CS265/CME309, Fall 2019. Instructor: Gregory Valiant

Problem Set 4

Electronic submission via Gradescope due **11:59pm Tuesday 10/22**. You are strongly encouraged to submit a homework with a partner—that is, submit one homework with both of your names.

[You may discuss these problems with classmates. Feel free to look at wikipedia, course notes, etc. for reference material, but do not try to specifically search online for solutions to the problems. ***Your submission must be the original work of you and your partner, and you must understand everything that is written on your submission.*** *We strongly suggest that you write solutions using LaTex—see the course website for a latex solution template.]*

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* $\boldsymbol{E}[X] \leq 1$ *then* $\lim_{t\to\infty} \Pr[Z_t = 0] = 1$.
   - *If* $\boldsymbol{E}[X] > 1$ *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!!]

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

   (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.]

   (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.

(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.]

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.