Random Graphs — Week 11

Yuval Filmus

January 8, 2020

1 Kruskal's algorithm

Suppose that the weight $w(e)$ of each edge $e \in K_n$ is chosen uniformly in $[0,1]$. What is the expected weight of a minimum spanning tree?

Recall Kruskal's algorithm for constructing a minimum spanning tree: go over the edges in increasing order of weight, and add each edge which connects two different connected components.

Let G_p denote the graph consisting of all edges whose weight is less than p, so that $G_p \sim G(n,p)$. If this graph consists of $\kappa(G_p)$ connected components, then Kruskal's algorithm will add exactly $\kappa(G_p)-1$ edges outside of G_p, that is, exactly $\kappa(G_p)-1$ edges whose weight is at least p. In other words, if we denote by T the tree constructed by Kruskal's algorithm, then

$$\#\{e \in T : w(e) \geq p\} = \kappa(G_p) - 1.$$

Now suppose that we choose p uniformly in $[0,1]$. An edge $e \in T$ satisfies $w(e) \geq p$ with probability exactly $w(e)$. Hence

$$\sum_{e \in T} w(e) = \int_0^1 \#\{e \in T : w(e) \geq p\}\,dp = \int_0^1 [\kappa(G_p) - 1]\,dp.$$

In particular, denoting by $w(T)$ the total weight of the tree,

$$\mathbb{E}[w(T)] = \int_0^1 \mathbb{E}[\kappa(G_p) - 1]\,dp.$$

Recall that the threshold of connectivity is around $\frac{\log n}{n}$. More accurately, for *constant* c, the probability that G_p is connected when $p = \frac{\log n + c}{n}$ tends to $e^{-e^{-c}}$. In particular, if $p \geq \frac{3\log n}{n}$, then the probability that G_p is connected is $1 - o(1/n)$ (this requires some calculation), and so

$$\mathbb{E}[\kappa(G_p) - 1] \leq n \Pr[G_p \text{ not connected}] = o(1).$$

Hence the contribution of $p \geq \frac{3\log n}{n}$ to $\mathbb{E}[w(T)]$ is negligible.

In their classical paper on the evolution of random graphs, Erdős and Rényi showed that $G(n,p)$ undergoes a phase transition around $p = 1/n$. When $p \ll 1/n$, most

components are small trees, and when $p = c/n$ for $c > 1$, there is in addition a so-called *giant component*, occupying a linear number of vertices. This suggests that when trying to estimate $\kappa(G_p)$, we concentrate on trees.

Let T_k be the number of tree components of size k. Using Cayley's formula, we get

$$\mathop{\mathbb{E}}_{G_p}[T_k] = \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)+\binom{k}{2}-(k-1)}.$$

We are only interested in $p = O(\log n/n)$. When k is small enough, we will have

$$\mathop{\mathbb{E}}_{G_p}[T_k] \sim \frac{n^k}{k!} k^{k-2} p^{k-1} (1-p)^{kn}.$$

For this to hold, we need $n^{\underline{k}} \sim n^k$ (which holds as long as $k = o(\sqrt{n})$), and $(1-p)^{\Theta(k^2)} \sim 1$ (which holds as long as $k = o(\sqrt{1/p}) = o(\sqrt{n/\log n})$.

Every other component contains a spanning tree as well as an additional edge. The number of such arrangements is at most $k^{k-2}\binom{k}{2} \leq \frac{1}{2}k^k$. Hence using a very rough union bound, we can bound the expected number of nontree components C_k of size k by

$$\mathop{\mathbb{E}}_{G_p}[C_k] \leq \frac{1}{2}\binom{n}{k} k^k p^k (1-p)^{k(n-k)} \leq \frac{1}{2}(1-p)^{-k^2} \cdot (npe^{1-np})^k,$$

using $\binom{n}{k} \leq (en/k)^k$ and $1 - p \leq e^{-p}$. When $k = o(\sqrt{n/\log n})$, we have $(1-p)^{-k^2} \sim 1$ as before. The function ce^{1-c} attains its maximum value 1 at the point $c = 1$, and this shows that when $k = o(\sqrt{n/\log n})$, it is always the case that

$$\mathop{\mathbb{E}}_{G_p}[C_k] \leq \frac{1}{2} + o(1).$$

Finally, we have to deal with large k, but this is easy: there are at most n/K components of size larger than K. Therefore

$$\mathbb{E}[\kappa(G_p) - 1] = (1 + o(1)) \sum_{k=1}^{K} \frac{n^k}{k!} k^{k-2} p^{k-1} (1-p)^{kn} + O(K + n/K).$$

Choosing (say) $K = n^{1/3}$, we get $K + n/K = o(\frac{n}{\log n})$, and so

$$\mathbb{E}[w(T)] = (1 + o(1)) \int_0^{\frac{3\log n}{n}} \sum_{k=1}^{K} \frac{n^k}{k!} k^{k-2} p^{k-1} (1-p)^{kn}\, dp + o(1).$$

We would like to extend the range of the integral all the way to 1. Indeed, if $p \geq \frac{3\log n}{n}$ then $(1-p)^n \leq e^{-pn} \leq 1/n^3$, and so

$$\int_{\frac{3\log n}{n}}^{1} \frac{n^k}{k!} k^{k-2} p^{k-1} (1-p)^{kn}\, dp \leq \frac{n^k}{k!} k^{k-2} \cdot \frac{1}{n^{3k}} = \frac{k^{k-2}}{k!} \cdot \frac{1}{n^{2k}} \leq \frac{1}{n^k}.$$

This is neglibigle even when summed over $k = 1, \ldots, K$, and so we get

$$\mathbb{E}[w(T)] = (1 + o(1)) \sum_{k=1}^{K} \frac{n^k}{k!} k^{k-2} \int_0^1 p^{k-1} (1-p)^{kn}\, dp + o(1).$$

It is a classical calculation that $\int_0^1 p^a(1-p)^b\,dp = \frac{a!b!}{(a+b+1)!}$, and so

$$\mathbb{E}[w(T)] = (1+o(1))\sum_{k=1}^{K} \frac{n^k}{k!} k^{k-2} \frac{(k-1)!(kn)!}{(k(n+1))!} + o(1).$$

For the relevant values of k, we have $(k(n+1))!/(kn)! \sim (kn)^k$, and so

$$\mathbb{E}[w(T)] = (1+o(1))\sum_{k=1}^{K} \frac{n^k}{k!} k^{k-2} \cdot \frac{(k-1)!}{(kn)^k} + o(1) = (1+o(1))\sum_{k=1}^{K} \frac{1}{k^3} + o(1).$$

Since the series $\sum_{k=1}^{\infty} 1/k^3$ converges and $K \to \infty$, we can finally conclude that

$$\mathbb{E}[w(T)] = \sum_{k=1}^{\infty} \frac{1}{k^3} = \zeta(3).$$

Apéry showed that the constant $\zeta(3)$ is irrational.

2 Dijkstra's algorithm

Suppose that the weight $w(e)$ of each edge $e \in K_n$ is chosen according to the unit-mean exponential distribution. What is the expected minimum distance between two fixed vertices? Given a vertex x, what is the expected minimum distance to the farthest away vertex y?

We are considering here the exponential distribution since it is easier to deal with. However, the same result holds for the uniform distribution as well, and similarly, the result on Kruskal's algorithm also holds for the exponential distribution. The reason is that the only relevant weights are very close to zero, and the two distributions have very similar density functions around zero.

We can generate an exponential random variable as follows: at each infinitesimal interval of length ϵ, we stop the process with probability ϵ, and otherwise we continue. If instead of stopping we just 'mark' the spot, then after N intervals we have roughly N marks, showing that the expected time until the first mark is 1. This point of view will be useful in the sequel.

Recall Dijkstra's algorithm for constructing a shortest path tree rooted at a given vertex 1. At each point in time, we consider all possible ways of adding an edge to the tree. We add the edge which results in the shortest distance from the root.

Let us see what happens when we run Dijkstra's algorithm starting at the vertex 1. The relevant edge weights are $w(1,2),\ldots,w(1,n)$, and we think of them as generated according to the process above. At each interval of length ϵ, we stop the process with probability $(n-1)\epsilon$ (using the estimate $\epsilon^2 = 0$). Equivalently, at each interval of length $\epsilon/(n-1)$, we stop the process with probability ϵ, speeding up the original process by a factor of $n-1$. Hence the first edge is added at time $1/(n-1)$ in expectation. The vertex it is pointing at is uniform over $2,\ldots,n$, but let's assume for simplicity that it is vertex 2.

We continue running the process for $w(1,3),\ldots,w(1,n)$, but now new weights come into play: $w(2,3),\ldots,w(2,n)$. Since there are $2(n-2)$ processes in play, the additional time we have to wait before the second edge is added is $1/[2(n-2)]$ in expectation.

Continuing in this fashion, we see that the expected total running time of the process, which is also the expected distance to the farthest vertex, is

$$\sum_{k=1}^{n-1} \frac{1}{k(n-k)} = \frac{1}{n} \sum_{k=1}^{n-1} \left[\frac{1}{k} + \frac{1}{n-k} \right] = \frac{2H_{n-1}}{n} \sim \frac{2\ln n}{n}.$$

We can also compute the variance of the distance to the farthest vertex. The variance of an exponential random variable of mean μ is μ^2. Since variance of independent random variables is additive, the variance of the distance to the farthest vertex is

$$\sum_{k=1}^{n-1} \frac{1}{(k(n-k))^2} = \frac{1}{n^2} \sum_{k=1}^{n-1} \left(\frac{1}{k} + \frac{1}{n-k} \right)^2 = \frac{2}{n^2} \sum_{k=1}^{n-1} \frac{1}{k^2} + \frac{2}{n^2} \sum_{k=1}^{n-1} \frac{1}{k(n-k)} = \frac{2}{n^2} \left(\frac{\pi^2}{6} - O\left(\frac{1}{n}\right) \right) + \frac{4H_{n-1}}{n^3} = \frac{\pi^2}{3n^2} - O\left(\frac{1}{n}\right).$$

Applying Chebyshev's inequality, we deduce that for any $f(n) = \omega(1)$, with high probability the distance to the farthest distance is in the interval $[\frac{2\ln n - f(n)}{n}, \frac{2\ln n + f(n)}{n}]$.

In the analysis above, we assumed that the vertices are discovered in the order $1,2,\ldots,n$, but in reality, after the first vertex, the other $n-1$ vertices are discovered in random order. In particular, the position of vertex 2 in this order is uniform, and so the expected distance from vertex 1 to vertex 2 is

$$\frac{1}{n-1} \sum_{\ell=1}^{n-1} \sum_{k=1}^{\ell} \frac{1}{k(n-k)} = \frac{1}{n(n-1)} \sum_{\ell=1}^{n-1} \sum_{k=1}^{\ell} \left[\frac{1}{k} + \frac{1}{n-k} \right] = \frac{1}{n(n-1)} \sum_{k=1}^{n-1} \left[\frac{n-k}{k} + \frac{n-k}{n-k} \right] = \frac{1}{n-1} \sum_{k=1}^{n-1} \frac{1}{k} = \frac{H_{n-1}}{n-1} \sim \frac{\ln n}{n}.$$

The variance is $O(\frac{1}{n^2})$ in this case as well, though the calculation is slightly more complicated, and left to the reader.

One can also show that the expected maximum distance between two vertices is $\sim \frac{3\ln n}{n}$, but the argument is more involved.