Graphs and Networks Lecture 4

Erdös-Rényi Random Graphs: The Giant Component

Daniel A. Spielman September 14, 2010

4.1 Introduction

Recall that all the "Real-World" graphs we examined in Lecture 2 had one component containing a large constant fraction of the vertices. The second-largest component was smaller by many orders of magnitude. This property is shared by Erdös-Rényi random graphs and by many other graph models. In this lecture, we will see (mostly) why Erdös-Rényi random graphs have this property. The large component is called the "Giant Component".

Consider drawing a random graph from the distribution $\mathcal{G}(n,p)$ with $p = c/(n-1)$ for a constant c that stays fixed as n grows. The expected degree of each vertex in such a graph is c as n. For $c < 1$, we will see that every component of such a graph is probably small, having at most $O(\log n)$ vertices. On the other hand, for $c > 1$ we will see that such a graph probably has a connected component containing a constant fraction of the vertices. Moreover, it is unlikely that any other component has more than $O(\log n)$ vertices.

This is an example of a *threshold* phenomenon. Generally speaking, a graph property is a *threshold* phenomenon if the probability that it happens jumps from 0 to 1 as the parameter c passes a threshold. There are many other graph properties that exhibit similar jumps. I will list some at the end of the lecture if there is time.

We will begin our study of the giant component by examining the simplest such phenomenon: the giant component in the Galton-Watson branching process. This does not involve a graph. Just an organism reproducing.

4.2 Concentration and Chernoff Bounds

We begin our discussion by introducing one of the workhorses of probabilistic combinatorics: the Chernoff bounds. These are the main reason you see terms like $O(\log n)$ popping up everywhere in probability. Basically, the Chernoff (and Hoeffding) bounds are quantitative versions of the central limit theorem. They say that sums of independent random variables are exponentially concentrated about their means. The forms of the statement depend upon the types of random variables involve. When studying Erdös-Rényi-random graphs, we will just be interested in Bernoulli random variables. The Chernoff bounds we will use[1] follow.

[1]Many forms of Chernoff bounds may be found. It is often convenient to prove one's own. The form we use here appears in [MU05]. Other useful forms and derivations may be found in [AS00, MR95, DP09].

Theorem 4.2.1. *Let $X_1, \ldots, X_n$ be independent Bernoulli (that is, 0/1 valued) random variables where $Pr[X_i = 1] = p_i$. Let $X = \sum X_i$ and let $\mu = \sum p_i$ be the expectation of X. Then, for all $0 < \delta < 1$*

$$Pr[X \leq (1-\delta)\mu] \leq \exp(-\mu\delta^2/2)$$

and

$$Pr[X \geq (1+\delta)\mu] \leq \exp(-\mu\delta^2/3).$$

For an example of how the Chernoff bounds are used, let's examine the degrees of vertices in graphs drawn from the distribution $\mathcal{G}(n,p)$. For now, consider $p = c\ln n/(n-1)$ for some $c > 6$. The expected degree of a vertex is $\mu = c\ln n$. We will now see that it is unlikely that the degree of a vertex is much more or much less than this. If we set $\delta = \sqrt{6/c} < 1$, we find that the probability that the degree of a vertex exceeds $(1+\delta)\mu$ is at most

$$\exp(-c(\ln n)\delta^2/3) = \exp(-c(\ln n)(6/c)/3) = \exp(-2\ln n) = n^{-2}.$$

So, the probability that there is *any* vertex in the graph with degree greater than $(c+\sqrt{6c})\ln n$ is at most n^{-1}. Applying the first form of the Chernoff bound, we can also show that the probability that there is *any* vertex in the graph with degree less than $(c-\sqrt{4c})\ln n$ is at most n^{-1}.

4.3 The Galton-Watson Process, binary case

Imagine a single-cell organism that reproduces by division. Also imagine that there is some fixed probability p that the organism will survive to reproduce. Under these assumptions, we will compute the probability that the descendants of this organism die out or survive forever.

To be more concrete, let's start with one organism and let's assume this it does survive to reproduce. If it does reproduce, it divides into two organisms, each of which has a probability p of surviving to divide themselves. So, the expected number of organisms in the first generation that survive to reproduce is $2p$. Similarly, the expected number of organisms in the second generation that survive to reproduce is $4p^2$. One way of seeing this is to identify 4 potential organisms in the second generation: the first and second child of each of the first and second children of the original. Each of these 4 potential organisms exists and survives only if their parent organism survives and they survive themselves. The probability of each of these events is p^2. We may similarly compute that the expected number of organisms in the kth generation is

$$2^k p^k = (2p)^k.$$

For $p < 1/2$ this number goes to zero, whereas for $p > 1/2$ it goes to infinity. This is clearly some type of threshold phenomenon.

We can use this expectation calculation to prove it is unlikely that the descendants of the organism will survive for a long time when $p < 1/2$. Let X^k be the random variable counting the number of descendants of the organism in the kth generation. The descendants are still around in the kth generation if and only if $X^k \geq 1$. However, Markov's inequality tells us that

$$\Pr\left[X^k \geq 1\right] \leq \mathbf{E}\left[X^k\right] \leq (2p)^k \xrightarrow[k\to\infty]{} 0$$

But, what about when $p > 1/2$? The expected number of descendants goes to infinity. But, what does that tell us about the chance that the number of descendants is in fact infinite? This is less obvious.

4.4 $p > 1/2$

Let $\theta_k(p)$ be the probability that an organism has at least k generations of descendants. This is the same as the probability that at least one of its descendants survives and has at least $k-1$ generations of descendants. Let A be the event that the descendants of the first child live to at least $k-1$ generations and B be the same event for the second child. The probability that the first child survives and has at least $k-1$ generations of descendants is $p\theta_{k-1}(p)$. As $\theta_k(p)$ is the probability of A or B, we may compute it using the formula

$$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \text{ and } B].$$

As A and B are independent in our mathematical abstraction, we have

$$\Pr[A \text{ and } B] = \Pr[A] \Pr[B].$$

This gives

$$\theta_k(p) = 2p\theta_{k-1}(p) - (p\theta_{k-1}(p))^2. \tag{4.1}$$

In particular,

$$\theta_0(p) = 1 \qquad \text{and} \qquad \theta_1(p) = 2p - p^2.$$

As k grows large, we expect $\theta_k(p)$ to approach a limit. If it does, it should be a number q that satisfies the equation

$$q = 2pq - (pq)^2.$$

This equation has one obvious solution: $q = 0$. For $p > 1/2$ we will see that the other solution dominates. To find the other solution, divide by q to get

$$\begin{aligned} 1 &= 2p - p^2 q \\ p^2 q &= 2p - 1 \\ q &\stackrel{\text{def}}{=} \frac{2p-1}{p^2}. \end{aligned}$$

Note that for $p > 1/2$ this is a constant strictly larger than 0. We will now show by induction that

$$\theta_k(p) \geq q$$

for all $k \geq 0$. To see this, we examine the function

$$f(x) = 2px - (px)^2,$$

as

$$\theta_k(p) = f(\theta_{k-1}(p)).$$

We will base our induction in the case $k = 0$, for which we have

$$1 = \theta_0(p) \geq q$$

(with equality only when $p = 1$). To perform the induction, we will show that $f(x) \geq q$ for $x \in [q, 1]$. We first compute the derivative of f with respect to x and find

$$f'(x) = 2p - 2p^2x = 2p(1 - px) > 0$$

for $x \in (0, 1]$. This means that f is an increasing function on $(0, 1]$. As $f(q) = q$, we may conclude that $f(x) \geq q$ for $x \geq q$. Thus,

$$\theta_k(p) \geq q$$

for all $k \geq 0$.

With a little more work one can show that

$$\liminf_{k\to\infty} \theta_k(p) = q.$$

One consequence of this is that with probability at least q the descendants of the organism never die out. That is, they exist for an infinite number of generations.

4.5 The Number of Descendants

We will now do a more detailed analysis in which we examine the number of descendants of an organism. We will perform this analysis in a more general setting. Each organism will divide into k others. We set the probability that an organism survives to reproduce to $p = c/k$. In the sub-critical regime ($c < 1$) we will see that it is very unlikely that the organism has too many descendants. In the super-critical regime ($c > 1$) we will see that once the number of descendants of an organism becomes sufficiently large it is likely to be infinite.

Remark I should have said during lecture that, just as in the case with $k = 2$, we can prove that for $c > 1$ there is a constant probability of an organism spawning an infinite number of generations. Maybe I'll put this on the first problem set.

We will find it useful to assign a number of every cell that survives to reproduce. We number the first cell 1. We must use consecutive numbers in a consistent manner, and must assign every cell a lower number than each of its descendants. For example, if there are j cells in the first generation that survive to reproduce, we could assign them numbers 2 through $j + 1$. We could then assign numbers to the cells in the second generation, and so on.

For each j such that cell j survives to reproduce, we introduce Bernoulli random variables $X_{j,1}, \ldots, X_{j,k}$ where $X_{j,i} = 1$ if the ith child of cell j survives to reproduce. Cell u is the last surviving member of the population precisely when

$$1 + \sum_{j=1}^{u} \sum_{i=1}^{k} X_{j,i} = u$$

and for all $v < u$

$$1 + \sum_{j=1}^{v} \sum_{i=1}^{k} X_{j,i} > v.$$

We will now use the Chernoff bounds to bound how unlikely this is in the sub-critical case. Define

$$X^{(u)} = \sum_{j=1}^{u} \sum_{i=1}^{k} X_{j,i}.$$

The expectation of $X^{(u)}$ is

$$\mu = ukp = uk\frac{c}{k} = uc.$$

For $c < 1$ this becomes significantly less than u and the Chernoff bounds will imply that $X^{(u)}$ is very unlikely to be more than u. Before we carry out the details of that argument, let me put one issue to rest. You might worry that $X^{(u)}$ is only defined if cell u actually survives to reproduce. You may then worry about what it means to take this sum if $X^{(u-1)} < u-1$. To make these notions precise, consider sampling all the variables $X_{j,i}$ for $1 \leq j \leq u$ and $1 \leq j \leq k$ without thinking about the Galton-Watson process. If it turns out that organism j does survive to reproduce, then and only then look at the variables $X_{j,i}$ to figure out which of its children survive to reproduce. If organism j never exists, then just throw away the unused variables[2].

Let Z be the number of descendants of the first organism, plus 1 for the first organism (or view Z as a descendant of itself). We can now say that

$$\Pr\left[Z > u\right] \leq \Pr\left[X^{(u)} \geq u\right] \leq \exp\left(-\frac{1}{3}\delta^2\mu\right),$$

where we set δ so that

$$\begin{aligned}(1+\delta)\mu &= u \\ (1+\delta)uc &= u \\ (1+\delta) &= \frac{1}{c} \\ \delta &= \frac{1}{c} - 1,\end{aligned}$$

which is greater than 0 in the sub-critical case. We conclude that

$$\Pr\left[Z > u\right] \leq \exp\left(-\frac{1}{3}\frac{(1-c)^2}{c}u\right).$$

So, the probability that there are more than u descendants decreases exponentially with u. This is why all the components of $\mathcal{G}(n,p)$ in the sub-critical case probably have logarithmic size.

Remark This bound is actually much stronger than the bound we proved on the number of generations spawned by the first organism, as the number of descendants can be exponential in the number of generations.

[2]This may worry you, but I assure you that you can make it formal.

In the super-critical case we will perform a similar analysis. We will show that it is very unlikely that $Z = u$ for any sufficiently larger u. By summing over all large u we will conclude that if Z is not small then it is probably infinite. Here the expectation of $X^{(u)}$ is also cu, but $c > 1$. We have

$$\Pr[Z = u] \leq \Pr\left[X^{(u)} \leq u\right] \leq \exp\left(-\frac{1}{2}\delta^2\mu\right),$$

where we set δ so that

$$\begin{aligned}(1-\delta)cu &= u \\ (1-\delta) &= \frac{1}{c} \\ \delta &= 1 - \frac{1}{c},\end{aligned}$$

which is greater than zero in the super-critical case. We thereby conclude that

$$\Pr[Z = u] \leq \exp\left(-\frac{1}{2}\frac{(c-1)^2}{c}u\right) = \exp\left(-\frac{1}{2}\frac{(c-1)^2}{c}\right)^u.$$

Define

$$\gamma = \exp\left(\frac{1}{2}\frac{(c-1)^2}{c}\right).$$

By summing an infinite series we can now bound the probability that Z is a large but finite number. We have

$$\Pr[u \leq Z < \infty] = \sum_{w=u}^{\infty} \Pr[Z = w] \leq \sum_{w=u}^{\infty} \gamma^{-w} = \frac{\gamma^{-u}}{1-\gamma^{-1}}.$$

So, this probability also decreases exponentially with u. This is part of why the second-largest component of $\mathcal{G}(n,p)$ in the super-critical case probably has logarithmic size. This is also particularly interesting as we know there is a constant probability that there are an infinite number of generations. In particular, this implies that

$$\sum_{w=1}^{\infty} \Pr[Z = w] < 1.$$

4.6 The Giant Component in a Graph

There is no way that we are going to get this far in this lecture. So, I will save it for next lecture. For now, I just say that it is a mathematically simple but conceptually interesting consequence of our analysis of the Galton-Watson process.

References

[AS00] Noga Alon and Joel Spencer. *The Probabilistic Method.* John Wiley & Sons, 2000.

[DP09] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, New York, NY, USA, 2009.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms.* Cambridge University Press, New York, NY, USA, 1995.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, New York, NY, USA, 2005.