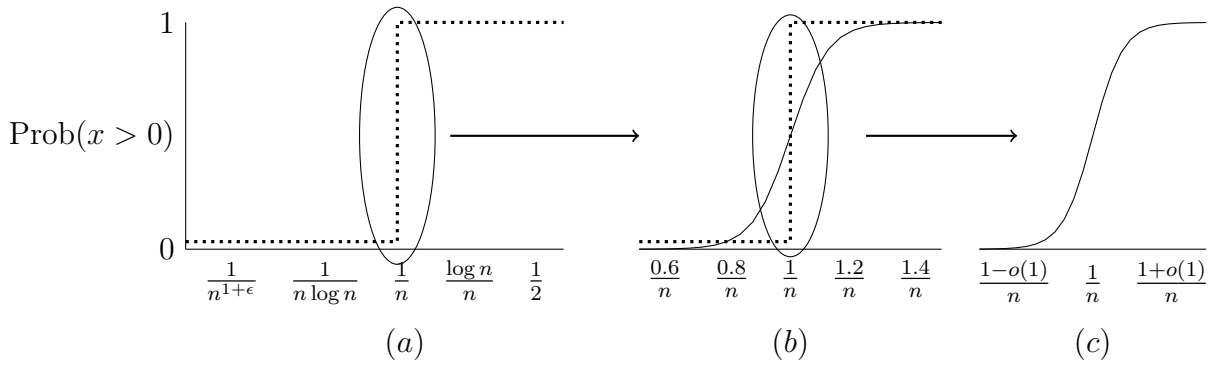


Figure 8.5: Figure 8.5(a) shows a phase transition at $p = \frac{1}{n}$. The dotted line shows an abrupt transition in Prob(x) from 0 to 1. For any function asymptotically less than $\frac{1}{n}$, Prob(x)>0 is zero and for any function asymptotically greater than $\frac{1}{n}$, Prob(x)>0 is one. Figure 8.5(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 8.5(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 $\frac{p_2(n)}{1/n} \to \infty$, then from (8.1), 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 vanishingly 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 8.3 (Second Moment method) *Let $x(n)$ be a random variable with $E(x) > 0$. If*

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

then x is almost surely greater than zero.

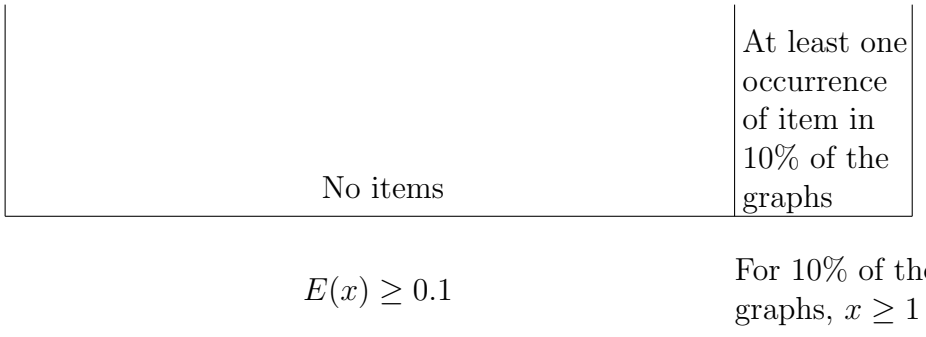


Figure 8.6: If the expected fraction of the number of graphs in which an item occurs did not go to zero, then $E(x)$, the expected number of items per graph, could not be zero. Suppose 10% of the graphs had at least one occurrence of the item. Then the expected number of occurrences per graph must be at least 0.1. Thus, $E(x) \to 0$ implies the probability that a graph has an occurrence of the item goes to zero. However, the other direction needs more work. If $E(x)$ is large, a second moment argument is needed to conclude that the probability that a graph picked at random has an occurrence of the item is nonnegligible, 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 $x > 0$.

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,

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

By Chebyshev inequality

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

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

Corollary 8.4 *Let x be a random variable with $E(x) > 0$. If*

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

then x is almost surely greater than zero.

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

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

∎

Second moment arguments are more difficult than first moment arguments since they deal with variance and without independence we do not have $E(xy) = E(x)E(y)$. In the triangle example, dependence occurs when two triangles share a common edge. However, if $p = \frac{d}{n}$, there are so few triangles that almost surely no two triangles share a common edge and the lack of statistical independence does not affect the answer. In looking for a phase transition, almost always the transition in probability of an item being present occurs when the expected number of items transitions.

Threshold for graph diameter two (two degrees of separation)

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. In other words, the property is that every pair of nodes has "at most two degrees of separation".

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 Hamiltonian 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 vanishingly small to unboundedly large. For $p < p_0$ almost surely a graph selected at random will not have a copy of the item. 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. So, what we will count is the number of pairs i and j that fail, i.e., the number of pairs i and j that have more than two degrees of separation. 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 8.5 *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 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 $c > \sqrt{2}$, $\lim_{n\to\infty} E(x) = 0$. 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 $\lim_{n\to\infty} E(x) = \infty$. 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.

$$E\left(x^2\right) = \sum_{\substack{i<j \\ k<l}} E\left(I_{ij}I_{kl}\right) + \sum_{\substack{\{i,j,k\} \\ i<j}} E\left(I_{ij}I_{ik}\right) + \sum_{i<j} E\left(I_{ij}^2\right). \tag{8.2}$$

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

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 not in $\{i, j, k, l\}$, at least one of the edges (i, u) or (j, u) is absent and, in addition, at least 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,

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

and the first sum is

$$\sum_{\substack{i<j \\ k<l}} E(I_{ij}I_{kl}) \le \frac{1}{4}n^{4-2c^2}\left(1 + o(1)\right),$$

where, the $\frac{1}{4}$ is because only a fourth of the 4-tupples (i, j, k, l) have $i < j$ and $k < l$.

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 (8.2).

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, $E\left(x^2\right) \le \frac{1}{4}n^{4-2c^2} + n^{3-2c^2} + n^{2-c^2}$ and $E\left(x\right) \cong \frac{1}{2}n^{2-c^2}$, from which it follows that for $c < \sqrt{2}$, $E\left(x^2\right) \le E^2\left(x\right)\left(1 + o(1)\right)$. By a second moment argument, Corollary 8.4, 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 $G(n,p)$ 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 8.6 *The disappearance of isolated vertices in $G(n,p)$ 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,

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

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

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

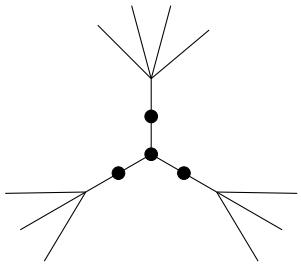


Figure 8.7: A degree three vertex with three adjacent degree two vertices. Graph cannot have a Hamilton circuit.

By the second moment argument, Corollary 8.4, 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. ∎

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 $E(x) \to 0$, we argued that with high probability, 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 high probability, 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.

A Hamilton circuit is a simple cycle that includes all the vertices. For example, in a graph of 4 vertices, there are three possible Hamilton circuits: $(1, 2, 3, 4)$, $(1, 2, 4, 3)$, and $(1, 3, 2, 4)$. Note that our graphs are undirected, so the circuit $(1, 2, 3, 4)$ is the same as the circuit $(1, 4, 3, 2)$.

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,

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

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 $\frac{1}{n} \log n$. For any $p(n)$ asymptotically greater, $G(n,p)$ 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 8.7. 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 with nonnegligible probability.

8.3 Giant Component

Consider $G(n,p)$ for $p = \frac{1+\epsilon}{n}$ where ϵ is a constant greater than zero. We now show that with high probability, such a graph contains a *giant component*, namely a component of size $\Omega(n)$. Moreover, with high probability, the graph contains only one such component, and all other components are much smaller, of size only $O(\log n)$. We begin by arguing existence of a giant component.

8.3.1 Existence of a giant component

To see that with high probability the graph has a giant component, do a depth first search (dfs) on $G(n,p)$ where $p = (1+\epsilon)/n$ with $0 < \epsilon < 1/8$. Note that it suffices to consider this range of ϵ since increasing the value of p only increases the probability that the graph has a giant component.

To perform the dfs, generate $\binom{n}{2}$ Bernoulli(p) independent random bits and answer

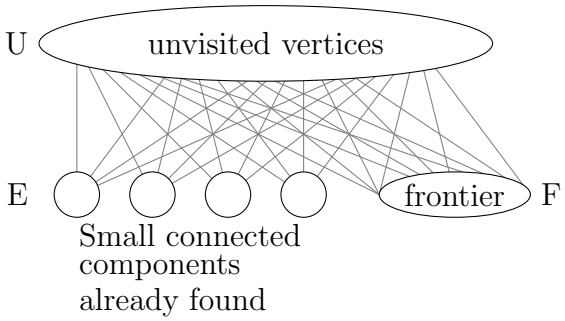


Figure 8.8: Picture after $\epsilon n^2/2$ edge queries. The potential edges from the small connected components to unvisited vertices do not exist in the graph. However, since many edges must have been found the frontier must be big and hence there is a giant component.

the t^{th} edge query according to the t^{th} bit. As the dfs proceeds, let

$E =$ set of fully explored vertices whose exploration is complete
$U =$ set of unvisited vertices
$F =$ frontier of visited and still being explored vertices .

Initially the set of fully explored vertices, E, and the frontier, F are empty and the set of unvisited vertices, U equals $\{1, 2, \ldots, n\}$. If the frontier is not empty and u is the active vertex of the dfs, the dfs queries each unvisited vertex in U until it finds a vertex v for which there is an edge (u, v) and moves v from U to the frontier and v becomes the active vertex. If no edge is found from u to an unvisited vertex in U, then u is moved from the frontier to the set of fully explored vertices E. If frontier is empty, the dfs moves an unvisited vertex from U to frontier and starts a new component. If both frontier and U are empty all connected components of G have been found. At any time all edges between the current fully explored vertices, E, and the current unvisited vertices, U, have been queried since a vertex is moved from the frontier to E only when there is no edge from the vertex to U.

Intuitively, after $\epsilon n^2/2$ edge queries a large number of edges must have been found since $p = \frac{1+\epsilon}{n}$. None of these can connect components already found with the set of unvisited vertices, and we will use this to show that with high probability the frontier must be large. Since the frontier will be in a connected component, a giant component exists with high probability. We first prove that after $\epsilon n^2/2$ edge queries the set of fully explored vertices is of size less than $n/3$.

Lemma 8.7 *After $\epsilon n^2/2$ edge queries, with high probability $|E| < n/3$.*

Proof: If not, at some $t \leq \varepsilon n^2/2$, $|E| = n/3$. A vertex is added to frontier only when an edge query is answered yes. So at time t, $|F|$ is less than or equal to the sum of $\varepsilon n^2/2$ Bernoulli(p) random variables, which with high probability is at most $\varepsilon n^2 p \leq n/3$. So,

at t, $|U| = n - |E| - |F| \geq n/3$. Since there are no edges between fully explored vertices and unvisited vertices, $|E|\ |U| \geq n^2/9$ edge queries must have already been answered in the negative. But $t > n^2/9$ contradicts $t \leq \epsilon n^2/2 \leq n^2/16$. Thus $|E| \leq n/3$. ∎

The frontier vertices in the search of a connected component are all in the component being searched. Thus if at any time the frontier set has $\Omega(n)$ vertices there is a giant component.

Lemma 8.8 *After $\epsilon n^2/2$ edge queries, with high probability the set F consists of at least $\epsilon^2 n/30$ vertices.*

Proof: After $\varepsilon n^2/2$ queries, say, $|F| < \varepsilon^2 n/30$. Thus

$$|U| = n - |E| - |F| = n - \frac{n}{3} - \frac{\epsilon^2 n}{30} \geq 1$$

and so the dfs is still active. Each positive answer to an edge query so far resulted in some vertex moving from U to F, which possibly later moved to E. The expected number of yes answers so far is $p\varepsilon n^2/2 = (1+\varepsilon)\varepsilon n/2$ and with high probability, the number of yes answers is at least $(\varepsilon n/2) + (\varepsilon^2 n/3)$. So,

$$|E| + |F| \geq \frac{\varepsilon n}{2} + \frac{\varepsilon^2 n}{3} \implies |E| \geq \frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}.$$

We must have $|E|\ |U| \leq \varepsilon n^2/2$. Now, $|E||U| = |E|(n - |E| - |F|)$ increases as $|E|$ increases from $\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}$ to $n/3$, so we have

$$|E||U| \geq \left(\frac{\varepsilon n}{2} + \frac{3\varepsilon^2 n}{10}\right)\left(n - \frac{\varepsilon n}{2} - \frac{3\varepsilon^2 n}{10} - \frac{\varepsilon^2 n}{30}\right) > \frac{\varepsilon n^2}{2},$$

a contradiction. ∎

8.3.2 No other large components

We now argue that for $p = (1+\epsilon)/n$ for constant $\epsilon > 0$, with high probability there is only one giant component, and in fact all other components have size $O(\log n)$.

We begin with a preliminary observation. Suppose that a $G(n,p)$ graph had at least a δ probability of having two (or more) components of size $\omega(\log n)$, i.e., asymptotically greater than $\log n$. Then, there would be at least a $\delta/2$ probability of the graph having two (or more) components with $\omega(\log n)$ vertices *inside the subset* $A = \{1, 2, \ldots, \epsilon n/2\}$. The reason is that an equivalent way to construct a graph $G(n,p)$ is to first create it in the usual way and then to randomly permute the vertices. Any component of size $\omega(\log n)$ will with high probability after permutation have at least an $\epsilon/4$ fraction of its vertices within the first $\epsilon n/2$. Thus, it suffices to prove that with high probability at most one component has $\omega(\log n)$ vertices within the set A to conclude that with high probability

the graph has only one component with $\omega(\log n)$ vertices overall.

We now prove that with high probability, a $G(n,p)$ graph for $p = (1+\epsilon)/n$ has at most one component with $\omega(\log n)$ vertices inside the set A. To do so, let B be the set of $(1-\epsilon/2)n$ vertices not in A. Now, construct the graph as follows. First, randomly flip coins of bias p to generate the edges within set A and the edges within set B. At this point, with high probability, B has at least one giant component, by the argument from Section 8.3.1, since $p = (1+\epsilon)/n \geq (1+\epsilon/4)/|B|$ for $0 < \epsilon \leq 1/2$. Let C^* be a giant component inside B. Now, flip coins of bias p to generate the edges between A and B *except* for those incident to C^*. At this point, let us name all components with $\omega(\log n)$ vertices inside A as $C_1, C_2, C_3, \ldots$. Finally, flip coins of bias p to generate the edges between A and C^*.

In the final step above, notice that with high probability, each C_i is connected to C^*. In particular, there are $\omega(n \log n)$ possible edges between any given C_i and C^*, each one of which is present with probability p. Thus the probability that this particular C_i is *not* connected to C^* is at most $(1-p)^{\omega(n \log n)} = 1/n^{\omega(1)}$. Thus, by the union bound, with high probability all such C_i are connected to C^*, and there is only one component with $\omega(\log n)$ vertices within A as desired.

8.3.3 The case of $p < 1/n$

When $p < 1/n$, then with high probability all components in $G(n,p)$ are of size $O(\log n)$. This is easiest to see by considering a variation on the above dfs that (a) begins with F containing a specific start vertex u_{start}, and then (b) when a vertex u is taken from F to explore, it pops u off of F, explores u fully by querying to find *all* edges between u and U, and then pushes the endpoints v of those edges onto F. Thus, this is like an explicit-stack version of dfs, compared to the previous recursive-call version of dfs. Let us call the exploration of such a vertex u a *step*. To make this process easier to analyze, let us say that if F ever becomes empty, we create a brand-new, fake "red vertex", connect it to each vertex in U with probability p, place the new red vertex into F, and then continue the dfs from there.

Let z_k denote the number of real (non-red) vertices discovered after k steps, not including u_{start}. For any given real vertex $u \neq u_{start}$, the probability that u is not discovered in k steps is $(1-p)^k$, and notice that these events are independent over the different vertices $u \neq u_{start}$. Therefore, the distribution of z_k is $\text{Binomial}\big(n-1, 1-(1-p)^k\big)$. Note that if $z_k < k$ then the process must have required creating a fake red vertex by step k, meaning that u_{start} is in a component of size at most k. Thus, it suffices to prove that $\text{Prob}(z_k \geq k) < 1/n^2$, for $k = c \ln n$ for a suitably large constant c, to then conclude by union bound over choices of u_{start} that with high probability *all* vertices are in components of size at most $c \ln n$.

To prove that $\text{Prob}(z_k \geq k) < 1/n^2$ for $k = c \ln n$, we use the fact that $(1-p)^k \geq 1-pk$ so $1-(1-p)^k \leq pk$. So, the probability that z_k is greater than or equal to k is at most the probability that a coin of bias pk flipped $n-1$ times will have at least k heads. But since $pk(n-1) \leq (1-\epsilon)k$ for some constant $\epsilon > 0$, by Chernoff bounds this probability is at most $e^{-c_0 k}$ for some constant $c_0 > 0$. When $k = c \ln n$ for a suitably large constant c, this probability is at most $1/n^2$, as desired.

8.4 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 when they form a connected component.

8.4.1 Emergence of Cycles

The emergence of cycles in $G(n,p)$ has a threshold when p equals to $1/n$. However, the threshold is not sharp.

Theorem 8.9 *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 $\binom{n}{k}\frac{(k-1)!}{2}$ possible 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 than $\frac{1}{n}$. A sharp threshold requires $E(x) \to 0$ for $p = d/n$, $d < 1$.

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	$\sqrt{\frac{2\ln n}{n}}$

Table 2: Thresholds for various properties

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. 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.

8.4.2 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 8.10 *The expected number of connected components of size k in $G(n,p)$ is at most*

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

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

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 8.11 *For $p = c\frac{\ln n}{n}$ with $c > 1/2$, 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 8.6 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 8.11. ∎

8.4.3 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 $p > c\frac{\ln n}{n}$ for sufficiently large constant c, the diameter of $G(n, p)$ is $O(\ln n)$ with high probability.

Consider a particular vertex v. Let S_i be the set of vertices at distance i from v. We argue that as i increases, with high probability $|S_1| + |S_2| + \cdots + |S_i|$ grows by at least a factor of two, 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 8.13 that a pair of $n/1000$ sized subsets, connected to two different vertices v and w, have an edge between them with high probability.

Lemma 8.12 *Consider $G(n, p)$ for sufficiently large n with $p = c\frac{\ln n}{n}$ 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 8.13 *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 8.2, 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 that this holds. So, for a fixed vertex v, S_1 as defined in Lemma 8.12 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 8.12 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. ∎

8.5 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 has a threshold, although not necessarily a sharp one.

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

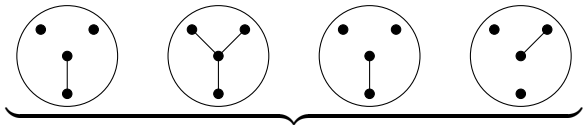
Lemma 8.14 *If Q is an increasing property of graphs and $0 \leq p \leq q \leq 1$, then the probability that $G(n,q)$ has property Q is greater than or equal to the probability that $G(n,p)$ has property Q.*

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 including 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 $G(n,p)$ on the same set of vertices. 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 $q = 1-(1-p)^m \leq 1-(1-mp) = mp$

$$\text{Prob}\left(G(n,mp) \text{ has } Q\right) \geq \text{Prob}\left(G(n,q) \text{ has } Q\right) \tag{8.3}$$

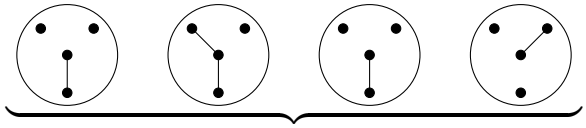
We now show that every increasing property Q has a phase transition. The transition occurs at the point $p(n)$ at which the probability that $G(n,p(n))$ has property Q is $\frac{1}{2}$. We will prove that for any function asymptotically less then $p(n)$ that the probability of having property Q goes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



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 8.9: 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.

Theorem 8.15 *Each 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.*

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 (11.4)

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

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. ∎

8.6 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 is 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.

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.

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 12.9 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 $f^i(x)$.

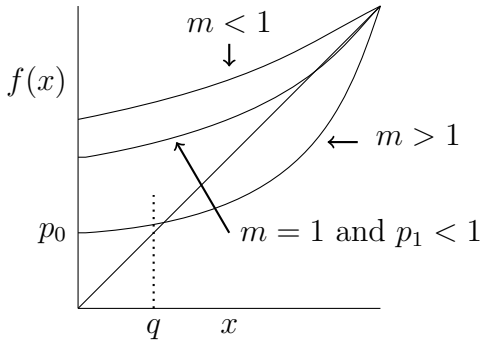


Figure 8.10: Illustration of the root of equation $f(x) = x$ in the interval [0,1).

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 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

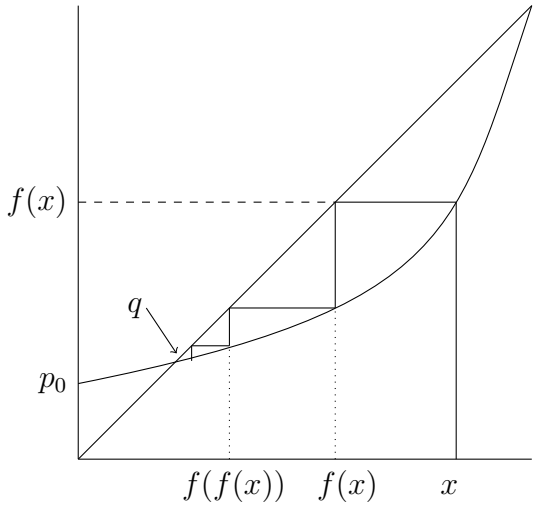


Figure 8.11: Illustration of convergence of the sequence of iterations $f_1(x), f_2(x), \ldots$ to q.

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 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 8.11 illustrates the proof which is given in Lemma 8.16. Similarly note that when $m < 1$ or $m = 1$ with $p_0 < 1$, $f_j(x)$ approaches one as j approaches infinity.

Lemma 8.16 *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, $0 \leq x < 1$. That is

$$\lim_{j \to \infty} f_j(x) = q + 0x + 0x^2 + \cdots$$

and there are no children with probability q and no finite number of children with probability zero. ∎

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.

Theorem 8.17 *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 can be viewed as the process of creating 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. For example, suppose that with probability $\frac{1}{2}$ it became extinct with size 3, with probability $\frac{1}{4}$ it became extinct with size 9, with probability $\frac{1}{8}$ it became extinct with size 27, etc. In such a case the expected size at extinction would be infinite even though the process dies out with probability one. We now show this does not happen.

Lemma 8.18 *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

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

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

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

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

The fact that $f'(q) < 1$ is illustrated in Figure 8.10. 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. ∎

8.7 CNF-SAT

Phase transitions occur not only in random graphs, but in other random structures as well. An important example is that of satisfiability of Boolean formulas in conjunctive normal form. A conjunctive normal form (CNF) formula over n variables $x_1, \ldots, x_n$ is an AND of ORs of *literals*, where a literal is a variable or its negation. For example, the following is a CNF formula over the variables $\{x_1, x_2, x_3, x_4\}$:

$$(x_1 \vee \bar{x}_2 \vee x_3)(x_2 \vee \bar{x}_4)(x_1 \vee x_4)(x_3 \vee x_4)(x_2 \vee \bar{x}_3 \vee x_4).$$

Each OR of literals is called a *clause*; for example, the above formula has five clauses. A *k-CNF* formula is a CNF formula in which each clause has size at most k, so the above formula is a 3-CNF formula. An assignment of true/false values to variables is said to

satisfy a CNF formula if it satisfies every clause in it. Setting all variables to true satisfies the above CNF formula, and in fact this formula has multiple satisfying assignments. A formula is said to be *satisfiable* it there exists at least one assignment of truth values to variables that satisfies it.

Many important problems can be converted into questions of finding satisfying assignments of CNF formulas. Indeed, the CNF-SAT problem of whether a given CNF formula is satisfiable is *NP-Complete*, meaning that any problem in the class NP can be converted into it. As a result, it is believed to be highly unlikely that there will ever exist an efficient algorithm for worst-case instances. However, there are solvers that turn out to work very well in practice on instances arising from a wide range of applications. There is also substantial structure and understanding of the satisfiability of *random* CNF formulas. The next two sections discuss each in turn.

8.7.1 SAT-solvers in practice

While the SAT problem is NP-complete, a number of algorithms have been developed that perform extremely well in practice on SAT formulas arising in a range of applications. Such applications include hardware and software verification, creating action plans for robots and robot teams, solving combinatorial puzzles, and even proving mathematical theorems.

Broadly, there are two classes of solvers: *complete* solvers and *incomplete* solvers. Complete solvers are guaranteed to find a satisfying assignment whenever one exists; if they do not return a solution, then you know the formula is not satisfiable. Complete solvers are often based on some form of recursive tree search. Incomplete solvers instead make a "best effort"; they are typically based on some local-search heuristic, and they may fail to output a solution even when a formula is satisfiable. However, they are typically much faster than complete solvers.

An example of a complete solver is the following DPLL (Davis-Putnam-Logemann-Loveland) style procedure. First, if there are any variables x_i that never appear in negated form in any clause, then set those variables to true and delete clauses where the literal x_i appears. Similarly, if there are any x_i that *only* appear in negated form, then set those variables to false and delete clauses where the literal $\bar{x}_i$ appears. Second, if there are any clauses that have only one literal in them (such clauses are called unit clauses), then set that literal as needed to satisfy the clause. E.g., if the clause was "$(\bar{x}_3)$" then one would set x_3 to false. Then remove that clause along with any other clause containing that literal, and shrink any clause containing the negation of that literal (e.g., a clause such as $(x_3 \vee x_4)$ would now become just (x_4), and one would then run this rule again on this clause). Finally, if neither of the above two cases applies, then one chooses some literal and recursively tries both settings for it. Specifically, choose some literal ℓ and recursively check if the formula is satisfiable conditioned on setting ℓ to true; if the answer

is "yes" then we are done, but if the answer is "no" then recursively check if the formula is satisfiable conditioned on setting ℓ to false. Notice that this procedure is guaranteed to find a satisfying assignment whenever one exists.

An example of an incomplete solver is the following local-search procedure called *Walksat*. Walksat begins with a random assignment of truth-values to variables. If this happens to satisfy the formula, then it outputs success. If not, then it chooses some unsatisfied clause C at random. If C contains some variable x_i whose truth-value can be flipped (causing C to be satisfied) without causing any *other* clause to be unsatisfied, then x_i's truth-value is flipped. Otherwise, Walksat either (a) flips the truth-value of the variable in C that causes the *fewest* other clauses to become unsatisfied, or else (b) flips the truth-value of a *random* x_i in C; the choice of whether to perform (a) or (b) is determined by flipping a coin of bias p. Thus, Walksat is performing a kind of random walk in the space of truth-assignments, hence the name. Walksat also has two time-thresholds T_{flips} and $T_{restarts}$. If the above procedure has not found a satisfying assignment after T_{flips} flips, it then restarts with a fresh initial random assignment and tries again; if that entire process has not found a satisfying assignment after $T_{restarts}$ restarts, then it outputs "no assignment found".

The above solvers are just two simple examples. Due to the importance of the CNF-SAT problem, development of faster SAT-solvers is an active area of computer science research. SAT-solving competitions are held each year, and solvers are routinely being used to solve challenging verification, planning, and scheduling problems.

8.7.2 Phase Transitions for CNF-SAT

We now consider the question of phase transitions in the satisfiability of *random* k-CNF formulas.

Generate a random CNF formula f with n variables, m clauses, and k literals per clause, where recall that a literal is a variable or its negation. Specifically, each clause in f is selected independently at random from the set of all $\binom{n}{k}2^k$ possible clauses of size k. Equivalently, to generate a clause, choose a random set of k distinct variables, and then for each of those variables choose to either negate it or not with equal probability. Here, the number of variables 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 $k = 3$. Unsatisfiability is an increasing property since adding more clauses preserves unsatisfiability. By arguments similar to Section 8.5, 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 $1 - \frac{1}{2^k}$ because 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

$$2^n \left(1 - \tfrac{1}{2^k}\right)^{n2^k \ln 2}.$$

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

$$2^n \left(1 - \tfrac{1}{2^k}\right)^{n2^k \ln 2} \leq 2^n e^{-n \ln 2} = 2^n 2^{-n} = 1.$$

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 shortest clause and delete the clause since it is now satisfied. In more detail, 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. Another variation is the "pure literal" heuristic. It 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. ∎

Theorem 8.19 *If the number of clauses in a random 3-CNF formula grows as cn where c is a constant less than 2/3, then with probability* $1 - o(1)$*, the Shortest Clause (SC) Heuristic finds a satisfying assignment.*

The proof of this theorem will take the rest of the section. A general impediment to proving that simple algorithms work for random instances of many problems is conditioning. At the start, the input is random and has properties enjoyed by random instances. But, as the algorithm is executed, the data is no longer random; it is conditioned on the steps of the algorithm so far. In the case of SC and other heuristics for finding a satisfying assignment for a Boolean formula, the argument to deal with conditioning is relatively simple.

We supply some intuition before giving the proof. Imagine maintaining a queue of 1 and 2-clauses. A 3-clause enters the queue when one of its literals is set to false and it becomes a 2-clause. SC always picks a 1 or 2-clause if there is one and sets one of its literals to true. At any step when the total number of 1 and 2-clauses is positive, one of the clauses is removed from the queue. Consider the arrival rate, that is, the expected number of arrivals into the queue at a given time t. For a particular clause to arrive into the queue at time t to become a 2-clause, it must contain the negation of the literal being set to true at time t. It can contain any two other literals not yet set. The number of such clauses is $\binom{n-t}{2}2^2$. So, the probability that a particular clause arrives in the queue at time t is at most

$$\frac{\binom{n-t}{2}2^2}{\binom{n}{3}2^3} \leq \frac{3}{2(n-2)}.$$

Since there are cn clauses in total, the arrival rate is $\frac{3c}{2}$, which for $c < 2/3$ is a constant strictly less than one. The arrivals into the queue of different clauses occur independently (Lemma 8.20), the queue has arrival rate strictly less than one, and the queue loses one or more clauses whenever it is nonempty. This implies that the queue never has too many clauses in it. A slightly more complicated argument will show that no clause remains as a 1 or 2-clause for $\omega(\ln n)$ steps (Lemma 8.21). This implies that the probability of two contradictory 1-length clauses, which is a precursor to a 0-length clause, is very small.

Lemma 8.20 *Let T_i be the first time that clause i turns into a 2-clause. T_i is ∞ if clause i gets satisfied before turning into a 2-clause. The T_i are mutually independent over the randomness in constructing the formula and the randomness in SC, and for any t,*

$$Prob(T_i = t) \leq \frac{3}{2(n-2)}.$$

Proof: For the proof, generate the clauses in a different way. The important thing is that the new method of generation, called the method of "deferred decisions", results in the same distribution of input formulae as the original. The method of deferred decisions is tied in with the SC algorithm and works as follows. At any time, the length of each clause (number of literals) is all that we know; we have not yet picked which literals are in each clause. At the start, every clause has length three and SC picks one of the clauses uniformly at random. Now, SC wants to pick one of the three literals in that clause to set to true, but we do not know which literals are in the clause. At this point, we pick uniformly at random one of the $2n$ possible literals. Say for illustration, we picked $\bar{x}_{102}$. The literal $\bar{x}_{102}$ is placed in the clause and set to true. The literal x_{102} is set to false. We must also deal with occurrences of the literal or its negation in all other clauses, but again, we do not know which clauses have such an occurrence. We decide that now. For each clause, independently, with probability $3/n$ include either the literal $\bar{x}_{102}$ or its negation x_{102}, each with probability $1/2$. In the case that we included $\bar{x}_{102}$ (the literal we had set to true), the clause is now deleted, and if we included x_{102} (the literal we had set to false), we decrease the residual length of the clause by one.

At a general stage, suppose the fates of i variables have already been decided and $n - i$ remain. The residual length of each clause is known. Among the clauses that are not yet satisfied, choose a random shortest length clause. Among the $n - i$ variables remaining, pick one uniformly at random, then pick it or its negation as the new literal. Include this literal in the clause thereby satisfying it. Since the clause is satisfied, the algorithm deletes it. For each other clause, do the following. If its residual length is l, decide with probability $l/(n - i)$ to include the new variable in the clause and if so with probability $1/2$ each, include it or its negation. If the literal that was set to true is included in a clause, delete the clause as it is now satisfied. If its negation is included in a clause, then just delete the literal and decrease the residual length of the clause by one.

Why does this yield the same distribution as the original one? First, observe that the order in which the variables are picked by the method of deferred decisions is independent of the clauses; it is just a random permutation of the n variables. Look at any one clause. For a clause, we decide in order whether each variable or its negation is in the clause. So for a particular clause and a particular triple i, j, and k with $i < j < k$, the probability that the clause contains the i^{th}, the j^{th}, and k^{th} literal (or their negations) in the order

determined by deferred decisions is:

$$\left(1-\tfrac{3}{n}\right)\left(1-\tfrac{3}{n-1}\right)\cdots\left(1-\tfrac{3}{n-i+2}\right)\tfrac{3}{n-i+1}$$
$$\left(1-\tfrac{2}{n-i}\right)\left(1-\tfrac{2}{n-i-1}\right)\cdots\left(1-\tfrac{2}{n-j+2}\right)\tfrac{2}{n-j+1}$$
$$\left(1-\tfrac{1}{n-j}\right)\left(1-\tfrac{1}{n-j-1}\right)\cdots\left(1-\tfrac{1}{n-k+2}\right)\tfrac{1}{n-k+1}=\tfrac{3}{n(n-1)(n-2)},$$

where the $(1-\cdots)$ factors are for not picking the current variable or negation to be included and the others are for including the current variable or its negation. Independence among clauses follows from the fact that we have never let the occurrence or nonoccurrence of any variable in any clause influence our decisions on other clauses.

Now, we prove the lemma by appealing to the method of deferred decisions to generate the formula. $T_i = t$ if and only if the method of deferred decisions does not put the current literal at steps $1, 2, \ldots, t-1$ into the i^{th} clause, but puts the negation of the literal at step t into it. Thus, the probability is precisely

$$\tfrac{1}{2}\left(1-\tfrac{3}{n}\right)\left(1-\tfrac{3}{n-1}\right)\cdots\left(1-\tfrac{3}{n-t+2}\right)\tfrac{3}{n-t+1}\leq\tfrac{3}{2(n-2)},$$

as claimed. Clearly the T_i are independent since again deferred decisions deal with different clauses independently. ∎

Lemma 8.21 *There exists a constant c_2 such that with probability $1-o(1)$, no clause remains a 2 or 1-clause for more than $c_2 \ln n$ steps. I.e., once a 3-clause becomes a 2-clause, it is either satisfied or reduced to a 0-clause in $O(\ln n)$ steps.*

Proof: Say that t is a "busy time" if there exists at least one 2-clause or 1-clause at time t, and define a time-window $[r+1, s]$ to be a "busy window" if time r is not busy but then each $t \in [r+1, s]$ is a busy time. We will prove that for some constant c_2, with probability $1-o(1)$, all busy windows have length at most $c_2 \ln n$.
Fix some r and s and consider the event that $[r+1, s]$ is a busy window. Since SC always decreases the total number of 1 and 2-clauses by one whenever it is positive, we must have generated at least $s-r$ new 2-clauses between r and s. Now, define an indicator variable for each 3-clause which has value one if the clause turns into a 2-clause between r and s. By Lemma 8.20 these variables are independent and the probability that a particular 3-clause turns into a 2-clause at a time t is at most $3/(2(n-2))$. Summing over t between r and s,

$$\text{Prob}\left(\text{a 3-clause turns into a 2-clause during } [r,s]\right) \leq \frac{3(s-r)}{2(n-2)}.$$

Since there are cn clauses in all, the expected sum of the indicator variables is $cn\frac{3(s-r)}{2(n-2)} \approx \frac{3c(s-r)}{2}$. Note that $3c/2 < 1$, which implies the arrival rate into the queue of 2 and 1-clauses is a constant strictly less than one. Using Chernoff bounds, if $s - r \geq c_2 \ln n$ for

appropriate constant c_2, the probability that more than $s - r$ clauses turn into 2-clauses between r and s is at most $1/n^3$. Applying the union bound over all $O(n^2)$ possible choices of r and s, we get that the probability that any clause remains a 2 or 1-clause for more than $c_2 \ln n$ steps is $o(1)$. ∎

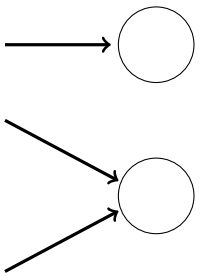
Now, assume the $1 - o(1)$ probability event of Lemma 8.21 that no clause remains a 2 or 1-clause for more than $c_2 \ln n$ steps. We will show that this implies it is unlikely the SC algorithm terminates in failure.

Suppose SC terminates in failure. This means that at some time t, the algorithm generates a 0-clause. At time $t - 1$, this clause must have been a 1-clause. Suppose the clause consists of the literal w. Since at time $t - 1$, there is at least one 1-clause, the shortest clause rule of SC selects a 1-clause and sets the literal in that clause to true. This other clause must have been $\bar{w}$. Let t_1 be the first time either of these two clauses, w or $\bar{w}$, became a 2-clause. We have $t - t_1 \leq c_2 \ln n$. Clearly, until time t, neither of these two clauses is picked by SC. So, the literals which are set to true during this period are chosen independent of these clauses. Say the two clauses were $w + x + y$ and $\bar{w} + u + v$ at the start. x, y, u, and v must all be negations of literals set to true during steps t_1 to t. So, there are only $O\left((\ln n)^4\right)$ choices for x, y, u, and v for a given value of t. There are $O(n)$ choices of w, $O(n^2)$ choices of which two clauses i and j of the input become these w and $\bar{w}$, and n choices for t. Thus, there are $O\left(n^4(\ln n)^4\right)$ choices for what these clauses contain and which clauses they are in the input. On the other hand, for any given i and j, the probability that clauses i and j both match a given set of literals is $O(1/n^6)$. Thus the probability that these choices are actually realized is therefore $O\left(n^4(\ln n)^4/n^6\right) = o(1)$, as required.

8.8 Nonuniform Models of Random Graphs

So far we have considered the $G(n, p)$ random graph model in which all vertices have the same expected degree, and moreover degrees are concentrated close to their 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 scales as $1/d^\alpha$ for some constant $\alpha > 0$.

One way to generate such graphs is to 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 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.



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 a random 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 8.12: Probability of encountering a degree d vertex when following a path in a graph.

8.8.1 Giant Component in 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 $(i-2)\,i\lambda_i$ 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, and use p_i to denote the probability that a vertex has degree i, i.e., in analog to λ_i. It turns out that 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,

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

■

8.9 Growth Models

Many graphs that arise in the outside world started as small graphs that grew over time. In a model for such graphs, vertices and edges are added to the graph over time. In such a model there are many ways in which to select the vertices for attaching a new edge. One is to select two vertices uniformly at random from the set of existing vertices. Another is to select two vertices with probability proportional to their degree. This latter method is referred to as preferential attachment. A variant of this method would be to add a new vertex at each unit of time and with probability δ add an edge where one end of the edge is the new vertex and the other end is a vertex selected with probability proportional to its degree. The graph generated by this latter method is a tree with a power law degree distribution.

8.9.1 Growth Model Without Preferential Attachment

Consider a growth model for a random graph without preferential attachment. Start with zero vertices. At each unit of time a new vertex is created and with probability δ two vertices chosen at random are joined by an edge. The two vertices may already have an edge between them. In this case, we add another edge. So, the resulting structure is a multi-graph, rather then a graph. Since at time t, there are t vertices and in expectation only $O(\delta t)$ edges where there are t^2 pairs of vertices, it is very unlikely that there will be many multiple edges.

The degree distribution for this growth model is calculated as follows. The number of vertices of degree k at time t is a random variable. Let $d_k(t)$ be the expectation of the number of vertices of degree k at time t. The number of isolated vertices increases by one at each unit of time and decreases by the number of isolated vertices, $b(t)$, that are picked to be end points of the new edge. $b(t)$ can take on values 0,1, or 2. Taking expectations,

$$d_0(t+1) = d_0(t) + 1 - E(b(t)).$$

Now $b(t)$ is the sum of two 0-1 valued random variables whose values are the number of degree zero vertices picked for each end point of the new edge. Even though the two random variables are not independent, the expectation of $b(t)$ is the sum of the expectations of the two variables and is $2\delta\frac{d_0(t)}{t}$. Thus,

$$d_0(t+1) = d_0(t) + 1 - 2\delta\frac{d_0(t)}{t}.$$

The number of degree k vertices increases whenever a new edge is added to a degree $k-1$ vertex and decreases when a new edge is added to a degree k vertex. Reasoning as above,

$$d_k\,(t+1) = d_k(t) + 2\delta\frac{d_{k-1}(t)}{t} - 2\delta\frac{d_k(t)}{t}. \tag{8.4}$$

Note that this formula, as others in this section, is not quite precise. For example, the same vertex may be picked twice, so that the new edge is a self-loop. For $k << t$, this problem contributes a minuscule error. Restricting k to be a fixed constant and letting $t \to \infty$ in this section avoids these problems.

Assume that the above equations are exactly valid. Clearly, $d_0(1) = 1$ and $d_1(1) = d_2(1) = \cdots = 0$. By induction on t, there is a unique solution to (8.4), since given $d_k(t)$ for all k, the equation determines $d_k(t+1)$ for all k. There is a solution of the form $d_k(t) = p_k t$, where p_k depends only on k and not on t, provided k is fixed and $t \to \infty$. Again, this is not precisely true since $d_1(1) = 0$ and $d_1(2) > 0$ clearly contradict the existence of a solution of the form $d_1(t) = p_1 t$.

Set $d_k(t) = p_k t$. Then,

$$(t+1)\,p_0 = p_0 t + 1 - 2\delta\frac{p_0 t}{t}$$

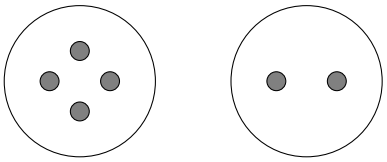


Figure 8.13: In selecting a component at random, each of the two components is equally likely to be selected. In selecting the component containing a random vertex, the larger component is twice as likely to be selected.

$$p_0 = 1 - 2\delta p_0$$
$$p_0 = \frac{1}{1+2\delta}$$

and

$$(t+1)\,p_k = p_k t + 2\delta \frac{p_{k-1}t}{t} - 2\delta \frac{p_k t}{t}$$
$$p_k = 2\delta p_{k-1} - 2\delta p_k$$

$$\begin{aligned} p_k &= \frac{2\delta}{1+2\delta} p_{k-1} \\ &= \left(\frac{2\delta}{1+2\delta}\right)^k p_0 \\ &= \frac{1}{1+2\delta}\left(\frac{2\delta}{1+2\delta}\right)^k. \end{aligned} \tag{8.5}$$

Thus, the model gives rise to a graph with a degree distribution that falls off exponentially fast with the degree.

The generating function for component size

Let $n_k(t)$ be the expected number of components of size k at time t. Then $n_k(t)$ is proportional to the probability that a randomly picked component is of size k. This is not the same as picking the component containing a randomly selected vertex (see Figure 8.13). Indeed, the probability that the size of the component containing a randomly selected vertex is k is proportional to $kn_k(t)$. We will show that there is a solution for $n_k(t)$ of the form $a_k t$ where a_k is a constant independent of t. After showing this, we focus on the generating function $g(x)$ for the numbers $ka_k(t)$ and use $g(x)$ to find the threshold for giant components.

Consider $n_1(t)$, the expected number of isolated vertices at time t. At each unit of time, an isolated vertex is added to the graph and an expected $\frac{2\delta n_1(t)}{t}$ many isolated

vertices are chosen for attachment and thereby leave the set of isolated vertices. Thus,

$$n_1(t+1) = n_1(t) + 1 - 2\delta\frac{n_1(t)}{t}.$$

For $k > 1$, $n_k(t)$ increases when two smaller components whose sizes sum to k are joined by an edge and decreases when a vertex in a component of size k is chosen for attachment. The probability that a vertex selected at random will be in a size k component is $\frac{kn_k(t)}{t}$. Thus,

$$n_k(t+1) = n_k(t) + \delta\sum_{j=1}^{k-1}\frac{jn_j(t)}{t}\frac{(k-j)n_{k-j}(t)}{t} - 2\delta\frac{kn_k(t)}{t}.$$

To be precise, one needs to consider the actual number of components of various sizes, rather than the expected numbers. Also, if both vertices at the end of the edge are in the same k-vertex component, then $n_k(t)$ does not go down as claimed. These small inaccuracies can be ignored.

Consider solutions of the form $n_k(t) = a_k t$. Note that $n_k(t) = a_k t$ implies the number of vertices in a connected component of size k is $ka_k t$. Since the total number of vertices at time t is t, ka_k is the probability that a random vertex is in a connected component of size k. The recurrences here are valid only for k fixed as $t \to \infty$. So $\sum_{k=0}^{\infty} ka_k$ may be less than 1, in which case, there are nonfinite size components whose sizes are growing with t. Solving for a_k yields $a_1 = \frac{1}{1+2\delta}$ and $a_k = \frac{\delta}{1+2k\delta}\sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}$.

Consider the generating function $g(x)$ for the distribution of component sizes where the coefficient of x^k is the probability that a vertex chosen at random is in a component of size k.

$$g(x) = \sum_{k=1}^{\infty} ka_k x^k.$$

Now, $g(1) = \sum_{k=0}^{\infty} ka_k$ is the probability that a randomly chosen vertex is in a finite sized component. For $\delta = 0$, this is clearly one, since all vertices are in components of size one. On the other hand, for $\delta = 1$, the vertex created at time one has expected degree $\log n$ (since its expected degree increases by $2/t$ and $\sum_{t=1}^{n}(2/t) = \Theta(\log n)$); so, it is in a nonfinite size component. This implies that for $\delta = 1$, $g(1) < 1$ and there is a nonfinite size component. Assuming continuity, there is a $\delta_{critical}$ above which $g(1) < 1$. From the formula for the $a_i's$, we will derive the differential equation

$$g = -2\delta x g' + 2\delta x g g' + x$$

and then use the equation for g to determine the value of δ at which the phase transition for the appearance of a nonfinite sized component occurs.

Derivation of $g(x)$

From

$$a_1 = \frac{1}{1+2\delta}$$

and

$$a_k = \frac{\delta}{1+2k\delta}\sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}$$

derive the equations

$$a_1\,(1+2\delta) - 1 = 0$$

and

$$a_k\,(1+2k\delta) = \delta\sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}$$

for $k \geq 2$. The generating function is formed by multiplying the k^{th} equation by kx^k and summing over all k. This gives

$$-x + \sum_{k=1}^{\infty} k a_k x^k + 2\delta x \sum_{k=1}^{\infty} a_k k^2 x^{k-1} = \delta \sum_{k=1}^{\infty} kx^k \sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}.$$

Note that

$$g(x) = \sum_{k=1}^{\infty} k a_k x^k \text{ and } g'(x) = \sum_{k=1}^{\infty} a_k k^2 x^{k-1}.$$

Thus,

$$-x + g(x) + 2\delta x g'(x) = \delta \sum_{k=1}^{\infty} kx^k \sum_{j=1}^{k-1} j(k-j)a_j a_{k-j}.$$

Working with the right hand side

$$\delta \sum_{k=1}^{\infty} kx^k \sum_{j=1}^{k-1} j(k-j)a_j a_{k-j} = \delta x \sum_{k=1}^{\infty}\sum_{j=1}^{k-1} j(k-j)(j+k-j)x^{k-1} a_j a_{k-j}.$$

Now breaking the $j+k-j$ into two sums gives

$$\delta x \sum_{k=1}^{\infty}\sum_{j=1}^{k-1} j^2 a_j x^{j-1}(k-j)a_{k-j}x^{k-j} + \delta x \sum_{k=1}^{\infty}\sum_{j=1}^{k-1} j a_j x^j (k-j)^2 a_{k-j}x^{k-j-1}.$$

Notice that the second sum is obtained from the first by substituting $k-j$ for j and that both terms are $\delta x g' g$. Thus,

$$-x + g(x) + 2\delta x g'(x) = 2\delta x g'(x) g(x).$$

Hence,

$$g' = \frac{1}{2\delta}\frac{1-\frac{g}{x}}{1-g}.$$

Phase transition for nonfinite components

The generating function $g(x)$ contains information about the finite components of the graph. A finite component is a component of size $1, 2, \ldots$, which does not depend on t. Observe that $g(1) = \sum_{k=0}^{\infty} k a_k$ and hence $g(1)$ is the probability that a randomly chosen vertex will belong to a component of finite size. If $g(1) = 1$ there are no infinite components. When $g(1) \neq 1$, then $1 - g(1)$ is the expected fraction of the vertices that are in nonfinite components. Potentially, there could be many such nonfinite components. But an argument similar to Part 3 of Theorem **??** concludes that two fairly large components would merge into one. Suppose there are two connected components at time t, each of size at least $t^{4/5}$. Consider the earliest created $\frac{1}{2}t^{4/5}$ vertices in each part. These vertices must have lived for at least $\frac{1}{2}t^{4/5}$ time after creation. At each time, the probability of an edge forming between two such vertices, one in each component, is at least $\delta\Omega(t^{-2/5})$ and so the probability that no such edge formed is at most $\left(1 - \delta t^{-2/5}\right)^{t^{4/5}/2} \leq e^{-\Omega(\delta t^{2/5})} \to 0$. So with high probability, such components would have merged into one. But this still leaves open the possibility of many components of size t^{ε}, $(\ln t)^2$, or some other slowly growing function of t.

We now calculate the value of δ at which the phase transition for a nonfinite component occurs. Recall that the generating function for $g(x)$ satisfies

$$g'(x) = \frac{1}{2\delta} \frac{1 - \frac{g(x)}{x}}{1 - g(x)}.$$

If δ is greater than some $\delta_{critical}$, then $g(1) \neq 1$. In this case the above formula at $x = 1$ simplifies with $1 - g(1)$ canceling from the numerator and denominator, leaving just $\frac{1}{2\delta}$. Since ka_k is the probability that a randomly chosen vertex is in a component of size k, the average size of the finite components is $g'(1) = \sum_{k=1}^{\infty} k^2 a_k$. Now, $g'(1)$ is given by

$$g'(1) = \frac{1}{2\delta} \tag{8.6}$$

for all δ greater than $\delta_{critical}$. If δ is less than $\delta_{critical}$, then all vertices are in finite components. In this case $g(1) = 1$ and both the numerator and the denominator approach zero. Appling L'Hopital's rule

$$\lim_{x \to 1} g'(x) = \frac{1}{2\delta} \left. \frac{\frac{xg'(x) - g(x)}{x^2}}{g'(x)} \right|_{x=1}$$

or

$$(g'(1))^2 = \frac{1}{2\delta} \left(g'(1) - g(1)\right).$$

The quadratic $(g'(1))^2 - \frac{1}{2\delta}g'(1) + \frac{1}{2\delta}g(1) = 0$ has solutions

$$g'(1) = \frac{\frac{1}{2\delta} \pm \sqrt{\frac{1}{4\delta^2} - \frac{4}{2\delta}}}{2} = \frac{1 \pm \sqrt{1-8\delta}}{4\delta}. \tag{8.7}$$

The two solutions given by (8.7) become complex for $\delta > 1/8$ and thus can be valid only for $0 \leq \delta \leq 1/8$. For $\delta > 1/8$, the only solution is $g'(1) = \frac{1}{2\delta}$ and an infinite component exists. As δ is decreased, at $\delta = 1/8$ there is a singular point where for $\delta < 1/8$ there are three possible solutions, one from (8.6) which implies a giant component and two from (8.7) which imply no giant component. To determine which one of the three solutions is valid, consider the limit as $\delta \to 0$. In the limit all components are of size one since there are no edges. Only (8.7) with the minus sign gives the correct solution

$$g'(1) = \frac{1-\sqrt{1-8\delta}}{4\delta} = \frac{1-\left(1-\frac{1}{2}8\delta-\frac{1}{4}64\delta^2+\cdots\right)}{4\delta} = 1+4\delta+\cdots = 1.$$

In the absence of any nonanalytic behavior in the equation for $g'(x)$ in the region $0 \leq \delta < 1/8$, we conclude that (8.7) with the minus sign is the correct solution for $0 \leq \delta < 1/8$ and hence the critical value of δ for the phase transition is $1/8$. As we shall see, this is different from the static case.

As the value of δ is increased, the average size of the finite components increase from one to

$$\left.\frac{1-\sqrt{1-8\delta}}{4\delta}\right|_{\delta=1/8} = 2$$

when δ reaches the critical value of $1/8$. At $\delta = 1/8$, the average size of the finite components jumps to $\left.\frac{1}{2\delta}\right|_{\delta=1/8} = 4$ and then decreases as $\frac{1}{2\delta}$ as the giant component swallows up the finite components starting with the larger components.

Comparison to static random graph

Consider a static random graph with the same degree distribution as the graph in the growth model. Again let p_k be the probability of a vertex being of degree k. From (8.5)

$$p_k = \frac{(2\delta)^k}{(1+2\delta)^{k+1}}\ .$$

Recall the Molloy Reed analysis of random graphs with given degree distributions which asserts that there is a phase transition at $\sum_{i=0}^{\infty} i(i-2)p_i = 0$. Using this, it is easy to see that a phase transition occurs for $\delta = 1/4$. For $\delta = 1/4$,

$$p_k = \frac{(2\delta)^k}{(1+2\delta)^{k+1}} = \frac{\left(\frac{1}{2}\right)^k}{\left(1+\frac{1}{2}\right)^{k+1}} = \frac{\left(\frac{1}{2}\right)^k}{\frac{3}{2}\left(\frac{3}{2}\right)^k} = \frac{2}{3}\left(\frac{1}{3}\right)^k$$

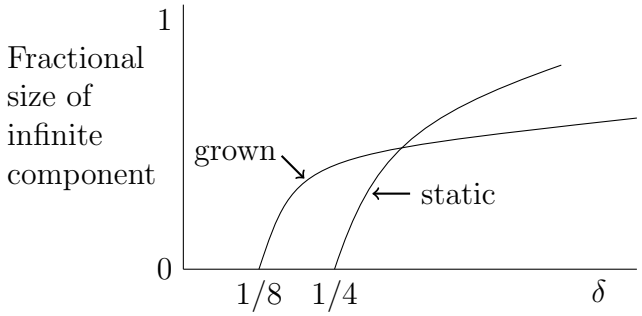


Figure 8.14: Comparison of the static random graph model and the growth model. The curve for the growth model is obtained by integrating g'.

and

$$\sum_{i=0}^{\infty} i(i-2)\tfrac{2}{3}\left(\tfrac{1}{3}\right)^i = \tfrac{2}{3}\sum_{i=0}^{\infty} i^2\left(\tfrac{1}{3}\right)^i - \tfrac{4}{3}\sum_{i=0}^{\infty} i\left(\tfrac{1}{3}\right)^i = \tfrac{2}{3}\times\tfrac{3}{2} - \tfrac{4}{3}\times\tfrac{3}{4} = 0.$$

Recall that $1+a+a^2+\cdots = \frac{1}{1-a}$, $a+2a^2+3a^3\cdots = \frac{a}{(1-a)^2}$, and $a+4a^2+9a^3\cdots = \frac{a(1+a)}{(1-a)^3}$.

See references at end of the chapter for calculating the fractional size S_{static} of the giant component in the static graph. The result is

$$S_{static} = \begin{cases} 0 & \delta \le \frac{1}{4} \\ 1 - \frac{1}{\delta+\sqrt{\delta^2+2\delta}} & \delta > \frac{1}{4} \end{cases}$$

8.9.2 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 expected 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

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

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.

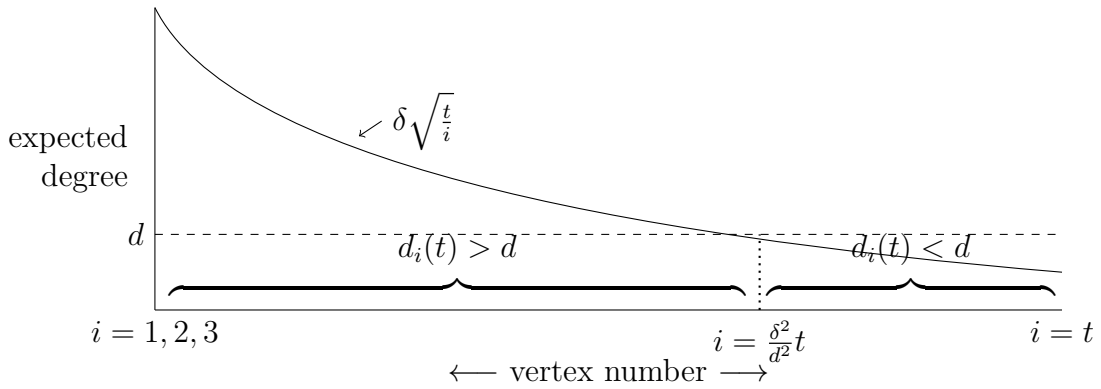


Figure 8.15: Illustration of degree of i^{th} vertex at time t. At time t, vertices numbered 1 to $\frac{\delta^2}{d^2}t$ have degrees greater than d.

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.

8.10 Small World Graphs

In the 1960's, Stanley Milgram carried out an experiment that indicated that most pairs of 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 via bidirectional local edges. 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 are too-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.

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

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

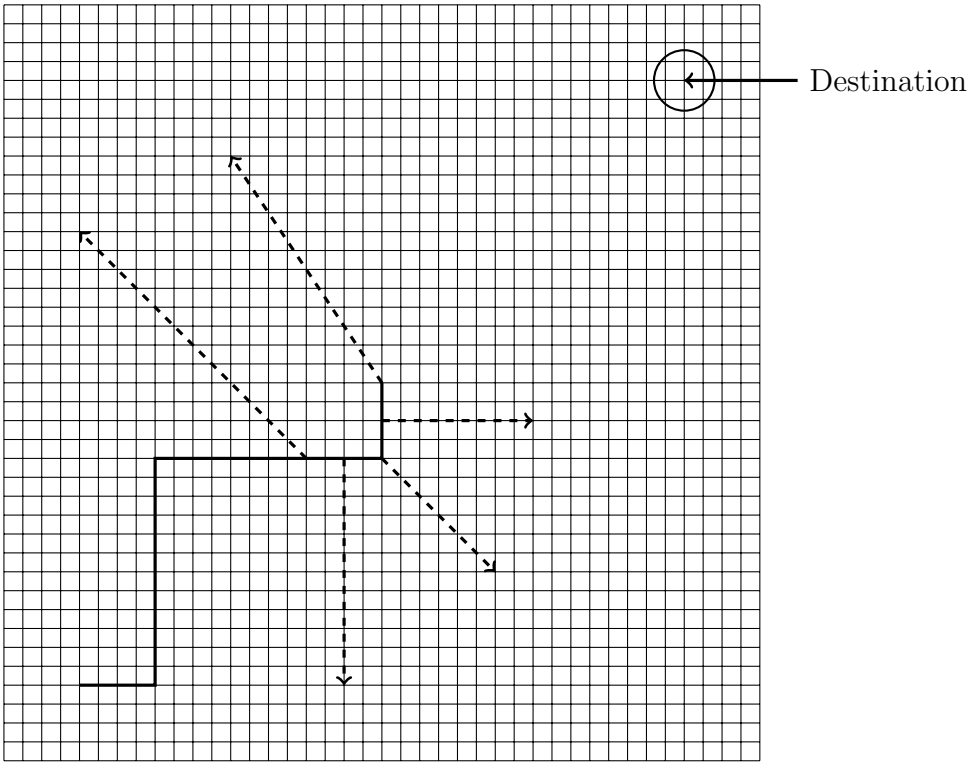


Figure 8.16: For $r < 2$, on a short path you are unlikely to encounter a long-distance edge that takes you close to the destination.

$r > 2$ 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 8.17: Effects of different values of r on the expected length of long-distance edges and the ability to find short paths.

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 half of 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. The probability that a particular edge of length greater than or equal to $n^{\frac{r+2}{2r}}$ is chosen is upper bounded by $cn^{-\left(\frac{r+2}{2}\right)}$ for some constant c. Since there are n^2 long edges, the expected number of edges of length at least $n^{\frac{r+2}{2r}}$ is at most $cn^2 n^{-\frac{(r+2)}{2}}$ 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 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 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 8.18. Recall that the normalizing constant c_r is upper bounded by $O(\ln n)$, and

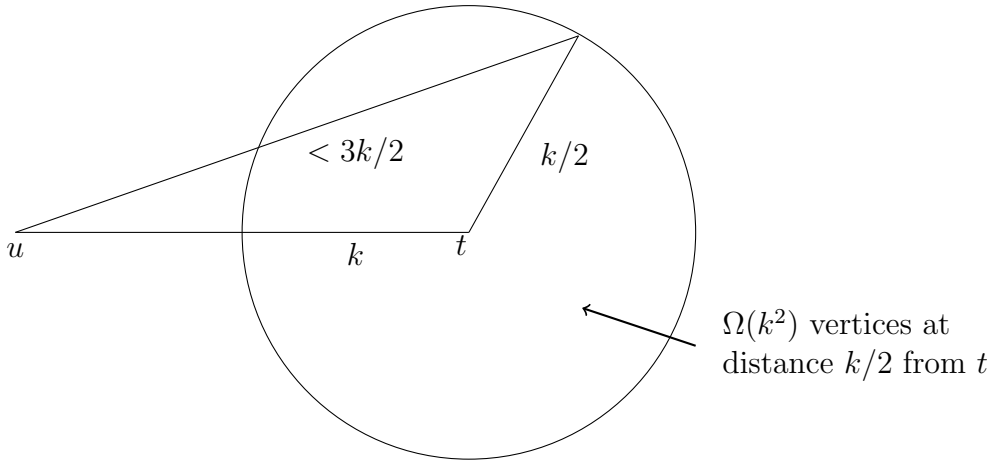


Figure 8.18: Small worlds.

hence, the constant of proportionality is lower bounded by some constant times $1/\ln n$. The probability that the long-distance edge from u goes to one of these vertices is at least

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

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

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

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 $r < 2$.

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 8.13 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 8.13. 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.

8.11 Bibliographic Notes

The $G(n,p)$ random graph model is from Erdös Rényi [ER60]. Among the books written on properties of random graphs a reader may wish to consult Frieze and Karonski [FK15], Jansen, Luczak and Ruciński [JLR00],or Bollobás [Bol01]. Material on phase transitions can be found in [BT87]. The argument for existence of a giant component is