CS271 Randomness & Computation **Fall 2022**

Lecture 16: October 18

Instructor: Alistair Sinclair

Disclaimer: *These notes have not been subjected to the usual scrutiny accorded to formal publications. They may be distributed outside this class only with the permission of the Instructor.*

16.1 The Giant Component in $\mathcal{G}_{n,p}$

In an earlier lecture we briefly mentioned the threshold for the existence of a "giant" component in a random graph, i.e., a connected component containing a constant fraction of the vertices. We now derive this threshold rigorously, using both Chernoff bounds and the useful machinery of *branching processes*. We work with our usual model of random graphs, $\mathcal{G}_{n,p}$, and look specifically at the range $p = \frac{c}{n}$, for some constant c. Our goal will be to prove:

Theorem 16.1 *For $G \in \mathcal{G}_{n,p}$ with $p = \frac{c}{n}$ for constant c, we have:*

1. *For $c < 1$, then a.a.s. the largest connected component of G is of size $O(\log n)$.*
2. *For $c > 1$, then a.a.s. there exists a single largest component of G of size $\beta n(1 + o(1))$, where β is the unique solution in $(0, 1)$ to $\beta + e^{-\beta c} = 1$. Moreover, the next largest component in G has size $O(\log n)$.*

Here, and throughout this lecture, we use the phrase "a.a.s." (asymptotically almost surely) to denote an event that holds with probability tending to 1 as $n \to \infty$.

This behavior is shown pictorially in Figure 16.1. For $c < 1$, G consists of a collection of small components of size at most $O(\log n)$ (which are all "tree-like"), while for $c > 1$ a single "giant" component emerges that contains a constant fraction of the vertices, with the remaining vertices all belonging to tree-like components of size $O(\log n)$.

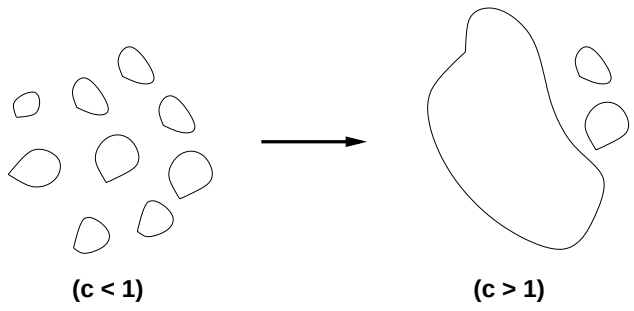

Figure 16.1: Evolution of the giant component

In the boundary case $c = 1$ things are more complicated. In this case the largest component is of size $O(n^{2/3})$, but we will not prove this here. Moreover, if $c = 1 + c'n^{-1/3}$ then the size of the largest component varies smoothly with c'. In other words, the "width" of the phase transition is $n^{-1/3}$.

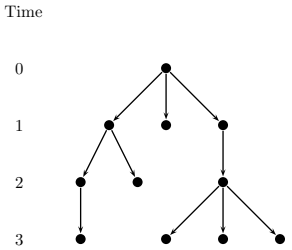

Figure 16.2: Galton-Watson Branching Process

Before proving this theorem, we develop some basic machinery from Galton-Watson branching processes that is of independent interest.

16.2 The Galton-Watson Branching Process

Let X be a random variable that takes non-negative, integer values. The branching process defined by X starts with a single node at time 0. At each subsequent time step, every node from the previous time step gives rise to a random number of children determined by X, independently of the other nodes. (See Figure 16.2.)

The random variable Z_i counts the number of nodes at time i. By definition, $Z_0 = 1$. Z_i is then distributed as the sum of Z_{i-1} independent copies of X. We can define the "extinction" of the process as the event that the number of nodes is eventually zero. More formally,

$$\Pr[\text{extinction}] = \lim_{n\to\infty} \Pr[Z_n = 0].$$

We will need the following basic fact about branching processes (see [GS01] or [Will91] for more background).

Theorem 16.2 *For a branching process defined by a non-negative integer-valued r.v. X satisfying the conditions $\Pr[X = 1] < 1$ and $\Pr[X = 0] > 0$, we have:*

- *If $\mathrm{E}[X] \le 1$ then $\lim_{n\to\infty} \Pr[Z_n = 0] = 1$, i.e., the process dies out a.a.s.*
- *If $\mathrm{E}[X] > 1$ then $\lim_{n\to\infty} \Pr[Z_n = 0] = p^* < 1$, where p^* is the unique solution in $(0, 1)$ to $f(x) = x$, where $f(x)$ is the probability generating function*

$$f(x) = \sum_{i\ge 0} \Pr[X = i]x^i.$$

The conditions $\Pr[X = 1] < 1$ and $\Pr[X = 0] > 0$ rule out trivial extreme cases in which the theorem is actually false. In particular, if $\Pr[X = 1] = 1$ then $\mathrm{E}[X] \le 1$ but clearly the process continues forever. And if $\Pr[X = 0] = 0$ then clearly the process continues forever with probability 1 regardless of the value of $\mathrm{E}[X]$.

For a topical illustration, note that in the context of epidemics, $\mathrm{E}[X]$ is the so-called "reproduction number" R_0, i.e., the expected number of currently healthy people that a typical sick person infects. The above theorem explains why getting this number below 1 is so important in containing the spread of the disease.

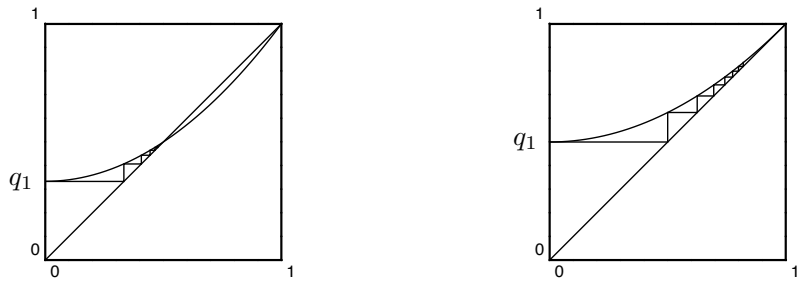

Figure 16.3: Generating functions for case 1 and case 2 respectively

16.2.1 Proof of Theorem 16.2

Let f_n be the probability generating function of the random variable Z_n, *i.e.*,

$$f_n(x) = \sum_{i \geq 0} \Pr[Z_n = i] x^i$$

Note that $f_1(x) = f(x)$ (where $f(x)$ is the generating function of X, as in the theorem above) since Z_1 has the same distribution as X. For $n > 1$, Z_n is distributed as the sum of Z_{n-1} many independent copies of X. By standard properties of generating functions we therefore have $f_n(x) = f(f_{n-1}(x))$ for all $n > 1$. (**Exercise:** Verify this, e.g., by comparing coefficients.)

Let the probability of extinction at time n be $q_n := \Pr[Z_n = 0] = f_n(0)$. Then we have the following recursive relation:

$$q_n = f(q_{n-1}) \text{ for all } n \geq 1,$$

where $q_0 = 0$. [This can also be proved directly as follows. If at time 1 the number of nodes is k, then there will be extinction at time n if and only if each of the k offspring gives rise to 0 children after $n-1$ more levels. Consequently $q_n = \Pr[Z_n = 0] = \sum_{k \geq 0} \Pr[Z_1 = k] \Pr[Z_{n-1} = 0]^k = \sum_{k \geq 0} \Pr[X = k] q_{n-1}^k = f(q_{n-1})$.]

As the probability of extinction at time n is at least as large as the extinction probability at time $n-1$, the sequence q_n is monotonically increasing with $0 < q_n \leq 1$ for all n, *i.e.,*

$$0 < q_1 \leq q_2 \leq q_3 \leq \cdots \leq 1.$$

Since the sequence (q_i) is increasing and bounded, it must converge to a limit; thus as $n \to \infty$, $q_n \to q^*$ where $0 < q^* \leq 1$. Also, the fact that f is continuous and $q_n = f(q_{n-1})$ for all $n \geq 1$ implies that q^* is a fixed point of the function $f(x)$, i.e., $q^* = f(q^*)$.

Observe that f is a strictly increasing convex function from $[0, 1]$ to $[0, 1]$ with $f(1) = 1$ and $f(0) > 0$. We have two different cases as depicted in Figure 16.3.

Case 1. The graph of $f(x)$ first crosses the line $y = x$ at a point $a < 1$. In this case a is the unique fixed point for $f(x)$ in $(0, 1)$ and q_n converges to a. (This can be seen by iterating f graphically, as in the left-hand panel in Figure 16.3.) Hence the extinction probability q^* is a. This corresponds to the second case in Theorem 16.2 above.

Case 2. The graph of $f(x)$ first crosses the line $y = x$ at 1. Note that $f(1) = 1$. $f(x)$ does not have a fixed point in $(0, 1)$. In this case q_n converges to 1 and the extinction probability q^* is 1. This corresponds to the first case in Theorem 16.2 above.

The two cases above are distinguished by the derivative of $f(x)$ evaluated at $x = 1$. Note that $f'(1) = \mathrm{E}[X]$. If $\mathrm{E}[X] > 1$ we are in case 1, while if $\mathrm{E}[X] \leq 1$ we are in case 2. The statement of Theorem 16.2 follows.

16.2.2 Sketch of application to random graphs

The number of neighbors of a node v in $G \in \mathcal{G}_{n,p}$ is distributed as $\mathrm{Bin}(n-1, p)$. Now consider exploring the connected component of v by revealing first the neighbors of v, then the (new) neighbors of each of these neighbors, and so on. This is just like a branching process, except that the number of offspring is not uniformly $\mathrm{Bin}(n-1, p)$ at every node: rather, it is $\mathrm{Bin}(n-m, p)$, where m is the number of nodes we have revealed so far. However, as long as m is not too large we might expect that this difference is not significant, so we can use the branching process based on $X \sim \mathrm{Bin}(n, p)$ to analyze the components of G.

What does Theorem 16.2 tell us about this branching process? Note that since $p = \frac{c}{n}$ we have $\mathrm{E}[X] = c$. Hence when $c < 1$ we would expect the process to die out a.s., which is in line with our claim that all components are small in this case. When $c > 1$ we would expect the process to continue for a long time with constant probability, which again is in line with our claim that a constant fraction of the nodes are in a giant component of size βn.

Let's see what the value of β should be. To do this, we use the fact that the probability generating function of $\mathrm{Bin}(n, \frac{c}{n})$ converges pointwise to that of $\mathrm{Poisson}(c)$. Hence the extinction probability for our branching process is the same as for that defined by a $\mathrm{Poisson}(c)$ r.v. But the probability generating function for the latter is

$$f(x) \quad = \quad \sum_i \frac{c^i e^{-c}}{i!} x^i = e^{c(x-1)}.$$

Therefore, the extinction probability p^* is the solution to the equation

$$e^{c(x-1)} = x. \tag{16.1}$$

Writing $\beta = 1 - p^*$ for the probability that a vertex is in the giant component, this equation becomes $e^{-c\beta} = 1 - \beta$, exactly as claimed in Theorem 16.1. This completes our intuition for Theorem 16.1, though of course the above argument is not rigorous. In the next section, we will turn this intuition into a proof.

16.3 Proof of Theorem 16.1

Following the above sketch and [JLR00], we will analyze the emergence of the giant component by linking it to the branching process described in the previous section. In order to analyze the size of the component containing a vertex v, we consider the branching process that starts with v and explores the graph in breadth-first manner. To explore from a vertex u, we reveal all its neighbors; we then say that u is "saturated" and the neighbors are "explored." The number of newly explored neighbors of u is a binomial random variable $\mathrm{Bin}(n-k, c/n)$, where k is the number of vertices already explored. Thus our exploration of the graph can be viewed as a *non-uniform* branching process where the offspring distribution depends on the history of the process. This process is depicted in Figure 16.4.

We will analyze this branching process (or, more specifically, uniform versions of it) in order to determine the size of the component containing the start vertex v. Recall from Theorem 16.2 that the behavior of the branching process depends on whether the mean of the offspring distribution, $\mathrm{E}X$, is less than or greater than 1. We begin with the first part of Theorem 16.1, which corresponds to the sub-critical case $c < 1$.

Proof of Theorem 16.1, Part 1: Consider the event that a particular vertex v is in a component of at least k vertices. This event happens only if the branching process starting at vertex v finds at least $k-1$ new

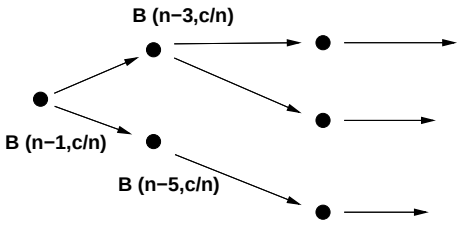

Figure 16.4: Branching process in $\mathcal{G}_{n,p=c/n}$

vertices after saturating k vertices. Note that this non-uniform branching process is stochastically dominated by a uniform branching process with offspring distribution $\mathrm{Bin}(n, c/n)$. This we can analyze by brute force as follows:

$$\begin{aligned}
\Pr[v \text{ in component of size } \geq k] &\leq \Pr\left[\sum_{i=1}^{k} X_i \geq (k-1)\right] \quad \text{where the } X_i \text{ are iid } \mathrm{Bin}(n, c/n) \\
&= \Pr\left[\sum_{i=1}^{k} X_i - ck \geq (1-c)k - 1\right] \\
&\leq \exp\left(\frac{-((1-c)k-1)^2}{c^2k^2(2+((1-c)k-1)/ck)}ck\right) \\
&= \exp\left(-\frac{(1-c)^2k}{2} + O(1)\right).
\end{aligned}$$

In the third line here we applied Angluin's form of the Chernoff bound with $\mu = ck$, $\beta = \frac{(1-c)k-1}{ck}$ to the sum $\sum_i X_i$, which is itself binomial $\mathrm{Bin}(nk, c/n)$.

Setting $k = \frac{3}{(1-c)^2}\ln n$, the above probability is bounded above by $O(n^{-\frac{3}{2}})$, so taking a union bound gives

$$\Pr\left[\exists \text{ a } v \text{ in component of size } \geq \frac{3}{(1-c)^2}\ln n\right] \leq O(n \cdot n^{-\frac{3}{2}}) = O(n^{-\frac{1}{2}}) \to 0$$

■

We turn now to Part 2 of Theorem 16.1, which is rather more involved. Fix $c > 1$, and define $k^- = c' \ln n, k^+ = n^{2/3}$ (for some constant c' that will be defined later). The heart of the analysis lies in the following claim.

Claim 16.3 *For all vertices v, a.a.s either:*

(i) The branching process starting at v terminates within k^- steps; or

(ii) $\forall k$ st $k^- \leq k \leq k^+$, after k steps of the branching process starting at v, there are at least $(c-1)k/2$ explored but unsaturated vertices.

Proof: For $k^- \leq k \leq k^+$, to ensure that there are at least $(c-1)k/2$ unsaturated vertices, note that it is sufficient to prove that there are at least $(c-1)k/2 + k = (c+1)k/2$ vertices explored from v in total. Call a vertex v *k-bad* if, after k steps, the branching process at v dies or has found fewer than $(c+1)k/2$ vertices. Then, noting that the branching process in this case is stochastically bounded below by a uniform branching

process with offspring distribution $\mathrm{Bin}(n - \frac{(c+1)k^+}{2}, c/n)$, we get

$$\begin{aligned}
\Pr[v \text{ is } k\text{-bad}] &\leq \Pr\left[\sum_{i=1}^{k} X_i \leq \frac{(c+1)k}{2} - 1\right] \quad \text{where the } X_i \text{ are iid } \mathrm{Bin}\Big(n - \tfrac{(c+1)k^+}{2}, c/n\Big) \\
&\leq \Pr\left[\sum_{i=1}^{k} X_i \leq ck - (c-1)k/2\right] \\
&\leq \exp\left(-\frac{(c-1)^2 k}{8c}\right),
\end{aligned}$$

where in the third line we have used the Angluin form of the Chernoff bound applied to the binomial r.v. $\sum_i X_i$, with $\mu = ck(1 - \frac{(c+1)k^+}{2n}) = ck(1 + o(1)) \approx ck$ and $\beta = (c-1)/2c$. (The small error in approximating μ by ck here can be absorbed into the bound.) Now by a union bound over k we get

$$\begin{aligned}
\Pr[v \text{ is bad}] &\leq \sum_{k=k^-}^{k^+} \exp\left(\frac{-(c-1)^2 k}{8c}\right) \\
&\leq n^{2/3} \exp\left(\frac{-(c-1)^2 k^-}{8c}\right) \\
&\leq n^{-4/3},
\end{aligned}$$

if we choose $k^- = \frac{16c \ln n}{(c-1)^2}$. Finally, by a union bound over v we get

$$\Pr[\exists v \text{ st } v \text{ is bad }] \leq n^{-1/3} \to 0 \quad \text{as } n \to \infty,$$

which completes the proof of the claim. ∎

Thus, a.a.s., the branching process starting from any vertex v either terminates within $k^- = O(\log n)$ steps, or goes on for at least k^+ steps. Call vertices of the first type "small" vertices, and the others "large" vertices. We will first prove that there is a unique component containing all the large vertices, and then we will estimate the number of small vertices. This will reveal the size of the giant component, as required.

Claim 16.4 *A.a.s., there exists a unique component containing all the large vertices.*

Proof: Consider two large vertices $u \neq v$. Run branching processes from u and from v separately. Let $U(v)$ denote the set of unsaturated vertices starting from v after k^+ steps; then $|U(u)| \geq \frac{c-1}{2}k^+$ and $|U(v)| \geq \frac{c-1}{2}k^+$. If the branching processes for k^+ steps from u and v have a vertex in common, then we are done. Otherwise, we will show that there is an edge between $U(v)$ and $U(u)$ whp:

$$\begin{aligned}
\Pr[\nexists \text{ an edge between } U(u) \text{ and } U(v)] &\leq (1-p)^{(\frac{c-1}{2}k^+)^2} \\
&\leq e^{-p(\frac{c-1}{2}k^+)^2} \text{ using } (1-p)^x \leq e^{-px} \\
&= e^{-\frac{(c-1)^2 c}{4} n^{\frac{1}{3}}} \text{ substituting } k^+ = n^{\frac{2}{3}}, p = \tfrac{c}{n} \\
&= o(n^{-2}).
\end{aligned}$$

Finally, we take a union bound over pairs u, v to conclude the proof:

$$\Pr[\text{for } \textit{any} \text{ pair of large vertices } u, v, \nexists \text{ an edge between } U(u) \text{ and } U(v)] = o(1).$$

∎

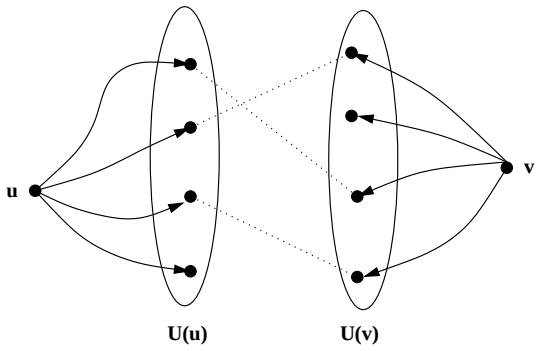

Figure 16.5: Illustration of the proof of Claim 16.4

We have established that there is a unique large component, and that all other vertices lie in components of size at most $O(\log n)$. It remains only to determine the size of the large component. We will do this by showing that the number of small vertices is αn, where α is a constant equal to $1-\beta$ (where β is the constant appearing in Part 2 of Theorem 16.1). This will immediately imply that the number of large vertices (and thus the size of the giant component) is βn.

Claim 16.5 *A.a.s. the number of small vertices is* $(1+o(1))(1-\beta)n$.

Proof: From the definition of "small" vertices, and following our usual strategy of bounding the branching process from v by uniform processes, we may conclude

$$\Pr\left[\text{b.p. with } \mathrm{Bin}(n,c/n) \text{ from } v \text{ dies in } k^- \text{ steps}\right] \le \Pr\left[v \text{ is small}\right] \le \Pr\left[\text{b.p. with } \mathrm{Bin}(n-k^-,c/n) \text{ from } v \text{ dies out}\right]$$

Moreover, since we know from Claim 16.3 that the b.p. on the lhs does not die out after k^- steps whp, we may write the lhs as $\Pr[\text{b.p. with } \mathrm{Bin}(n,c/n) \text{ from } v \text{ dies out}] + o(1)$. Using the notation $d(m,p)$ to denote the probability that a b.p. with offspring probability $\mathrm{Bin}(m,p)$ dies out, we may express the above more compactly as

$$d(n,c/n) + o(1) \le \Pr[v \text{ is small}] \le d(n-k^-,c/n).$$

Now recall from the Poisson example in Section 16.2.2 that $d(n,c/n) \to \alpha = 1-\beta$, where $\beta = \beta(c)$ is the unique solution in $(0,1)$ to the equation $\beta + e^{-\beta c} = 1$. Also, since $k^- \ll n$, the same holds for $d(n-k^-,c/n)$. [**Exercise:** verify this!] Therefore, by the above sandwiching, the same holds also for $\Pr[v \text{ is small}]$.

Now let $Z = \sum_v Z_v$ be the number of small vertices, where $Z_v = 1$ if v is small and 0 otherwise. Then $\mathrm{E}[Z_v] = \Pr\left[v \text{ is small}\right] \to \alpha(c)$ as $n \to \infty$ and thus $\mathrm{E}[Z] = \sum_v \mathrm{E}[Z_v] = (1+o(1))\alpha n$ asymptotically. This gives us the result we want in expectation. To get it in probability, we need to look at the second moment:

$$\begin{aligned}
\mathrm{E}[Z^2] &= \mathrm{E}[(\sum_v Z_v)^2] = \sum_v \mathrm{E}[Z_v^2] + \sum_{u \ne v} \mathrm{E}[Z_u Z_v] \\
&= \mathrm{E}[Z] + \sum_{u \ne v} \Pr\left[\text{both } u,v \text{ are small}\right] \\
&= \mathrm{E}[Z] + \sum_v \Pr\left[v \text{ is small}\right] \sum_{u \ne v} \Pr\left[u \text{ is small} \,|\, v \text{ is small}\right]. \qquad (16.2)
\end{aligned}$$

Now we may write

$$\begin{aligned}
&\sum_{u \neq v} \Pr\left[u \text{ is small} \,|\, v \text{ is small}\right] \\
= &\sum_{\substack{u \neq v \\ u,\, v \text{ in one comp.}}} \Pr\left[u \text{ is small} \,|\, v \text{ is small}\right] + \sum_{\substack{u \neq v \\ u,\, v \text{ in different comp.}}} \Pr\left[u \text{ is small} \,|\, v \text{ is small}\right] \\
\leq \; & k^- + nd(n - k^-, c/n). \qquad (16.3)
\end{aligned}$$

Here we have used the fact that there are at most k^- vertices in the same component as v and that, for vertices that are in a different component from v,

$$\Pr\left[u \text{ is small} \,|\, v \text{ is small}\right] = \Pr\left[u \text{ is small in } G \backslash \{u' : u' \text{ is in the same comp. as } v\}\right] \sim d(n - k^-, c/n).$$

Plugging (16.3) into (16.2) and noting that, since $k^- \ll n$, $d(n - k^-, \frac{c}{n}) \sim d(n, \frac{c}{n}) \to \alpha(c)$ as $n \to \infty$, we get

$$\mathrm{E}[Z^2] \leq \mathrm{E}[Z] + n^2\alpha^2(1 + o(1)) = \mathrm{E}[Z]^2(1 + o(1)),$$

since $\mathrm{E}[Z] = n\alpha(1 + o(1))$.

Hence $\frac{\mathrm{Var}[Z^2]}{(\mathrm{E}[Z])^2} = \frac{\mathrm{E}[Z^2]}{(\mathrm{E}[Z])^2} - 1 = o(1)$, so by Chebyshev,

$$\Pr\left[|Z - \mathrm{E}[Z]| > \gamma \mathrm{E}[Z]\right] \leq \frac{\mathrm{Var}[Z]}{\gamma^2 \cdot \mathrm{E}[Z]^2} = \frac{1}{\gamma^2} \cdot o(1) = o(1)$$

for a sufficiently slowly growing function $\gamma = \gamma(n)$. Hence $Z = (1 + o(1))\mathrm{E}[Z]$ a.a.s., which completes the proof of the second part of Theorem 16.1. ■

References

[GS01] G.R. Grimmett and D.R. Stirzaker, *Probability and Random Processes* (3rd ed.), Oxford Univ Press, 2001.

[JLR00] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley, 2000.

[Will91] D. Williams, *Probability with Martingales*, Cambridge Univ Press, 1991.