Random Graphs — Week 5

Yuval Filmus

November 30, 2019

Today we will discuss *connectivity*. A graph is connected as long as it doesn't contain a *cut*. A set S forms a cut if there are no edges between S and its complement. We can assume that $|S| \leq n/2$. What cuts are the most hard to get rid of?

$$\sum_{|S|=k} \Pr[S \text{ is a cut}] = \binom{n}{k}(1-p)^{k(n-k)}.$$

When k is constant and p is small, we can approximate

$$\sum_{|S|=k} \Pr[S \text{ is a cut}] \sim \frac{n^k}{k!}e^{-pkn} = \frac{(ne^{-pn})^k}{k!}.$$

(In the sequel, it will be useful to notice that $1-p \leq e^{-p}$ always.)

This calculation suggests that we want ne^{-pn} to be small, and in that case, the most likely cut is that of size 1, or in other words, an isolated vertex.

We can make a guess at the threshold by solving the equation $ne^{-pn} = 1$, or $n = e^{pn}$. Taking logarithms, we find out that $p = \frac{\log n}{n}$. In slightly more detail, if we want $ne^{-pn} = o(1)$, then we need $p = \frac{\log n + c}{n}$, where $c = \omega(1)$. Indeed, for such a value of p we would have

$$\sum_{v \in [n]} \Pr[v \text{ is isolated}] \leq ne^{-pn} = e^{-c} \to 0.$$

Similarly, for any constant k we get that with high probability there are no cuts of size k. With some work, one can show that in fact

$$\sum_{1 \leq |S| \leq n/2} \Pr[S \text{ is a cut}] = o(1),$$

and so if $p = \frac{\log n + \omega(1)}{n}$, then with high probability the graph is connected. Later on we'll see a better way of doing this calculation, so we don't bother working it out here.

First, though, let us see what happens if $p = \frac{\log n - \omega(1)}{n}$. Is it true that the graph is disconnected with high probability? Given that isolated vertices are the hardest to get rid of, this suggests trying to show that for such p, with high probability there are isolated vertices. Indeed, denoting by X the number of isolated vertices, we can easily compute

$$\begin{aligned}
\mathbb{E}[X] &= n(1-p)^{n-1}, \\
\mathbb{E}[X^2] &= n(1-p)^{n-1} + n(n-1)(1-p)^{2(n-2)+1}, \\
\mathbb{V}[X] &= n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3} - n^2(1-p)^{2(n-1)} = \\
&\quad n[(1-p)^{n-1} - (1-p)^{2n-3}] + n^2(1-p)^{2n-3}p, \\
\frac{\mathbb{V}[X]}{\mathbb{E}[X]^2} &\le \frac{1}{n(1-p)^{n-1}} + \frac{p}{1-p}.
\end{aligned}$$

If $n(1-p)^{n-1} \to \infty$ and $p \to 0$, then $\mathbb{V}[X]/\mathbb{E}[X]^2 = o(1)$, and so with high probability there is an isolated vertex. It's not hard to check that this is indeed the case for $p = \frac{\log n - \omega(1)}{n}$.

Now let us go back to estimating the probability of a cut when $p = \frac{\log n + c}{n}$. If the graph is not connected, then there must be a connected component of size $k \le n/2$. Picking a spanning tree for the component, we see that there must be a tree on some $k \le n/2$ vertices which is disconnected from the rest of the graph. For a given set of k vertices, there are k^{k-2} potential trees (this is Cayley's formula), and so a union bound shows that the probability that the graph is disconnected is at most

$$\sum_{k=1}^{n/2} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}.$$

When k is constant and p is not too close to 1, the kth summand is

$$\Theta(n^k p^{k-1} (1-p)^{kn}) = \Theta((npe^{-pn})^k / p) \le \Theta(n) \cdot \left(\frac{\log n + c}{n} e^{-c} \right)^k.$$

When $k \ge 2$, this bound is $o(1)$ even when c is *constant*. This suggests the tantalizing possibility, that when c is constant, the only reason for $G(n,p)$ to be disconnected is if it contains an isolated vertex! To show this, we have to bound the sum

$$\sum_{k=2}^{n/2} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}.$$

Bounding $\binom{n}{k} \le n^k/k!$, $k^{k-2} \le k^k$, $p^{-1} = O(n)$, and $1 - p \le e^{-p}$, the kth summand is at most

$$O(n) \cdot \frac{k^k}{k!} n^k p^k e^{-pkn} e^{pk^2}.$$

Stirling's approximation states that $k! \sim \sqrt{2\pi k}(k/e)^k$, implying that $k^k/k! = O(e^k)$. Therefore the kth summand is at most

$$O(n) \cdot \left(e \cdot n \cdot \frac{\log n + c}{n} \cdot \frac{e^{-c}}{n} e^{(\log n + c)(k/n)} \right)^k = O(n) \cdot \left(e^{1-c} \frac{\log n + c}{n} e^{(\log n + c)(k/n)} \right)^k.$$

When $k \ge 2$ is constant, the factor $e^{(\log n + c)(k/n)}$ is very small, and so the kth summand is $o(1)$. For all $k \le n/2$, the kth summand is at most

$$O(n) \cdot \left(e^{1-c} \frac{\log n + c}{n} e^{(\log n + c)/2} \right)^k = O(n) \cdot \left(e^{1-c/2} \frac{\log n + c}{\sqrt{n}} \right)^k,$$

which is $o(1/n)$ for $k \geq 5$. Since there are $O(n)$ values of k between 5 and $n/2$, this completes the proof that

$$\sum_{k=2}^{n/2} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} = o(1)$$

for $p = \frac{\log n + c}{n}$, where c is constant.

It follows that for such p, the probability that $G(n,p)$ is connected is the same (up to $o(1)$ error) as the probability that $G(n,p)$ contains no isolated vertices. This prompts us to estimate the latter probability. Above we have estimated the first two moments of the number X of isolated vertices. More generally, for every fixed ℓ,

$$\mathbb{E}\left[\binom{X}{\ell}\right] = \sum_{|S|=\ell} \Pr[\text{All vertices in } S \text{ are isolated}] = \binom{n}{\ell}(1-p)^{\ell(n-\ell)+\binom{\ell}{2}} \sim \frac{n^\ell}{\ell!} e^{-pn\ell} = \frac{e^{-c\ell}}{\ell!}.$$

If we denote $\mathbb{E}[X] = e^{-c}$ by λ, then this shows that $\mathbb{E}[\binom{X}{\ell}] \sim \lambda^\ell/\ell!$. The distribution of X is thus roughly Poisson, and we conclude that the probability that there are no isolated vertices tends to $e^{-e^{-c}}$.

In total, we have proved the following celebrated theorem:

Theorem 1. *If $p = \frac{\log n + c}{n}$, where c is constant, then the probability that $G(n,p)$ is connected tends to $e^{-e^{-c}}$.*

We can actually say a bit more: with high probability, the graph becomes connected once the last isolated vertex disappears. To make this statement precise, consider the coupling from Week 1: each edge has a weight distributed uniformly in $[0,1]$, and G_t consists of all edges whose weight is at most t. The precise statement is that with high probability, the minimal t such that G_t is connected is the same as the minimal t at which there are no isolated vertices.

Here is a rough sketch of the argument. Let $t_- = \frac{\log n - \log\log n}{n}$ and $t_+ = \frac{\log n + \log\log n}{n}$. The first step is to show that with high probability, the graph gets connected during the interval $[t_-, t_+]$. Furthermore, with high probability, at time t_- there aren't so many isolated vertices. The probability that an edge gets added in $[t_-, t_+]$ is $t_+ - t_- = 2\log\log n/n$, and so with high probability no edges added in this interval touch two isolated vertices (because there aren't so many of them). The result follows.