Chapter 3: Random Graphs

Large graphs appear in many contexts such as the World Wide Web, the internet, social networks, journal citations and other places. What is different about the modern study of large graphs from traditional graph theory and graph algorithms is that here one seeks statistical properties of these very large graphs rather than an exact answer to questions. This is perhaps akin to the switch Physics made in the late 19th century in going from Mechanics to Statistical Mechanics. Just as the Physicists did, one formulates clean abstract models of graphs that may not be completely realistic in every situation, but admit a very nice mathematical development that can guide what really might happen in practical situations. Perhaps the most important such model is the $G(n, p)$ model of a random graph. In this chapter, we study several random graph models and the properties of the random graphs generated by these models.

3.1 The G(n,p) model

One of the earliest models of a random graph is the $G(n, p)$ model due to Erdös and Rényi. The model 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 (v,w) of distinct vertices, the edge (v,w) is present with probability p. The presence of each edge is statistically independent of all other edges. The graph-valued random variable with these parameters is denoted by $G(n, p)$. Here $G(n, p)$ is a random variable. When we refer to ``the graph $G(n, p)$'', we mean one realization of the random variable. The interesting thing about this model and similar ones is that even though pairs of vertices independently chose edges with no ``collusion'' certain global properties of the graph emerge from such behavior. Also note that in many cases, p is a function of n such as $p(n) = \frac{d}{n}$ for some constant d. In this case, the expected degree of a vertex of the graph is d.

3.1.1 Degree Distribution

One of the simplest quantities to observe in a real graph is the distribution of the degrees of vertices. It is also very simple to study these distributions in $G(n, p)$ since the degree of each vertex is the sum of *n*-1 independent random variables. Since we will be dealing with graphs where *n*, the number of vertices, is large, from here on we replace *n*-1 by *n* to simplify formulas. Consider the $G(n, p)$ random graph model for *p* greater than $c\frac{\log n}{n}$. The degree distribution is binomial and is concentrated about the average degree, falling off exponentially fast as one moves away from the average. However, graphs that appear in many applications do not have binomial degree distributions. Rather their degree distribution is 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. Clearly 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 frequency of a vertex having a given degree decreases as a power of the degree as in $\mathrm{P}(d) = c\frac{1}{d^r}$ for some small constant *r*. Later, we shall consider a random graph model giving rise to such degree distributions.

Consider the graph $G(n, p)$. Since *p* is the probability of an edge being present, the expected degree of a vertex is $d = pn$. The actual degree distribution is binomial with

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

The quantity $\binom{n}{k}$ is the number of ways *k* edges can be chosen and $p^k(1-p)^{n-k}$ is the probability that the *k* selected edges are present and the remaining *n-k* are not.

Example 3.1: 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 a multiplicative constant of $1 \pm \varepsilon$ times n/2. To see this, note that the probability that a vertex is of degree *k* is

$$\mathrm{Prob}(k) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k} = \binom{n}{k}\frac{1}{2^n}.$$

This probability distribution has a mean of *m*=*n*/2 and variance $\sigma^2 = \frac{n}{4}$. Near the mean, the binomial distribution is well approximated by the normal density $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(k-m)^2}{\sigma^2}}$ or

$\frac{1}{\sqrt{\pi n/2}} e^{-\frac{(k-n/2)^2}{n/2}}$. See appendix. The standard deviation is $\frac{\sqrt{n}}{2}$ and essentially all of the probability mass is within an additive term $\pm c\sqrt{n}$ of $n/2$ for some constant c and thus is certainly within a multiplicative factor of $1 \pm \varepsilon$ of $n/2$ for sufficiently large n.

■

The following theorem claims that for $p \geq \frac{\log n}{n}$, the degree distribution of the random graph $G(n, p)$ is tightly concentrated about its expected value.

Theorem 3.1: Let v be a vertex of the random graph $G(n, p)$. When $p \geq \frac{\log n}{n}$, for any c, $0 \leq c \leq \sqrt{np}$,

$$\text{Prob}\left(\left|np - \deg(v)\right| \geq c\sqrt{np}\right) \leq 4e^{-c^2/8})$$

Proof: The theorem follows from a Chernoff bound since the degree of v is just the sum of $n-1$ independent $0-1$ valued random variables.

■

ADD MATERIAL ON LOW DEGREE DISTRIBUTIONS?

3.1.2 Existence of triangles in G(n, d/n)

What is the expected number of triangles in $G\left(n, \frac{d}{n}\right)$? As the number of vertices, n, 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 and 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 cancelling the rate of grow 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 $\frac{d^3}{6}$. There are $\binom{n}{3}$ triples of vertices. Each triple has probability $\left(\frac{d}{n}\right)^3$ of being a triangle. Even though the events are not statistically independent, by linearity of expectation (see Appendix), 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

$$E(\#) = \binom{n}{3}\left(\frac{d}{n}\right)^3 = \frac{d^3}{6}$$

3.1.3 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 temperature or pressure increases. Some examples of this are the abrupt appearance of cycles when p reaches $\frac{1}{n}$ and the disappearance of isolated vertices when p reaches $\frac{\log n}{n}$.

For many properties of random graphs, such a threshold exists where an abrupt transition from not having 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*. We shall soon see that every monotone property has a threshold. This is true not only for monotone properties of $G(n, p)$, but for monotone 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, we say $p(n)$ is a *sharp threshold*.

In establishing phase transitions for various properties, we will often use a variable x to denote the number of occurrences of some item in a graph. By showing that the expected value of x is zero, we will conclude that a graph picked at random has no occurrence of the item. However, when the expect value of x is large, we cannot conclude that a graph picked at random will likely have a copy since the items may all appear on a small fraction of the graphs. We willneed to resort to a technique called the second moment method For a non-negative random variable x, if one can show that in the limit $\mathrm{Var}(x) = o((E(x))^2)$, then x is almost surely non-zero.

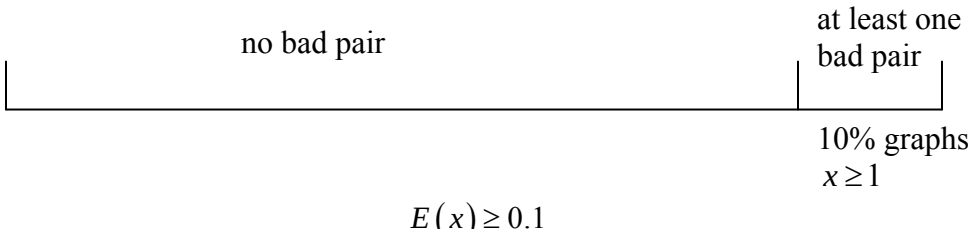

Figure 3.1: If the expected number of graphs with diameter greater than two did not go to zero, then $E(x)$, the expected number of bad pairs per graph, could not be zero. Suppose 10% of the graphs had at least one bad pair. Then the expected number of bad pairs per graph must be at least 0.1. Thus, $E(x) = 0$ implies the probability that a graph has a bad pair is zero. However, the other direction needs more work. If $E(x)$ were not zero, a second moment argument is

needed to conclude that the probability that a graph had a bad pair was non zero since there could be a large number of bad pairs concentrated on a vanishingly small fraction of all graphs. The second moment argument claims that for a non negative random variable x with $E(x) > 0$, if $\text{var}(x) \le E^2(x)$ or alternatively if $E(x^2) \le E^2(x)$, then almost surely $x > 0$. We will return to this later.

Threshold for diameter two

We now present the first example of a 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 we consider is that of a random graph having diameter (maximum length of a shortest path between a pair of nodes) less than or equal to two. We will see that when $p = \frac{c\sqrt{\ln n}}{\sqrt{n}}$, 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 3.2: Let $p = \frac{c\sqrt{\ln n}}{\sqrt{n}}$. For $c < \sqrt{2}$, $G(n, p)$ almost surely has diameter greater than two and for $c > \sqrt{2}$, $G(n, p)$ almost surely has diameter less than or equal to two.

Proof: If *G* has diameter greater than two, then there exists a pair of vertices *u* and *v* that are not adjacent to each other and no other vertex of *G* is adjacent to both *u* and *v*. This motivates calling such a pair *bad*.

Introduce a set of indicator random variables x_{ij}, where x_{ij} is 1 if and only if the pair (i, j) is bad. Let

$$x = \sum_{i,j} x_{ij}$$

be the number of bad pairs of vertices. A graph has diameter 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 diameter at most two. See Figure 3.1.

The probability that a given vertex is adjacent to both vertices in a pair of vertices (u,v) is p^2 and 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 (u,v) is $(1 - p^2)^{n-2}$ and the probability that *u* and *v* are not connected is $1 - p$. Since there are $\binom{n}{2}$ pairs of vertices, the expected number of bad pairs is

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

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

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

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

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

$$\text{var}(x) = E\left[\sum_{ij}\left(x_{ij}\right) - E\left(\sum_{ij} x_{ij}\right)\right]^2 = E\left[\left(\sum_{ij}\left(x_{ij}\right) - E\left(\sum_{ij} x_{ij}\right)\right)\left(\sum_{kl}\left(x_{kl}\right) - E\left(\sum_{kl} x_{kl}\right)\right)\right].$$
$$= \sum_{i,j,k,l} E\left[\left(x_{ij} - E\left(x_{ij}\right)\right)\left(x_{kl} - E\left(x_{kl}\right)\right)\right]$$

The summation has n^4 terms. However, if *i* and *j* are distinct from *k* and *l*, then by independence

$$E\left[\left(x_{ij} - E\left(x_{ij}\right)\right)\left(x_{kl} - E\left(x_{kl}\right)\right)\right] = 0.$$

Thus, we need only consider the summation over n^3 terms and

$$\text{var}(x) = \sum_{\substack{ijk \\ i\neq j \\ i\neq k}}\left[\left(x_{ij} - E\left(x_{ij}\right)\right)\left(x_{ik} - E\left(x_{ik}\right)\right)\right]$$

Now,

$$E\left[\left(x_{ij} - E\left(x_{ij}\right)\right)\left(x_{ik} - E\left(x_{ik}\right)\right)\right] = E\left(x_{ij}x_{ik} - x_{ij}E\left(x_{ik}\right) - x_{ik}E\left(x_{ij}\right) + E\left(x_{ij}\right)E\left(x_{ik}\right)\right)$$
$$= E\left(x_{ij}x_{ik}\right) - E\left(x_{ij}\right)E\left(x_{ik}\right)$$
$$\le E\left(x_{ij}x_{ik}\right)$$

Thus,

$$\text{var}(x) \le \sum_{\substack{ijk \\ i\neq j \\ i\neq k}} E\left(x_{ij}x_{ik}\right)$$
$$\le \sum_{\substack{ijk \\ i\neq j \\ i\neq k \\ j\neq k}} E\left(x_{ij}x_{ik}\right) + \sum_{\substack{ij \\ i\neq j}} E\left(x_{ij}^2\right) \qquad \text{Eq. 3.1}$$

Since $x_{ij}x_{ik}=1$ if and only if both pairs $(i,j),(i,k)$ are bad, which happens if and only if for every other vertex $u \neq i,j,k$, either there is no edge between i and u or there is an edge (i,u) and both edges (j,u), (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 u is $\left(1-2p^2\right)^n$. Substituting $p=\frac{c\sqrt{\ln n}}{\sqrt{n}}$ yields

$$\left(1-\frac{2c^2\ln n}{n}\right)^n = e^{-2c^2\ln n} = n^{-2c^2}$$

Summing over all distinct triples yields n^{3-2c^2} for the first summation in Eq. 3.1.

For the second summation in Eq. 3.1, since the value of x_{ij} is zero or one, $E\left(x_{ij}^2\right)=E\left(x_{ij}\right)$. Thus

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

Since $\text{var}(x) \le n^{3-2c^2}+n^{2-c^2}$, $E(x) \le n^{2-c^2}$, and $E^2(x) \le n^{4-2c^2}$, it follows that $\text{var}(x) \le E^2(x)$. Thus by the second moment argument a graph almost surely has at least one bad pair of vertices and thus has diameter greater than two.

■

Disappearance of isolated vertices

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

Theorem 3.3: 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

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

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, then it follows that almost all graphs have no isolated vertices,. On the other hand, if the expected number of isolated vertices goes to infinity, then we need a second moment argument to show that almost all graphs have an isolated vertex and that the isolated vertices are not bunched on some vanishingly small set of graphs with almost all graphs having none.

Write $x=x_1+x_2+\cdots+x_n$ where x_i is the indicator variable indicating whether vertex *i* is an isolated vertex. Then $E\left(x^2\right)=\sum_{i=1}^{n}E\left(x_i^2\right)+\sum_{i\neq j}E\left(x_ix_j\right)$ and since x_i equals 0 or 1, $x_i^2=x_i$. Since all elements in the second sum are equal

$$E\left(x^2\right) = E\left(x\right) + n\left(n-1\right)E\left(x_1 x_2\right)$$
$$= E\left(x\right) + n\left(n-1\right)\left(1-p\right)^{2(n-1)-1}$$

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

Now

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

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

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

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

■

Threshold for graph connectivity

As p increases from $p = 0$, small components form. At $p = \frac{1}{n}$ a giant component emerges and sallows up smaller components starting with the larger components and ending up sallowing up isolated vertices forming a single connected component at $p = \frac{\ln n}{n}$ at which point the graph becomes connected.

To show that the graph becomes connected at $p = \frac{\ln n}{n}$, we prove that if $p = \frac{c\ln n}{n}$, for c>1, almost surely, there is no connected component with k vertices for any k, $2 \le k \le \frac{n}{2}$. This clearly suffices since if the graph is disconnected, it has a component of size between 1 and n/2 and we have already ruled out components of size 1 (isolated vertices). For $c < 1$, there are isolated vertices and the graph is not connected.

The probability that k vertices form a 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 all in. The at most is because $G\left(n,p\right)$ may contain more than one spanning tree on these nodes.

EXPLAIN BETTER There are k^{k-2} spanning trees on k nodes. The probability of all the $k-1$ edges of one spanning tree being in 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 $k^{k-2}p^{k-1}(1-p)^{kn-k^2}$. Let x_k be the number of components of size *k*. Then

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

Substitute $p = \frac{c\ln}{n}$ into the above inequality 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

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

It is good to keep in mind that the leading terms here are the last two and in fact at $k = n,$ they cancel each other so that our argument does not prove the fallacious statement that there is no connected component of size n, since in fact there is. Now let

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

We see by differentiating that $f'(k) = 1 + \ln\ln n - \frac{2}{k} + \ln c - c\ln n + \frac{2ck\ln n}{n}$ and that $f''(k) = \frac{2}{k^2} + \frac{2c\ln n}{n} > 0$, so the function $f(k)$ attains its maximum over the range at one of the extreme points 2 or $\frac{n}{2}$. By direct calculation, it is easy to see that $f(2) > f(n/2)$ and that $f(2) \approx (1-2c)\ln n$. Thus, $E(x_k) \le n^{1-2c}$ and $E\left(\sum_{k=2}^{n/2} x_k\right) \le n^{2-2c}$. Since c>1, it follows that the expected number of components of size *k*, $2 \le k \le \frac{n}{2}$ goes to zero and thus there are no components other than isolated vertices and the giant component which is of size great than $\frac{n}{2}$. By Theorem 3.3, $\frac{c\ln n}{n}$ with c=1 is a sharp threshold for the disappearance of isolated vertices, and since there are no components except for isolated vertices and the giant component, $\frac{\ln n}{n}$ is a sharp threshold for connectivity.

Emergence of cycles

Another threshold property is the emergence of cycles in $G(n,p)$ when *p* becomes asymptotically equivalent to $\frac{1}{n}$.

Theorem 3.4: The threshold for the existence of cycles in $G(n,p)$ is $p = \frac{1}{n}$.

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 in one of $k-1$ ways, a third in one of *k-2* ways, etc. Since a cycle and its reverse are the same cycle, divide by 2. Thus, there are $\binom{n}{k}\frac{(k-1)!}{2}$ cycles of length *k* and

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

When p is asymptotically less than $\frac{1}{n}$, $\lim_{n\to\infty} np=0$ and $\lim_{n\to\infty}\sum_{k=3}^{n}(np)^k=0$. Thus $E(x)=0$ and a graph almost surely has no cycles.

Having considered the case when *p* is asymptotically less than $\frac{1}{n}$, we now consider the case when p asymptotically equals $\frac{1}{n}$. Let $p=\frac{c}{n}$ for some constant c. When $p=\frac{c}{n}$ for some constant c,

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

$$=\tfrac{1}{2}\sum_{k=3}^{n}\tfrac{n(n-1)\cdots(n-k+1)}{n^k}\tfrac{(c)^k}{k}.$$

$E(x)$ converges if c<1 and diverges if $c\ge 1$. If $c<1$, $E(x)\cong\frac{1}{2}\sum_{k=3}^{n}\frac{c^k}{k}$ and $\lim_{n\to\infty}E(x)$ equals a constant. If $c=1$, $E(x)=\frac{1}{2}\sum_{k=3}^{n}\frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{1}{k}$. Consider only the first $\log n$ terms of the sum. Then in the limit as n goes to infinity

$$\lim_{n\to\infty}E(x)\ge\lim_{n\to\infty}\tfrac{1}{2}\sum_{k=3}^{\log n}\tfrac{1}{k}\ge\lim_{n\to\infty}(\log\log n)=\infty.$$

Thus, for $p=\frac{c}{n}$, c<1, $E(x)$ converges to a nonzero constant and graphs will have with some non zero probability a constant number of cycles independent of the size of the graph. For c>1 they will have an unbounded number of cycles increasing with n. A second moment argument can be used to show that for $p=\frac{c}{n}$, a graph will have a cycle with probability tending to one.

■

Property	Threshold
Cycles	$1/n$
giant component	$1/n$
giant component + isolated vertices	$\frac{1}{4}\frac{\ln n}{n}$
connectivity, disappearance of isolated vertices	$\frac{\ln n}{n}$
diameter two	$\sqrt{\frac{2\ln n}{n}}$

3.1.4 Phase transitions for monotonic properties

For many graph properties such as connectivity, 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 G if when a graph G has the property any graph obtained by adding edges to G must also have the property.

Lemma 3.1: If Q is an increasing property, and $0 \le p \le q \le 1$, then the probability that $G(n, p)$ has property Q is at most the probability that $G(n, p)$ has property Q.

Proof: For this and the next proof, we use an interesting relationship between $G(n, p)$ and $G(n, q)$. Generate $G(n, q)$ as follows. First generate $G(n, p)$. Recall that this really means, generate a graph on n vertices with independent edge probabilities p. Now generate independently another graph $G\left(n, \frac{q-p}{1-p}\right)$ and take the union by putting in an edge if only if at least one of the two graphs has that edge. Call this graph H. We claim that 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 in/out independently. Now, the lemma follows since whenever $G(n, p)$ has the property Q, H also has the property Q.

■

Threshold

Let Q be an increasing property for $G(n, p)$. As p increases, the graph goes from not having property Q to having property Q. Recall the definition of a phase transition we made in Section 3.1.4. If the transition from not having the property to having the property occurs within a multiplicative constant, the property has a threshold. Thresholds arise not only for a graph $G(n, p)$ but also in other random structures. We now show that every increasing property of $G(n, p)$ has a threshold.

To show that $p(n)$ is a threshold for a property Q, we need to show that the probability of property Q goes from 0 to 1 within a range that is bounded by a multiplicative constant of p(n). Define $p(n, \varepsilon)$ to be the function $p(n)$ such that the probability that $G(n, p)$ has property Q equals ε. Such a $p(n)$ may not be unique, in this case we take the *lim inf* of this. This detail is only technical. To simplify notation we write $p(\varepsilon)$ for $p(n, \varepsilon)$ remembering that $p(\varepsilon)$ is a function of n and not just a constant.

We now show that for any ε, $0 < \varepsilon < \frac{1}{2}$, the range $\left[p(\varepsilon), p(1-\varepsilon) \right]$ in which the probability of property Q goes from ε to $1-\varepsilon$ is bounded by a multiplicative constant *m* in the sense that $p(1-\varepsilon) \le mp(\varepsilon)$. This then implies that the function $p(\frac{1}{2})$ is a threshold for Q..

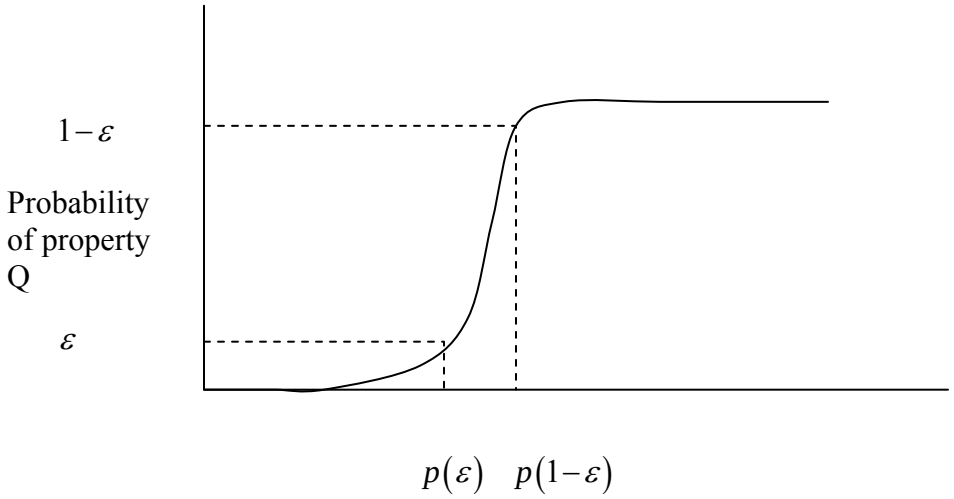

Figure 3.2: The figure illustrates the range $\left[p(\varepsilon), p(1-\varepsilon) \right]$ in which the probability of property Q goes from ε to $1-\varepsilon$. The horizontal axis is labeled with functions. We assume that we have some set of functions that can be linearly ordered such as $\log\log n$, $\log n$, $\log^2 n$, $\frac{1}{n^2}$, $\frac{1}{n}$, $\frac{1}{\sqrt{n}}$, 1.

Theorem 3.5: Let Q be an increasing property of $G(n, p)$. Then Q has a threshold.

Proof: Let $0 < \varepsilon < \frac{1}{2}$ and let *m* be an integer such that $(1-\varepsilon)^m \le \varepsilon$. Note that *m* depends only upon ϵ, not on n. Recall that $p(\varepsilon)$ is the value of p such that the probability that $G(n, p)$ has property Q equals ε. We now show that $p(1-\varepsilon) \le mp(\varepsilon)$, that is, $G(n, p)$ goes from having property Q with probability ε to having property Q with probability $1-\varepsilon$ within a multiplicative factor *m* and hence there is a threshold.

Consider the union H of *m* independent copies of $G(n, p(\varepsilon))$. By definition, (i) edges are independently in/out in H and (ii) the probability that an edge is in in H is

$$q = 1-(1-p(\varepsilon))^m \le mp(\varepsilon).$$

$\left(1-p\left(\varepsilon\right)\right)^m$ is the probability that an edge is not selected in any of the m copies and thus $q=1-\left(1-p\left(\varepsilon\right)\right)^m$ is the probability that it is selected in at least one of the copies.

If one or more of the m independent copies of $G\left(n,p\left(\varepsilon\right)\right)$ has the property Q, then H has the property Q. Thus, if H does not have property Q, then none of the $G\left(n,p\left(\varepsilon\right)\right)$ has the property Q. That is,

$$\text{Prob}\left[G\left(n,q\right) \text{ does not have } Q\right]=\left(\text{Prob}\left[G\left(n,p\left(\varepsilon\right)\right) \text{ does not have } Q\right]\right)^m=(1-\epsilon)^m\le\epsilon.$$

or

$$\text{Prob}\left[G\left(n,mp(\epsilon)\right) \text{ has Q }\right]\ge\text{Prob}\left[G\left(n,q\right) \text{ has } Q\right]\ge 1-\epsilon.$$

Note that we cannot say that $G\left(n,q\right)$ has the property Q only if one of the $G\left(n,p\left(\varepsilon\right)\right)$ has the property Q even though $G\left(n,q\right)$ is the union of the $G\left(n,p\left(\varepsilon\right)\right)$. $G\left(n,q\right)$ might have the property even though none of the $G\left(n,p\left(\varepsilon\right)\right)$ have the property. For example, if Q was the property that the graph has at least two edges.

We assert based on this that $p(1/2)$ is a threshold. First consider any $\tilde{p}(n)=o(p(1/2))$. We will show that the probability that $G\left(n,\tilde{p}(n)\right)$ has Q cannot go to any positive real. If the probability that $G\left(n,\tilde{p}(n)\right)$ has Q goes to a positive real ϵ in the limit, then for n large enough, the probability is at least $\epsilon/2$. The above argument can be modified to show that $p(1-\epsilon/2)\le mp(\epsilon/2)$ for some m depending only on ϵ, not on n. We would then have that $p(1/2)\le p(1-\epsilon)\le m\tilde{p}(n)$. This contradicts $\tilde{p}(n)=o(p(1/2))$. A similar argument shows that for any $\tilde{p}(n)$ such that $p(1/2)=o(\tilde{p}(n))$, $G\left(n,\tilde{p}(n)\right)$ almost surely has Q finishing the proof.

■

3.1.5 **Phase transitions for CNF-sat**

Phase transitions occur not only in random graphs, but in other random structures as well. In fact, any monotone property in a combinatorial structure has a phase transition. Here, we give an important example, that of satisfiability for a Boolean formula in conjunctive normal form.

Generate a random CNF formula *f* with *n* variables, *cn* clauses, and *k* literals per clause. Throughout this section, *n* will denote the number of variables in the initial formula we are given and the number of clauses will be *m=cn*, where, *c* is some constant. For fixed a fixed number of literals per clause, the probability that the function is satisfiable is a function of the number of variables and the number of clauses. When the number of clauses is very small, then intuitively, the function is almost surely satisfiable.
On the other hand if the function has a large number of clauses and only a small number of variables, then it is almost surely not satisfiable. As the number of clauses increases relative to

the number of variables, there is a sharp transition in the probability of satisfiability of the function. It is an open problem to precisely determine this threshold.

A random *k*-CNF formula with *n* variables and *cn* clauses has a threshold r_k such that if $c < r_k$, the formula almost sure is satisfiable and if $c > r_k$, the formula almost sure is not satisfiable. Researchers have been able to show that such a threshold exists. But its exact value is still unknown and is the subject of current research. To get an upper bound on r_k, observe that for any fixed truth assignment, a random clause is satisfiable with probability $1-\frac{1}{2^k}$. Thus, *cn* independent clauses are satisfiable 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 $2^n\left(1-2^{-k}\right)^{cn}$. If $c = 2^k \ln 2$, the expected number of satisfying assignments is

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

When *k* is moderately large $\left(1-\frac{1}{2^k}\right)^{2^k}$ can be approximated by $1/e$. Thus

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

For $c > 2^k \ln 2$, the expected number of satisfying assignments goes to zero as $n \to \infty$ and thus a random formula with *rn* clauses is almost surely not satisfiable.

The other direction, showing a lower bound for r_k is not that easy. It turns out that the second moment method cannot be directly applied because the variance is too high.
But, a simple algorithm called the unit clause heuristic yields a satisfying assignment with high probability if $c < \frac{2^k}{k}$, thus proving that $r_k \geq \frac{2^k}{k}$. Thus, the threshold is between $2^k/k$ and $2^k \ln k$. For k=3 these bounds are 2.667 and 5.5452.

The unit clause heuristic repeatedly executes the following iteration. If there is a clause with only one literal, then that literal must be set to true. Thus, the heuristic picks a one unit (one literal) clause and sets that literal to true. If no one literal clause exists, the heuristic selects a literal uniformly at random and sets it to true. If a literal w is set to true, the heuristic deletes all clauses containing w since they are satisfied and deletes $\bar{w}$ from any clause containing $\bar{w}$.

Example 3.2: Consider a 3-CNF formula with *n* variable and *c* clauses. With *n* variables there are *2n* literals since a variable and its complement are distinct literals. If for some constant α, the number of clauses is $c = \alpha n$, then the expected number of times a literal occurs can be calculated as follows. Each clause has three literals. Thus, each of the *2n* different literals occurs $\frac{3\alpha}{2}$ times on average. Suppose $\alpha = 5$. 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 is not easy to repeat – for one thing there is now conditioning so that the formula is no longer random.

■

The proof that if the density c (number of clauses / number of variables) is small enough, then the unit clause heuristic succeeds in finding satisfying assignment is not simple. The major difficulty in such proofs in general is the conditioning. While the formula is random at the start, the formula after one or more steps depends upon the literals set to true and so is not made up of independently chosen clauses. Many algorithms in this subject tend to be simple, like the unit clause heuristic. But the proofs of correctness need to deal with conditioning and tend to be complicated.

In the case of the unit clause heuristic, the argument to deal with conditioning is relatively simple and we go over it as illustrative of this class of arguments. Indeed, what we can show in this case is that the formula f_t we have after t literals have been set to be true is ``as independent as it can be'', in the sense that it is uniform independent conditioned on the number of clauses of different cardinalities in f_t. **EXPAND**

Before we state the result precisely, here is the argument. Imagine each clause in the original formula f is written on k cards, one literal of the clause per card. We turn the cards upside down, so the literals are not visible. Say the cards are arranged in m rows, one per clause. If a step sets a literal w to true, then all cards with w or $\overline{w}$ written on them are turned up; the rows with w in them are discarded (the clauses are already satisfied) and the $\overline{w}$ cards are thrown out, but the rows containing them are otherwise left as they are. The crucial claim is that what is left upside down is still random, since it was random at the start. One has to be a little careful though – the number of clauses of different sizes of course does not conform to the original distribution, which had precisely m clauses, all with k literals in them. But the only seen (non-random) effect of the steps is that we know the number of clauses of different sizes in f_t and of course the literals that have been set to be true so far. Conditioned on this information, f_t is random. This is precisely stated now.

Lemma 3.2: Suppose f is k-CNF formula chosen as follows: each clause is independently picked from among the $\binom{n}{k}2^k$ possible k-literal clauses on n variables with uniform probabilities. Suppose after t steps of the algorithm, the remaining set of variables is S, with $|S| = n - t$ and suppose the remaining formula f_t has $C_i(t)$ clauses with i literals for $i = 0,1,\ldots k$. If $|C_0(t)| \geq 1$, then we have failed in finding a satisfying assignment. Then, F_t is independent uniformly distributed conditioned on $C_i(t)$ - i.e., the distribution of f_t is the same as picking all independently and uniformly - $C_i(t)$ clauses from among the possible $\binom{n-t}{i}2^i$ clauses on the variables in S for $i = 0,1,\ldots k$.

■

With the Lemma on hand, we can proceed with the proof that the algorithm succeeds with high probability provided the density is low enough. We only deal with 3-SAT here, where there are 3 literals per clause to start with.

The argument (which we will not present in its full rigor) starts with the following assertion: if the (expected) number of new 1-literal clauses generated at each step is at most $1-\delta$ for some fixed $\delta > 0$, then since we always reduce the number of 1-literal clauses by 1 (if the number is non-zero), the number of 1-literal clauses will stay bounded by a constant. This follows from a basic queuing theory result which we state without proof.

Theorem 3.6: If a number x_t of jobs arrives into a queue at time t, where the x_t are independent, each Poisson with mean $1-\delta$ for some $\delta > 0$ and there is a ``server'' who at each time, serves one job if there is any in the queue, then the total number of unserved jobs with high probability remains bounded by a constant.

Further the set of one literal clauses are independent by Lemma 3.xxx. It is easy to see that if we pick independently $o(\sqrt{s})$ one literal clauses at random on s variables, then with high probability, we will not get two contradictory clauses (two clauses of the form w and $\bar{w}$) which is the only way to terminate without a satisfying assignment.

So it suffices to upper bound the number of one literal clauses generated. Let a_t be the number of new one literal clauses generated at time t conditioned on whatever happened in the first $t-1$ steps. We claim that

$$E(a_t \mid \text{what happened in previous steps}) = \frac{C_2(t-1)}{n-t+1}.$$

New one literal clauses can only be generated from 2-literal clauses and only when we set the negation of one of the two literals in the clause to true. By the lemma, any one of the remaining $2(n-t+1)$ literals is equally likely to be picked to be set to true at step t and hence the claim.

Thus, we now need to understand the behavior of $C_2(t)$. Let b_t denote the number of new 2-literal clauses generated at time t. Clearly,

$$C_2(t) = \sum_{t=1}^{t} b_t.$$

But, b_t is the sum of indicator random variables $x_1(t), x_2(t), \ldots x_m(t)$, where $x_j(t)$ is 1 or 0 depending on whether the jth clause in the input formula became a 2-literal clause precisely at time t. We claim that the $x_j(t)$ are independent Bernoulli random variables with probability of 1 being $\dfrac{\binom{n-t}{2}}{2\binom{n}{3}}$. This is because for $X_j(t)$ to be 1 (conditioned on what happened in the first $t-1$ steps), the clause must have 2 literals among the $n-t$ ``untouched'' variables giving us the numerator. Also, the third literal in the clause must be the negation of the one being set to true at step t; this gives us the two in the denominator.

Assume for the moment that the random variables take on their expected value. Detailed, but elementary calculations show that the conditions are satisfied that lets us use the queuing theory result to conclude that we end with a satisfying assignment provided c<2/3. If c were very close to 2/3, each variable on average appears in 2 clauses.

A rigorous argument involves a study of the actual distribution of the random variables and we do not present it here. But note that given the independence Lemma, the random variables have reasonable distributions.

For a CNF formula where the number of clauses is well below the threshold, in addition to the theoretical results, it has been empirically observed that one can efficiently determine satisfying assignments by any one of a number of algorithms. As the number of clauses increases there is a threshold where the formula is still almost surely satisfiable but it becomes computationally hard to find a satisfying assignment. As the number of clauses increases even further, a phase transition occurs and the formula almost surely is not satisfiable. When the number of clauses is above the threshold for unsatisfiability, even well above the threshold, but with a linear number of clauses, there is no known efficient algorithm to prove that the formula is not satisfiable. One difficulty is that to show satisfiability, it suffices to display one satisfying assignment, whereas, (unless NP=Co-NP), we do not know in general short (polynomial length) proofs of unsatisfiability. However, for random formulae with $\Omega(n^{3/2})$ clauses and n variables, efficient proofs of unsatisfiability are known.

It has also been empirically observed that as the number of clauses increases, the space of satisfying solutions changes character. For a small number of clauses the solution space is connected. Then a transition occurs and the solution space fractures into a set of disconnected regions of satisfiable assignments. As the number of clauses increases further, all of the disconnected regions of satisfiable assignments vanish. While there is a wealth of empirical results, rigorous proofs of such statements are still very hard to come by.