3.1.7 The Giant Component

When one looks at the connected components of large graphs that appear in various contexts, one observes that there is often one very large component. One example, is a graph formed from a data base of protean interactions where vertices correspond to proteins and edges correspond to pairs of proteins that interact. (By an interaction one means two amino acid chains that bind to each other for a function.) The graph has 2735 vertices and 3602 edges. At the time we looked at the data base, the associated graph had the number of components of various sizes shown in Table 1.

SIZE OF COMPONENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	…	1851
NUMBER OF COMPONENTS	48	179	50	25	14	6	4	6	1	1	1	0	0	0	0	1		1

Table 1: Size of components in the graph implicit in the database of interacting proteins. Science 1999 July 30 285 pp751-753.

There are a number of small components but only one component of size greater than 16 and that is a giant component of size 1851. As more proteins are added to the data base the giant component will grow even larger and eventually swallow up all the smaller components.

The existence of a giant component is not unique to the graph produced from the protein data set. Take any data set that one can convert to a graph and it is likely that the graph will have a giant component provided that the ratio of edges to vertices is a small number greater than one. Table 2 lists two other examples. This phenomenon of the existence of a giant component in many real world graphs deserves study.

ftp://ftp.cs.rochester.edu/pub/u/joel/papers.lst

Vertices are papers and edges mean that two papers shared an author.

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

http://www.gutenberg.org/etext/3202

Vertices represent words and edges connect words that are synonyms of one another

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

Table 2

Consider the random graph $G(n, p)$ where *p* equals d/n, d a constant independent of *n*. Note that the expected degree of each vertex is *d* (actually it is $d(n-1)/n$ which is very close to *d* for large *n*). Many large graphs have constant average degree, so this is an important range.

If d<1, almost surely, each connected component of the graph is of size $O(\ln n)$. At $d = 1$, the graph may have larger components, in fact, the expected number of components of size at least $\Omega(n^{2/3})$ is greater than 1. This does not imply that almost surely there is such a component since the components may be concentrated on a small fraction of the graphs. For d>1, the seminal result in random graphs says that there is a unique giant component of size $\Omega(n)$ and all other components are of size $O(\ln n)$. The overall process involves a phase transition at *d*=1 where the size of the connected components grows from $O(\ln n)$ with the emergence of the giant component that is of size $\Omega(n)$. The process involves a ``double jump'' where the size of components go from $O(\log n)$ to a number of components of size $\Omega\left(n^{\frac{2}{3}}\right)$ to a unique giant component of size $\Omega(n)$. More detailed studies are needed to see what happens at the ``critical point'' $d = 1$, i.e., when $d = 1 \pm o(1)$. Finally, note that we already showed that for $p \geq \frac{\log n}{n}$, the graph consists of a single connected component.

To compute the size of a connected component of $G(n, p)$ do a search of a component starting from a random vertex and generate an edge only when the search process needs to know if the edge exists. In order to maintain independence of edges, the search process only inquires about the existence of edges from the vertex being searched to vertices not yet discovered.

Consider a search of a connected component of $G\left(n, \frac{d}{n}\right)$ starting at a random vertex. Mark the start vertex discovered. Then select an unexplored vertex from the set of discovered vertices and add all vertices that are connected to this vertex to the set of discovered vertices and mark the selected vertex as explored. Discovered but unexplored vertices are called the frontier. The algorithm terminates when the frontier becomes empty.

Let z_i be the number of vertices discovered in the first *i* steps of the search. The distribution of z_i is $\text{binomial}\left[n-1, 1-\left(1-\frac{d}{n}\right)^i\right]$. To see this, observe that the probability a given vertex is not adjacent to the vertex searched in a given step is $1-\frac{d}{n}$. The probability that it is not discovered in any of *i* steps is $\left(1-\frac{d}{n}\right)^i$ and thus that it is discovered within *i* steps is $1-\left(1-\frac{d}{n}\right)^i$. From this it follows that the distribution of z_i is $\text{binomial}\left[n-1, 1-\left(1-\frac{d}{n}\right)^i\right]$.

Write $1-\left(1-\frac{d}{n}\right)^i$ as

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

Thus, the expected value of z_i is $s = n\left(1-e^{-\frac{d}{n}i}\right)$. At each unit of time a vertex is marked as explored removing it from the frontier. The expected size of the frontier after i steps of the search is

$$s-i = n\left(1-e^{-\frac{d}{n}i}\right)-i$$

and the normalized size of the frontier is

$$\frac{s-i}{n} = 1-e^{-\frac{d}{n}i}-\frac{i}{n}$$

Replacing $\frac{i}{n}$ by *x* yields $f(x) = 1-e^{-dx}-x$ where $f = \frac{s-i}{n}$ is the normalized size of the frontier and *x* is the normalized time.

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

Examining the expected normalized size $f(x) = 1-e^{-dx}-x$ of the frontier, we see that for $d > 1$ the second derivative of $f(x)$ is always negative and thus $f(x)$ is concave. The function $f(x)$ is zero at x=0 and negative at x=1. Thus, $f(x)$ has a unique root $\Theta^\dagger$ in (0,1] and is positive between x=0 and x= Θ. For x> Θ, $f(x)$ is negative. For d=1, $f(x)$ has a double root at x=0 and is negative for x>0.

† Θ is capital theta.

d	Θ
1.2	0.311
1.4	0.512
1.6	0.642
1.8	0.732
2.0	0.799

Table 3.2: The normalized size of the giant component Θ in $G\left(n, \frac{d}{n}\right)$ for various values of the expected degree *d*.

The actual size of the frontier is a random variable. What is the probability that the actual size of the frontier will differ from the expected size of the frontier by a sufficient amount so that the actual size of the frontier is zero? For small $i, i = O(\log n)$ and $d > 1$, the expected size of the frontier grows as $(d-1)i$. In this range of i, the binomial distribution can be approximated by the Poisson distribution $p(k) = e^{-di}\frac{(di)^k}{k!}$. Thus, the probability that the size of z_i differs from its expected value by $(d-1)i$ so that the size of the frontier is zero, drops off exponentially fast with i. This explains the distribution of sizes of the small components. See Illustration 3.1.

For larger *i*, near the point Θ at which the expected value of the size of the frontier is again zero, the expected value of the absolute value of the frontier increases linearly with $|i-\Theta|$. For the actual frontier to be zero would require that z_i deviate from its expected value by an amount proportional to the distance *i* is from Θ. For values of i near Θ, the binomial distribution can be approximated by a Gaussian distribution. The Gaussian distribution falls off exponentially fast with the square of the distance from its mean. Thus, for the frontier to be zero, the value of $k = |\Theta - i|$ must be small. The distribution falls off proportional to $e^{-\frac{k^2}{\sigma}}$ where σ is the standard deviation and is proportional to *n*. Thus, to have non vanishing probability, *k* must be at most $O\left(\sqrt{n}\right)$. This implies that the giant component is in the range $\left[\Theta - O\left(\sqrt{n}\right), \Theta + O\left(\sqrt{n}\right)\right]$. Thus, either the component is small, $O(\log n)$ or it is the giant component in the range $\left[\Theta - O\left(\sqrt{n}\right), \Theta + O\left(\sqrt{n}\right)\right]$.

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

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

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

Illustration 3.1.

This gives an intuitive explanation of the sizes of components. The argument above can be made rigorous to show that the frontier is highly unlikely to be zero for sizes in the range $[\Omega(\ln n), O(n)]$ showing that no component is likely to be in this range. But this does not prove that there is at least one large (larger than say $\Omega(n^{2/3})$) component when d>1, since all components may have stopped at $O(\ln n)$. To prove the existence of the giant, we re-examine the tree from the point of the exploring vertex rather than the explored vertex.

John: END OF REVISIONS. What follows up to Proof of Theorem 2 is unchanged from the version you have. Proof of Theorem 2 is changed. Also, the section on Critical Point d=1 is (edited version) of your notes on tree components.

Start a breadth first search at the vertex v_0. When the search stops with an empty frontier, the set of discovered vertices is precisely the vertices of the connected component in the graph G(n,p) containing the first vertex v_0.

Consider the random variable x_k that is the number of vertices discovered while exploring the k^{th} vertex. Set $x_0=1$. The total number of vertices discovered after the k^{th} vertex has been explored is $\sum_{i=0}^{k} x_i$. Let $m=n-\sum_{i=0}^{k-1} x_i$. Then x_k is the sum of m independent Bernoulli random variables, each with probability p of being 1. Thus

$E(x_k \mid x_1 + x_2 + \ldots + x_{k-1} = n - m) = mp \approx np = d$ provided $m \approx n$, i.e., provided that so far only a small fraction of the n vertices have been discovered. So, $E(\sum_{i=0}^{k} x_i) \approx dk$.

If d>1, then dk>k. Thus, (d-1)k which is the expected number of discovered but unexplored vertices is strictly positive. Thus, the frontier will likely be non-empty since we will show that the random variable - the size of the frontier- is concentrated around its expectation provided k is not too small, i.e., provided k is $\Omega(\ln n)$. Thus, the process will not stop as k increases above $\Omega(\ln n)$ at least until we reach a point when the above approximation $n \approx m$ fails to hold. We will see that in fact we can go up to $k = O(n)$.

Consider the case d<1. Assuming that we have already explored k vertices, we would have explored in expectation less vertices than we have already discovered since (d-1)k<0, so the expected size of the frontier is negative! Again for $k = \Omega(\ln n)$, the size of the frontier is concentrated about its mean, so it must be negative, leading to a contradiction. We formalize this intuitive line to show that for d<1, with high probability, there is no component of size $\Omega(\ln n)$. To make these arguments rigorous, requires understanding two things – one is the conditioning on $x_1 + x_2 + \ldots x_{k-1} = n - m$ in defining the distribution of x_k and the second is the $\approx$ sign above. The following theorem makes these ideas rigorous.

Theorem 1 Let p=d/n.

(i) If d<1, then the probability that G(n,p) has some component of size more than $\frac{4 \ln n}{(1-d)^2}$ is at most 1/n.

(ii) For d>1, the probability that there is a connected component of size between $c_2 \ln n$ and $c_3 n$ is at most 1/n, where c_2 and c_3 are fixed constants.

(iii) For d>1, the probability that there are two or more connected components, each of size more than $n^{2/3}$ is at most 1/n.

Remark: (ii) and (iii) together prove that with high probability, we have some components of size $O(\ln n)$ plus **possibly** one component of size $\Omega(n^{2/3})$. But, they still leave open the possibility that there is no component at all of size $\Omega(\ln n)$. It turns out that there is a unique ``giant component'' of size $\Omega(n)$ and indeed this is the most interesting result on random graphs. This will be proved as Theorem 3 later. What would remain to be proved for this is that there is in fact some large component.

Proof: First consider the case d<1. The connected component containing the starting vertex has at least *k* vertices if and only if $\sum_{i=1}^{k} x_i \geq k - 1$. The probability of this is very low. As

noted earlier, the distribution of x_i is Binomial $(n - \sum_{j=0}^{i-1} x_j, p)$. Introduce new mutually independent random variables $y_1, y_2, \ldots y_k$, each with the same distribution – Binomial (n, p); then it is clear that $x_i \le y_i$. Make the *nk* independent ``coin tosses'' needed for the y_i; then to determine x_i, which is in Binomial (s, p), use the results of the first s trials we already did for y_i. Make sure you understand this is correct, i.e., we get the correct distribution for the random variables x_i. Note that the y_i are mutually independent, even though x_i are not and also the y_i are not independent of the x_i.]

Since $x_i \le y_i$,

$$\text{Prob}\left(\sum_{i=0}^{k} x_i \ge k-1\right) \le \text{Prob}\left(\sum_{i=0}^{k} y_i \ge k-1\right).$$

Since the y_i are independent , $\sum_{i=1}^{k} y_i$ has the distribution Binomial (nk, p). Now we use the following known concentration result for binomially distributed random variables :

This will be moved to the appendix later.

If x has distribution Binomial (m, q), and t is any positive real, then

$$\text{Prob}\,(x \ge mq + t) \le \exp\left(-\frac{t^2}{2(mq + (t/3))}\right) \quad \text{.....(1)}$$

$$\text{Prob}\,(x \le mq - t) \le \exp\left(-\frac{t^2}{2mq}\right). \quad \text{.....(2)}$$

Will be moved to Appendix later.

Apply the first inequality to upper bound $\text{Prob}\,(\sum_{i=1}^{k} y_i \ge k-1)$. For this, set $m = nk$, $q = p$, and $t = k - 1 - kd = k(1-d) - 1$. Thus

$$\text{Prob}\,(\sum_{i=1}^{k} y_i \ge k-1) \le \exp\left(-\frac{((1-d)k-1)^2}{2(dk + (1-d)k/3)}\right) \le \exp\left(-\frac{(1-d)^2}{2}k\right).$$

This bounds the probability that a particular vertex v_0 is in a connected component of size at least k. Thus, the probability that there is some connected component of size at least k is at most $ne^{-(1-d)^2 k/2}$. Putting $k = \frac{4 \ln n}{(1-d)^2}$, we see by direct calculation that the probability that

there is some connected component of size at least k is at most 1/n. [This requires k to be at least $\Omega(\ln n)$ to ``eat up'' the factor of n.]

Consider the case when d>1. Here, for k in the range $\Omega(\ln n)$ to $O(n)$, the number of frontier vertices after k explorations will be at least $\frac{d-1}{2}k$ which is strictly greater than k, so the process will have to continue. This will be used to prove (ii) and (iii).

Lemma For k with $\Omega(\ln n) \le k \le O(n)$, the probability that after k explorations, there are fewer than $(d-1)k/2$ frontier vertices is at most $1/n^3$.

Proof : The frontier is of size less than $\frac{d-1}{2}k$ if and only if $\sum_{i=0}^{k} x_i < \frac{d+1}{2}k$. Again to avoid dependence, we introduce mutually independent random variables $z_1, z_2, \ldots z_k$ each with distribution Binomial $(n-\frac{d+1}{2}k, p)$ and we have $x_i \ge z_i$. Similar to the previous argument, we can argue now that

$$\text{Prob}(\sum_{i=0}^{k} x_i < \frac{d+1}{2}k) \le \text{Prob}(\sum_{i=0}^{k} z_i < \frac{d+1}{2}k).$$

We now use the inequality (2) to upper bound $\text{Prob}(\sum_{i=1}^{k} z_i < \frac{d+1}{2}k)$. For this note that $\sum_{I=1}^{k} z_i$ is binomial $(k(n-\frac{d+1}{2}k), p)$, so we put in (2) $m = k(n-\frac{d+1}{2}k)$, $q = p$. We need to simplify the value of t. Now we use the assumption that $k \le O(n)$. More specifically, we assume that

$k \le \frac{d-1}{9d(d+1)}n$. Using this we get:

$$t = mp - \frac{d+1}{2}k + 1 = k(n - \frac{d+1}{2}k)\frac{d}{n} - \frac{d+1}{2}k + 1 = dk - \frac{d+1}{2}k^2\frac{d}{n} - \frac{d+1}{2}k + 1$$
$$\ge \frac{d-1}{2}k - \frac{d-1}{18}k + 1 \ge \frac{4(d-1)k}{9}$$

Now using (2), we get

$$\text{Prob}\left(\sum_{i=0}^{k} x_i < \frac{d+1}{2}k\right) \le \text{Prob}\left(\sum_{i=0}^{k} z_i < \frac{d+1}{2}k\right) \le \exp\left(\frac{-(d-1)^2 k}{11d}\right),$$

Now we use the lower bound on k, namely $k \ge \Omega(\ln n)$ to ensure that the last expression is at most $1/n^3$.

■

Now to prove (ii) of the theorem, just observe that this bounds the probability that the bfs started from one particular vertex terminates at one particular value of k between $\Omega(\ln n)$ and $O(n)$ is at most $1/n^3$. So the probability that any one of the bfs's started at one of the n vertices terminates for any k in this range is at most 1/n, by the union bound. [Note: there is dependence between these events, but the union bound still applies.]

To prove (iii): suppose a pair of vertices u,v belong to two different connected components, each of size at least $n^{2/3}$. We will show that with high probability, they should have merged into one producing a contradiction. First run the bfs process starting at v for $n^{2/3}/2$ steps. By the above, with high probability, there are $\Omega(n^{2/3})$ frontier vertices. Further by the assumption, u does not belong to the connected component. Now temporarily stop the bfs tree of v and start and do a bfs tree starting at u, again for $n^{2/3}/2$ steps. [It is important to understand that this change of order of building $G(n,p)$ is OK. As we observed earlier, we can choose edges in any order as long as we maintain independence and make sure there is no conditioning.] The bfs tree from u also will have with high probability $\Omega(n^{2/3})$ frontier vertices. Now grow the u tree further. The probability that none of the edges between the two frontier sets is put in is $(1-p)^{O(n^{4/3})} \le e^{-O(dn^{1/3})}$ which clearly converges to zero. So almost surely, one of the edges is put in and u and v end up in the same connected component. This argument shows for a particular pair of vertices, u and v, the probability that they belong to different large connected components is very small. Now use the union bound to conclude that this does not happen for any of the $\binom{n}{2}$ pairs of vertices. The details are left to the reader.

■

The most interesting part perhaps is the fact that for d>1, there is a unique ``giant component'' consisting of a fraction of the vertices with other components of size $O(\ln n)$. We will prove this by showing in the next section (using Branching Processes) the following:

Theorem 2: Assume d>1. The probability that a breadth first search started at a vertex terminates before its connected component reaches a size $\Omega(n^{2/3})$ is at most a constant c_0 strictly less than 1 (for n large enough).

We show that the main result – Theorem 3 below simply follows from Theorem 1 and 2:

Theorem 3: Assume d>1. Then almost surely, at most $c_0 n + o(n)$ vertices lie in connected components of size $O(\ln n)$ and the rest of the vertices (there are at least $(1-c_0)n + o(n)$ of them) all lie in one giant component.

Proof: For vertex i, let x_i be an indicator random variable which is 1 if and only if it lies in a component of size $O(\ln n)$. First we see that from Theorem 2, $E(x_i) \le c_0$, so the expected number of vertices in $O(\ln n)$ size components is at most $c_o n$.

We use the second moment method to assert that the probability that the number of vertices in $O(\ln n)$ size components exceeds $c_0 n + o(n)$ goes to zero which proves the theorem. For this, let $x = \sum_i x_i$. We want to bound the variance of x.

$$\mathrm{Var}(x) = E\left((\sum_i x_i)x \right) - (Ex)^2 = \sum_i \mathrm{Prob}(x_i = 1) E(x \mid x_i = 1) - (Ex)^2$$

Conditioned on i belonging to a component of size $O(\ln n)$, the expected number of vertices in $O(\ln n)$ size components is at most

size of i 's component + n Prob(bfs started in a graph of size [n-size of i's component] terminates in $O(\ln n)$ steps)

which is at most

$O(\ln n)$ + n Prob(bfs started in a graph of size [n- $O(\ln n)$] terminates in $O(\ln n)$ steps) $\le c_0 n + o(n)$.

Plugging this in, we get that $\mathrm{Var}(x) \le c_0^2 n^2 + o(n^2)$ -$(c_0 n)^2 = o(n^2)$. Thus it follows (from Chebychev inequality) that almost surely at most $c_0 n + o(n)$ vertices belong to $O(\ln n)$ size components proving Theorem 3.

Branching Processes

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

We analyze a simple case of a branching process where the distribution of the number of children of each node is the same for every node in the tree. 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''.

An important tool in our analysis of branching processes is the generating function. The generating function for a non-negative random variable y where p_i is the probability that y

equals *i* is $f(x)=\sum_{i=0}^{\infty} p_i x^i$. The reader not familiar with generating functions should consult the appendix.

Let the random variable z_j be the number of children in the j^{th} generation and $f_j(x)$ be the generating function for z_j. Then $f_1(x)=f(x)$ is the generating function for the first generation. The generating function for the 2nd generation is $f_2=f(f(x))$. 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

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

Note that for x_1+x_2 to have value zero, both x_1 and x_2 must have value zero, for x_1+x_2 to have value one, exactly one of x_1 or x_2 must have value zero and the other have value one, and so on. In general, the generating function for the sum of *i* independent random variables, each with generating function $f(x)$, is $f^i(x)$. The second observation is that the coefficient of x^i in $f_j(x)$ is the probability of there being *i* children in the j^{th} generation. Since these *i* children will contribute $f^i(x)$ in the $j+1^{st}$ generation, $f_{j+1}(x)=f_j(f(x))$. Note that if $f_j(x)=a_0+a_1x+a_2x^2+\cdots$, then $f_j(f(x))=a_0+a_1f(x)+a_2f^2(x)+\cdots$.

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

We now focus 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 non negative root? The derivative of $f(x)$ at $x=1$ is $f'(1)=p_1+2p_2+3p_3+\cdots$. Let $m=f'(1)$ be the slope of $f(x)$ at $x=1$.

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 for $x\in[0,1)$, $f(x)>x$.

If $m=1$ and $p_1=1$, then $f(x)=x$. 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) in the interval.

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.

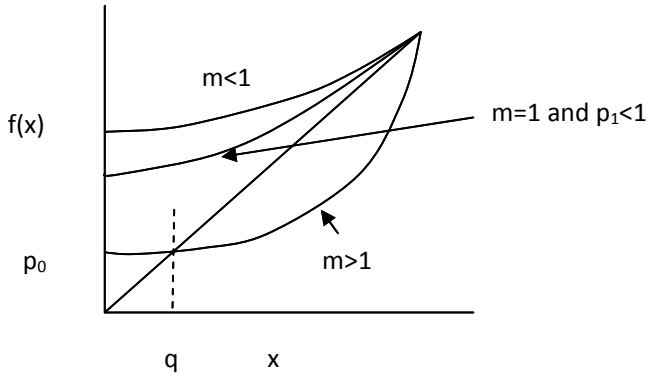


Figure XXX: Illustration of root of equation f(x)=x in the interval [0,1]

Recall that the random variable z_j is the number of children in the j[th] generation. Let q be the smallest non negative root of the equation f(x)=x. For $m < 1$ and for q=1 and $p_0 < 1$, q=1 and for m>1, q<1. 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, z_j, is zero. To see this, note that $\lim_{j\to\infty} f_j(x) = q$ for 0<x<1. The picture illustrates the proof of this. Similarly note that when $m < 1$, $f_j(x)$ approaches one as *j* approaches infinity.

Lemma: Assume $f'(1) > 1$. Let q be the unique root of f(x)=x in [0,1). Consider the j[th] iterate $f_j(x)$ as j goes to infinity. If $x \in [0, q)$ then $f_j(x)$ is monotonically increasing with j and converges to q. If $x \in (q, 1)$ then $f_j(x)$ is monotonically decreasing with j and converges to q. If x=q or x=1 then $f_j(x) = x$ for all j. Thus, the least non negative root of $f(x) = x$ is the extinction probability.

Proof: If $0 \le x \le q$ then x<f(x)≤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 $q \le x < 1$ then f(q)≤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, $0 \le x < 1$. Since in the limit $f_j(x) = q$, a constant independent of x, the extinction probability, $\text{Prob}(z_j = 0)$, equals q. For all finite i>0, $\text{Prob}(z_j = i) = 0$. Thus, 1-q is the probability that z_j grows without bound, i.e., immortality.

■

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 non-zero values of i.

Theorem 3.4: Consider a tree generated by a branching process. Let *y* be a random variable whose distribution is that of the number of children at each node.

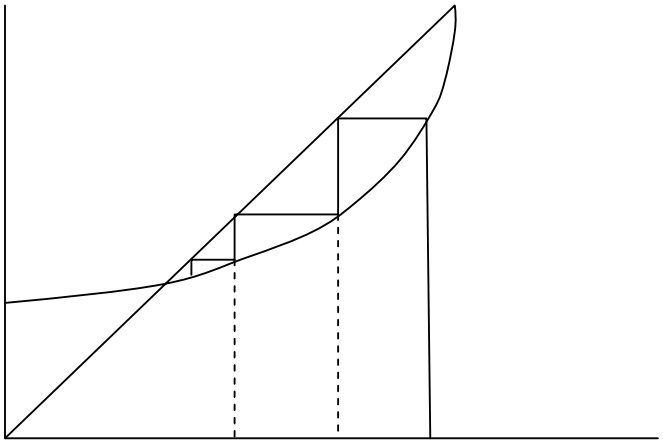
(i) If $E(y) \le 1$, the probability of extinction is 1 unless $\text{Prob}(y = 1) = 1$.

(ii) If $E(y) > 1$, the probability of extinction is the unique solution to $f(x) = x$ in $[0,1)$, where $f(x)$ is the generating function of y.

Proof: Follows from the above remarks.

■

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



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

Note that the branching process either dies out or goes to infinity. In biological systems there must be some other factor since processes seem to 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.

Proof of Theorem 2: In the branching process, we assume that the number of offspring has the same distribution at each node, whereas, in the bfs, the number of children is chosen from Binomial(s, p), where s is the number of as yet unexplored nodes. But the probability that the bfs terminates before $c_1 n^{2/3}$ is clearly at most the probability that the branching process with number of children at each node distributed as per Binomial$(n - c_1 n^{2/3}, p)$ becomes extinct. (Convince yourself of this.)

This finishes the proof of Theorem 2.

The Critical Point d=1

We now show that at d=1, the expected number of components (in fact even the expected number of tree components) of size at least $c_1 n^{2/3}$ goes to a positive real. This number can be increased by decreasing c_1. Thus, we can make the expected number of components in $G(n, \frac{1}{n})$ of this size an arbitrarily large constant. But note that this does not immediately

imply that almost surely, there is such a component. It may well be that with probability 1/1000, $G(n,p)$ has 1000 components of size $c_1 n^{2/3}$ and with probability 999/1000, it has none.

Let x_m be the number of components of G that are trees with at least m vertices. Now

$$E(x_m) = \sum_{k=m}^{n} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{\binom{k}{2}-(k-1)+k(n-k)}$$

[To specify a tree, first select *k* vertices out of the n vertices. Given the *k* vertices there are k^{k-2} possible ways to form a tree with *k* vertices **(**This is a basic fact which we do not prove here**).** Once a tree is formed each of *k-1* edges of the tree is present with probability *p*. Each of the other $\binom{k}{2}-(k-1)$ edges among the *k* vertices is absent with probability *1-p* and each of the $k(n-k)$ edges from the *k* vertices of the tree to the remaining *n-k* vertices is absent with probability *1-p*.]

We now make a series of approximations to derive a formula for $E(x_m)$ assuming again $p = \frac{d}{n}$ with $d \le 1$.

$$(1-p)^{\binom{k}{2}-(k-1)+k(n-k)} \cong \theta(1)\left[\left(1-\tfrac{d}{n}\right)^{n-\frac{k}{2}}\right]^k \cong \theta(1)\left[e^{\frac{kd}{2n}} e^{-d}\right]^k$$

Since $k \le n$ the term $\left(1-\frac{d}{n}\right)^{-\frac{3}{2}k+1}$ is bounded above and below by a constant and it was replaced in the above equation by the $\theta(1)$ term.

Next since $k! \cong \frac{k^{k+\frac{1}{2}}}{e^k}$ we can write $\binom{n}{k} = \frac{n(n-1)\cdots(n-k)}{k!} \cong \frac{n(n-1)\cdots(n-k)}{k^{k+\frac{1}{2}}} e^k$.

Substituting into $E(x_m)$ (and using the abbreviation $\cong$ again to mean within a constant factor), we get

$$E(x_m) \cong \sum_{k=m}^{n} \frac{n(n-1)\cdots(n-k)}{k^{k+\frac{1}{2}}} e^k k^{k-2} \left(\frac{d}{n}\right)^{k-1} \left(\frac{e^{\frac{kd}{2n}}}{e^d}\right)^k$$

$$\cong \tfrac{n}{d} \sum_{k=m}^{n} \frac{n(n-1)\cdots(n-k)}{n^k} \frac{1}{k^{\frac{5}{2}}} \left[e^{\frac{dk}{2n}} d e^{1-d}\right]^k$$

Now $\frac{n(n-1)\cdots(n-k)}{n^k} e^{\frac{dk^2}{2n}} = O(1)$. [We see this by using $1-\frac{t}{n} \le e^{-t/n}$ from which it follows that that the expression is at most exp($-\frac{k^2}{2n}+\frac{dk^2}{2n}$) which for $d \le 1$ is O(1).] Thus

$$E(x_m) \cong \sum_{k+m}^{n} \frac{n}{k^{\frac{5}{2}}} \left(de^{1-d}\right)^k \text{.....(1)}$$

Theorem: For $p = \frac{1}{n}$, the maximum size of a tree component is $O(n^{\frac{2}{3}})$. Also, for some constant c>0, the expected number of tree components of size at least $cn^{\frac{2}{3}}$ tends to a positive real.

Proof: For $p = \frac{1}{n}$, from (1), we see that the expected number of trees with at least m vertices is $\theta(1) n \sum_{k=m}^{n} \frac{1}{k^{\frac{5}{2}}}$. The summation, $\sum_{k=m}^{n} \frac{1}{k^{\frac{5}{2}}}$, is very close to $\int_{k=m}^{n} k^{-5/2} dk$ which is $(2/3)(m^{-3/2} - n^{-3/2})$. If $m > cn^{\frac{2}{3}}$, the summation converges to zero. Thus, with high probability no component is larger than $O(n^{\frac{2}{3}})$ proving the first statement. The second statement follows from the same calculation.

■

Exercises

Exercise 3.35: Find the average degree of some large graphs.

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

■

Exercise: Prove that the expected value of the absolute value of the size of the frontier increases linearly with *i* for i in the neighborhood of Θ.

■

Exercise: Consider the binomial distribution $\text{binomial}\left[n-1, 1-\left(1-\frac{d}{n}\right)^i\right]$ for $d>1$. Prove that as $n \to \infty$, the distribution goes to zero for all i except for i in the two ranges $[0, \log n]$ and $\left[\Theta-\sqrt{n}, \Theta+\sqrt{n}\right]$. **NEED TO GET CONSTAINTS IN RANGES CORRECT**

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

■

Exercise: Consider the branching process when m>1 with q as the root of the equation f(x)=x. Show that $\lim_{j\to\infty} f_j(x) = q$ for 0<x<1.

■

Exercise: We showed that $\lim_{j\to\infty} f_j(x) = q$ for 0<x<1 for the generating function f of a branching process when m>1. This implies (in the limit) $\text{Prob}\left(z_j = 0\right) = q$ and $\text{Prob}\left(z_j = i\right) = 0$

for all non-zero finite values of i. Shouldn't the probabilities add up to 1? Why is not a contradiction?

Exercise: Try to create a probability distribution which varies with the current population in which future generations neither die out nor grow to infinity.

Solution:

■

Exercise: Show that for a branching process with number of children distributed as $\text{Binomial}(n - c_1 n^{2/3}, \frac{d}{n})$, where d is a constant strictly greater than 1, the root of the generating function $f(x)$ in $(0,1)$ is at most a constant strictly less than 1. [In other words, it cannot for example be greater than $1-(1/n)$.]