In this problem set we characterize the extinction probability of the Galton-Watson branching process, and prove the threshold behavior of the size of the largest component of random graphs.

1. The Galton-Watson branching process corresponding to a random variable $X$ that takes non-negative integer values is defined as follows: at time $t = 0$, there is one node. At time $t = 1$, the number of nodes is distributed according to the random variable $X$, and in general, at time $t$, each of the nodes at time $t - 1$ has a number of children distributed according to (independent) copies of $X$. Let $Z_t$ denote the random variable denoting the number of nodes that exist at time $t$. We will prove the following theorem:

**Theorem 1.** Provided $\Pr[X = 1] < 1$ and $\Pr[X = 0] > 0$, then:

- If $E[X] \leq 1$ then $\lim_{t \to \infty} \Pr[Z_t = 0] = 1$.
- If $E[X] > 1$ then $\lim_{t \to \infty} \Pr[Z_t = 0] = p$ for $p \in (0, 1)$ with $p$ being the unique solution in $(0, 1)$ to the equation $p = \sum_{i \geq 0} \Pr[X = i] p^i$.

First, let us understand the relationship between the $Z_i$’s:

(a) Show that $Z_t$ is distributed according to the sum of $Z_1$ independent copies of $Z_{t-1}$.

**Solution:** The starting node’s children after $t$ time steps, are also children of the nodes at time 1 after $t - 1$ time steps. There are $Z_1$ nodes at time 1 and each one independently has $Z_{t-1}$ children after $t - 1$ time steps. So $Z_t$, the number of children the starting node has after $t$ time steps has the same distribution as $Z_1$ independent copies of $Z_{t-1}$. Another way to do this is to show that $Z_t$ and $\sum_{i=1}^{Z_1} Z_{t-1}^{-1}$ (where the $Z_{t-1}$ are independent) have the same moment generating functions.

Second, define $p_t = \Pr[Z_t = 0]$ to be the probability of extinction by time $t$:

(b) Prove that $p_t = \sum_{i \geq 0} \Pr[X = i] \Pr[Z_{t-1} = 0]^i$.

**Solution:** Using (a),

\[
\Pr[Z_t = 0] = \Pr[\sum_{i=1}^{Z_1} Z_{t-1} = 0]
\]
\[
= \sum_{i \geq 0} \Pr[Z_1 = i] \Pr[\sum_{j=1}^{Z_1} Z_{t-1} = 0 \mid Z_1 = i]
\]
\[
= \sum_{i \geq 0} \Pr[Z_1 = i] \Pr[\bigcap_{j=1}^{i} Z_{t-1}^{(j)} = 0 \mid Z_1 = i]
\]
\[
= \sum_{i \geq 0} \Pr[Z_1 = i] \Pr[Z_{t-1} = 0]^i.
\]
In the last step, we used the fact that the $Z_{t-1}^{(j)}$ are independent of $Z_1$ and of each other.

Since $p_1 \leq p_2 \leq \ldots$ is monotonically increasing and bounded by 1, by the Monotone Convergence Theorem, a limit $p = \lim_{t \to \infty} p_t$ exists. Define function $f(x) = \sum_{i \geq 0} \Pr[X = i]x^i$. By part (b) we know that $f(p_t) = p_{t+1}$, and combining with the definition of $p$, we conclude that $p = f(p)$. Let us explore some properties of $f$:

(c) Prove that $f(1) = 1$, $f'(1) = \mathbb{E}[X]$, and $f(x)$ is convex on the interval $(0, 1)$.

Solution: Evaluating $f(x)$ at 1 gives $\sum_{i \geq 1} \Pr[X = i] = 1$. Evaluating $f'(x)$ at $x = 1$ gives $\sum_{i \geq 1} \Pr[X = i]i = \mathbb{E}[X]$. The first derivative of $f(x)$ is $\sum_{i \geq 1} \Pr[X = i]ix^{i-1}$ which is non-decreasing on $[0, 1]$. So $f$ is convex on the interval $(0, 1)$.

Actually when $\mathbb{E}[X] \geq 1$, we can show that $f(x)$ is strictly convex. See that $\mathbb{E}[X] - 1 = \sum_{i \geq 0} i \Pr[X = i] - \sum_{i \geq 0} \Pr[X = i]$ which simplifies to $\sum_{i \geq 2} (i - 1) \Pr[X = i] - \Pr[X = 0]$. So $\mathbb{E}[X] \geq 1$ implies that $\sum_{i \geq 2} (i - 1) \Pr[X = i] \geq \Pr[X = 0] > 0$. So $\Pr[X = i] > 0$ for some $i \geq 2$. This means at least one of the terms in $f'(x)$ is positive and strictly increases with $x$. Since $f'(x)$ is strictly increasing, $f(x)$ is strictly convex.

Finally, to complete our proof:

(d) Show that if $\mathbb{E}[X] > 1$, $f(x) = x$ will have a unique solution in $(0, 1)$, and if $\mathbb{E}[X] \leq 1$, then there is no solution to $f(x) = x$ for $x \in (0, 1)$.

Solution: It’s helpful to look at the graph of $g(x) = f(x) - x$. We want to show that if $\mathbb{E}[X] > 1$ then $g(x)$ has a unique root in $(0, 1)$, and if $\mathbb{E}[X] \leq 1$ then $g(x)$ has no root in $(0, 1)$. We consider three cases-

i. $\mathbb{E}[X] > 1$ or $g'(1) > 0$: A positive slope at $x = 1$ means that for some small $\epsilon$, $g(1 - \epsilon) < g(1) = 0$. We also know that $g(0) = f(0) > 0$. Since $g(1 - \epsilon) < 0 < g(0)$, by the Intermediate Value Theorem, there is some $c$ in $(0, 1 - \epsilon)$ such that $g(c) = 0$. Since $g(x)$ is strictly convex on $[0, 1]$ it can’t have more than two roots in this interval. Since it has one root at $x = 1$, $c$ must be the unique root in $(0, 1)$.

ii. $\mathbb{E}[X] < 1$ or $g'(1) < 0$: Since $g(x)$ is convex, we know that $g(x)$ should lie above all of its tangents. For any $x \in (0, 1)$, $g(x) \geq g(1) + g'(1)(x - 1) = g'(1)(x - 1) > x - 1 > 0$. So $g(x) > 0$ for all $x \in (0, 1)$.

iii. $\mathbb{E}[X] = 1$ or $g'(1) = 0$: Since $g'(x)$ is strictly convex on $[0, 1]$, for all $x \in (0, 1)$, $g'(x) < g'(1) = 0$. This means $g(x)$ is strictly decreasing on $[0, 1]$ so $g(x) > g(1) = 0$ for all $x \in (0, 1)$.

For problem 2 and 3, we consider the sizes of the connected components of random graphs. Let $G_{n,p}$ denote the Erdos-Renyi random graph model, where each edge exists (independently) with probability $p = c/n$ for some constant $c$ that does not vary with $n$.

**Theorem 2.** Let $G$ be drawn from $G_{n,p}$, with $p = c/n$ for some constant $c$:

- If $c < 1$, with probability tending to 1 as $n \to \infty$, the largest connected component of $G$ has size $O(\log n)$. 

• If $c > 1$, with probability tending to 1 as $n \to \infty$, the largest connected component of $G$ has size $(1-p)n + o(n)$, where $p$ is the probability of extinction of the Galton-Watson branching process for the Poisson random variable with expectation $c$, and the second-largest component of $G$ has size $O(\log n)$.

2. In this problem we prove the $c < 1$ case of the above theorem.

(a) For a given vertex $v$, prove that

$$\Pr[v \text{ in connected component of size } \geq k] \leq \Pr[X \geq k-1],$$

where $X$ is distributed according to $Binomial[k \cdot n, c/n]$. [Hint: consider doing a breadth-first search of the neighborhood of $v$ in the graph.]

**Solution:** Consider doing a breadth-first search of $v$'s component. For the $i$th node explored in this search, the number of new nodes discovered is distributed according to $Binomial[n - \text{nodesDiscovered}, c/n]$. Note that for any $i$, the probability that the above binomial is at least $i$ is at most the probability that $Binomial[n, c/n]$ is at least $i$. The probability that $v$ is in a component of size at least $k$ is at most the probability that after $k-1$ nodes have been explored, there is at least one more node to explore, namely that the first $i-1$ explorations turned up at least $k-1$ nodes (i.e. at least $k$ nodes including $v$) which is at most

$$\Pr[\sum_{i=1}^{k-1} Bin[n, c/n] \geq k-1] = \Pr[Bin[n(k-1), c/n] \geq k-1] < \Pr[Bin[kn, c/n] \geq k-1].$$

(b) Assuming the above, using a union bound over Chernoff bounds, prove that

$$\Pr[\text{there is a connected component of size } \geq 10 \log n (1 - c)^2] \leq \frac{1}{n}.$$

This completes the proof.

**Solution:** To simplify things, we use the stronger statement that we proved above, namely that

$$\Pr[v \text{ in connected component of size } \geq k] \leq \Pr[X \geq k-1],$$

where $X$ is distributed according to $Binomial[(k-1) \cdot n, c/n]$. Letting $k = \frac{10 \log n}{(1-c)^2}$, for $k \geq 2$, sufficiently large $n$ we have that $k-1 \geq \frac{9 \log n}{(1-c)^2}$, and $E[X] = c(k-1)$. Hence we have that for $(1-c)/c < 1$,

$$\Pr[X \geq k-1] = \Pr[X \geq (1 + (1-c)/c)E[X]] \leq e^{-E[X] \frac{(1-c)^2}{3c^2}} \leq n^{-9/3} = n^{-3},$$

and hence the probability that any of the $n$ connected components have size exceeding $k-1$ is less than $1/n^2$. Note that in the above, used the Chernoff bound that for $\delta \in (0, 1)$, for $Y$ a sum of indicator random variables, $\Pr[X \geq (1 + \delta)E[X]] \leq e^{-E[X]\delta^2/3}$. Additionally, since the probability that the graph has a connected component of size at least $k$ is monotonically decreasing as $c$ decreases, it suffices to consider only the $c$ for which $(1-c)/c < 1$.}
3. In this problem, we prove the $c > 1$ case of the above theorem.

(a) Given a random node $v$ in the graph, prove that for any $\frac{100c \log n}{(c-1)^2} \leq k \leq n^{3/4}$, the probability that the connected component of $v$ has size $k$ is no more than $n^{-10}$. [Hint: given that the connected component of $v$ has size at least $k$, show that with high probability it will have size at least $k + 1$. Be mindful of the way you condition events.]

**Solution:** Consider exploring the connected component of $v$ as follows: at the 1st time step, we explore the neighborhood of $v$ (i.e. we discover the children of $v$, if it has any), and at each successive time step, we chose one vertex that has been discovered at a previous step of the algorithm (but whose neighborhood has not been previously chosen for exploration) and explore its neighborhood. If no such choice of vertex is possible (i.e. we have explored all the nodes that have been discovered), then we have fully explored $v$’s connected component. Hence, if $v$’s connected component has size at least $k$, then the above process must continue for at least $k$ steps. Note that at each step in the above process, the number of new (previously undiscovered) nodes is distributed according to $Bin[n - \text{discoveredNodes}, c/n]$, which dominates the distribution $Bin[n - n^{3/4}, c/n]$, as long as the number of discovered nodes is at most $n^{3/4}$. Hence, given that the process has continued for $k$ steps, but has not yet discovered more than $n^{3/4}$ nodes, the probability that we will not be able to continue it for a $k + 1$ step is at most $Pr[Bin[k(n - n^{3/4}), c/n] \leq k - 1] \leq Pr[Bin[k(n - n^{3/4}), c/n] \leq k]$. Now, the trick is to argue that with high probability, after exploring $n^{3/4}$ steps, by the previous part, they belong to components of size $\geq n^{3/4}$.

(b) Prove that we do not expect any connected components to have size in the interval $[\frac{100c \log n}{(c-1)^2}, n^{3/4}]$.

**Solution:** Given a random node $v$ in the graph,

$$Pr[\text{connected component of } v \text{ has size } k] \geq \frac{k}{n} Pr[\exists \text{connected component of size } k].$$

Hence, via a union bound over the at most $n$ different values of $k$ in this interval, it follows that with high probability (at least $1 - n^{-8}$) there will be no connected component with size in the range $[\frac{100c \log n}{(c-1)^2}, n^{3/4}]$.

(c) Prove that with probability tending to 1 as $n \to \infty$, there is at most one connected component of size $\geq n^{3/4}$. [Hint: conditioned on the neighborhood of both $v$ and $u$ having size at least $n^{3/4}$, show that the probability that they are not connected is tiny, then union bound over the at most $n$ such neighborhoods.]

**Solution:** Here, the high-level idea is to “grow” out components from two nodes, $v, w$. If each of them last $>> O(\log n)$ steps, by the previous part, they belong to components of size $> n^{3/4}$. Now, the trick is to argue that with high probability, after exploring $n^{3/4}$ steps of the component exploration process described in the previous part, with very
high probability, there will be $O((c - 1)n^{3/4})$ discovered nodes that have not yet been explored. Given this, with high probability either these node sets intersect, or at least one of the $O(n^{1.5})$ possible edges between these two node sets exist. We conclude via a union bound over all pairs of “large component”.

By the same argument as in the previous part, after exploring $n^{3/4}$ nodes in a connected component that started at node $v$, the probability that there are fewer than $\frac{(c - 1)n^{3/4}}{2}$ nodes that have been discovered, but are not yet explored, is at most $\Pr[X \leq n^{3/4} + \frac{(c - 1)n^{3/4}}{2}]$, where $X$ is distributed according to $Bin[n^{3/4}(n - n^{3/4}), c/n]$. Chernoff bounds yield that this probability is $< o(1/poly(n)) < 1/n^{10}$ for sufficiently large $n$.

For two nodes, $u, v$ that each have connected components of size $\geq n^{3/4}$, let $A, B$ be the sets of discovered but unexplored nodes after $n^{3/4}$ nodes were explored. If $A \cap B \neq \emptyset$, then they are part of a single component. Otherwise, with probability at least $2/n^{10}$, $|A|, |B| \geq \frac{(c - 1)n^{3/4}}{2}$, in which case out of the $(c - 1)^2 n^{1.5}/4$ potential edges connecting the two sets of unexplored nodes, the probability that none exist is $(1 - c/n)^{(c - 1)^2 n^{1.5}/4} < e^{-\Theta(\sqrt{n})} << n^{-10}$ for sufficiently large $n$. Hence, via a union bound over the at most $n^2$ nodes whose connected components have size at least $n^{3/4}$, with high probability all such connected components are connected, namely there is a single large connected component.

(d) Using the theorem proved in problem 1, show that the expected size of the large component is as claimed at the beginning of Theorem 2.

**Solution:** By the previous two parts, provided a node’s connected component has size at least $\frac{100c^2}{(c - 1)^2} \log n$, with high probability it will be part of the large connected component. Thus it suffices to bound estimate the probability that a given node is part of a connected component of size at least $\frac{100c^2}{(c - 1)^2} \log n$. Consider two Galton-Watson branching processes, one corresponding to the random variable $Bin(n, c/n)$, and one corresponding to the random variable $Bin(n - n^{3/4}, c/n)$. The probability that a given node’s component has size at least $\frac{100c^2}{(c - 1)^2} \log n$ lies between the probabilities that the above two branching processes do not go extinct, and hence the expected size of the large component is $n$ times this probability. By definition $Bin(n, c/n)$ converges to the Poisson distribution of expectation $c$ in the limit as $n \to \infty$, and $\Pr[Bin(n, c/n) = i] - \Pr[Bin(n - n^{3/4}, c/n) = i] \to 0$ as $n \to 0$, as can be seen by either writing out the probability density functions, or arguing that the extra $n^{3/4}$ coin tosses in the former distribution will yield 0 heads with probability $(1 - \frac{c}{n})^{n^{3/4}} \to 1$ as $n \to \infty$.

(e) **BONUS:** Show that the size of the large component is within $o(n)$ of its expectation with probability tending to 1 as $n \to \infty$. [Hint: bound the variance of the number of nodes that are in “small” components of size at most $\frac{100c^2 \log n}{(c - 1)^2}$, then use Chebyshev’s inequality.]

Spend a few minutes thinking about the theorem you have just proved, and the intuition behind why, with very high probability, there are never any medium-sized components.