CS265/CME309, Problem Set 4

SUNet ID(s):
Name(s):

By turning in this assignment, I agree by the Stanford honor code and declare that all of the writing is the work of my partner and I (discussion in larger groups is permissible).

Due by 11:59 PM on **Tuesday**, October 22th.

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 models the number of descendants that an individual has. The process is defined in terms of a random variable X that takes non-negative integer values—each individual will have a number of "children" drawn according to independent copies of X. The process, in terms of X 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 describing the number of nodes that exist at time t, namely the number of nodes that are "born" at time t. We will prove the following theorem:

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

   - *If* $\mathbf{E}[X] \leq 1$ *then* $\lim_{t\to\infty} \Pr[Z_t = 0] = 1$.
   - *If* $\mathbf{E}[X] > 1$ *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* $p = \sum_{i\geq 0} \Pr[X = i]p^i$.

   (a) (4 points) First, let us understand the relationship between the Z_i's. Show that Z_t is distributed according to the sum of Z_1 independent copies of Z_{t-1}.

   (b) (4 points) Define $p_t = \Pr[Z_t = 0]$ to be the probability of extinction by time t. Prove that $p_t = \sum_{i\geq 0} \Pr[X = i] \Pr[Z_{t-1} = 0]^i$.

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

   (d) (4 points) We now complete our proof of Theorem 1. Show that if $\mathbf{E}[X] > 1$, $f(x) = x$ will have a unique solution in $(0, 1)$, and if $\mathbf{E}[X] \leq 1$, then there is no solution to $f(x) = x$ for $x \in (0, 1)$.

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* $(1 - p)n \pm o(n)$, *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) (4 points) For a given vertex v, prove that

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

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.]

(b) (4 points) Assuming the above, using a union bound over Chernoff bounds, prove that

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

This completes the proof.

3. In this problem, we prove the $c > 1$ case of the above theorem.

(a) (6 points) Given a random node v in the graph, prove that for any k satisfying $\frac{100c \log n}{(c-1)^2} \leq k \leq n^{3/4}$, the probability that the connected component of v has size k is no more than n^{-10}. [Hint: consider a sort of breadth-first search that starts with a set that contains only v, then "marks" v and adds all the neighbors of v to the set , and then iteratively continues by "marking" an unmarked node of the set and adding all its neighbors to the set. Suppose we have "marked" k nodes, what is the chance that there are no more "unmarked" nodes in our set? Based on this, prove that, with high probability, if the connected component of v has size at least $k \in [\frac{100c \log n}{(c-1)^2}, n^{3/4}]$, then it will in fact have size at least $k+1$. Be mindful of the way you condition events!!]

SOLUTION: Following the suggested hint, let X_i denote the number of nodes added to v's connected component as a result of considering the neighbors of the ith node that we mark. (If v's component has $< i$ nodes, then let $X_i = 0$.) Provided we have an ith node to mark, X_i is distributed as $Bin(n-1-\sum_{j=1}^{i-1} X_j, c/n)$, since at the time we mark the ith node, we have not considered any of the potential edges between this node, and any of the $n-1-\sum_{j=1}^{i-1} X_j$ nodes that have not yet been added to v's component. [The "-1" in this expression is to count node v itself.] Pr[v's component has size k|v's component has size $\geq k$] = Pr[v's component has size k$|X_1 \geq 1, X_1 + X_2 \geq 2, \ldots, \sum_{j=1}^{k-1} \geq k-1$]. If this actually holds, then for all $i = 1, \ldots, k$, $\sum_{j=1}^{i-1} X_j \leq k$, and hence for all such i, X_i is a binomial consisting of at least $n-1-k$ tosses of a coint. Continuing in our analysis: $\Pr[\sum_{j=1}^{k} X_j = k-1 | X_1 \geq 1, X_1 + X_2 \geq 2, \ldots, \sum_{j=1}^{k-1} X_j \geq k-1]$ $\leq \Pr[\sum_{j=1}^{k} X_j \leq k-1 || X_1 \geq 1, X_1 + X_2 \geq 2, \ldots, \sum_{j=1}^{k-1} X_j \geq k-1]$. This conditioning only decreases this probability, as we are conditioning on the event that sums of these X_i's are *at least* certain values, hence this probability is at most

$$\Pr[\sum_{j=1}^{k} X_j \leq k] \leq \Pr[Bin(k(n-1-k), c/n) \leq k].$$

For any $k \in [\frac{100c \log n}{(c-1)^2}, n^{3/4}]$, for sufficiently large n it holds that the expected value of this binomial is $k(n-1-k)c/n = kc + o(kc)$, and hence letting Y denote a random variable distributed as $Bin(k(n-1-k), c/n)$, the standard Chernoff bound gives:

$$\Pr[Y \leq k] \leq \Pr[Y \leq (1-(c-1+o(1)))\mathbf{E}[Y]] \leq \Pr[Y \leq (1-(c-1)/2)\mathbf{E}[Y]] \leq e^{-\frac{(c-1)^2 \mathbf{E}[Y]}{4 \cdot 2}}.$$

Since $\mathbf{E}[Y] \geq k \geq \frac{100 \log n}{(1-c)^2}$, this probability is less than $n^{-100/8} \leq n^{-10}$.

(b) (2 points) 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: Since there are n different possible nodes, v, and $< n$ different values of k in the range $[\frac{100c \log n}{(c-1)^2}, n^{3/4}]$, a union bound over these $< n^2$ possible combinations, together with our probability of $\leq n^{-10}$ from the previous part, yields that with probability at least $1 - n^{-8}$ there are no connected components with sizes in this range.

(c) (4 points) 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.]

The following is NOT a solution: assuming both v and w have neighborhoods of size $\geq n^{3/4}$, then the probability there are no edges between these is at most $(1 - c/n)^{n^{3/2}}$. To see why this doesnt work, note that this argument could be applied to ANY sets of size $n^{3/4}$, but there DO exist sets of size $n^{3/4}$ that are disconnected, since a constant fraction of the nodes will have zero neighbors, and hence two subsets of these don't have any edges between them (since these are all degree zero nodes)....

SOLUTION: If we proceed as in the solution to Part (a), by the time we have marked $k = n^{3/4}$ vertices in v's component and $k = n^{3/4}$ vertices in w's component, then by the reasoning in Part(a), with probability at least $1 - e^{-\Theta(n^{3/4})}$, both v and w will have at least $(c-1)n^{3/4}/2$ unmarked nodes. Now, given this, if v and w's components do not already intersect then the probability that no edge exists between their unmarked node sets is at most $(1 - c/n)^{\left((c-1)n^{3/4}/2\right)^2} \leq e^{-\Theta(\sqrt{n})}$, and hence we can (with tons of room to spare) union bound over the $\leq n$ connected components to argue that they must be connected.

(d) (4 points) 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 (sketch): There are a few different ways to prove this. One way is to first argue that if Z_t is the Galton-Watson process corresponding to $Poisson(c)$, for $c > 1$, then

$$\lim_{t \to \infty} \Pr[Z_t = 0] = \Pr[Z_{(100 \log n)/c} = 0] + o(1).$$

(Namely, if the process is going to go extinct, it has probably already done so by time $t = O(\log n)$. Now, we just need to compute the probability that a given node is part of the big connected component. (By linearity of expectation, the expected size is just n times the probability each node is in the big component.) To show that this probability is at most a $o(1)$ off from the probability that the Galton-Watson branching process does not go extinct, we just need to compare the branching process with the breadth-first-search exploration of a connected component, up to depth $100 \log n/(1-c)^2$. (Since by Part(a), we know that there is only a $o(1)$ probability that a connected component has size greater than this, without being part of the big component).

SOLUTION 2: A slightly slicker approach is as follows: given Part (e)—that with probability $1 - o(1)$ the size of the largest connected component is $\alpha n + o(n)$, for some constant α that we need to compute. By linearity of expectation, α is the probability that each node is in the big connected component. For node v, this probability is simply the probability that at least one of v's neighbors is in the big component. This probability is just

$$\Pr[Bin(\alpha n + o(1), c/n) \geq 1] + o(1) = 1 - (1 - c/n)^{\alpha n + o(1)} + o(1) \to 1 - e^{-c\alpha}$$

for large n, where the $o(1)$ is from the probability that the big component does NOT have size $\alpha n + o(1)$. Putting this together, we have that α is the unique solution in the interval $(0, 1)$ to

$$\alpha = 1 - e^{-c\alpha}.$$

[And you can check that this is the same as the non-extinction probability of the specified Galton-Watson branching process.]

(e) BONUS +2: 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 $\frac{100c \log n}{(c-1)^2}$, then use Chebyshev's inequality.]

SOLUTION SKETCH: We just need to show that the variance of the size of the big connected component is $o(n^2)$, and then will be done by Chebyshev's inequality. Letting X_i denote the 0/1 random variable corresponding to whether or not the ith node is in the large connected component, to bound the variance we just need to bound $\mathbf{E}[(\sum_i X_i)^2] - \mathbf{E}[\sum_i X_i]^2$. The crucial part of this sum is the cross terms, the $O(n^2)$ terms of the form $\mathbf{E}[X_i X_j]$. Bounding these is fairly easy: from parts (a) and (c) up to a very small additive term $O(n^{-10})$, these events are determined by whether or not the "marking" process reaches $O(\log n)$ marked nodes. If we first do the marking process for node i, whether or not we have gotten to a component of size $O(\log n)$, we will have discovered at most $O(\log n)$ nodes, with high probability, and knowledge about this set does not significantly impact the marking process for node j, since with high probability the discovered neighborhood of node j (after marking at most $O(\log n)$ vertices from j) will be disjoint from node i's discovered neighborhood (of size $O(\log n)$).

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.