STAT 206: Random Graphs and Complex Networks	Spring 2003

Lecture 7: Critical behavior of the Erdős-Renyi model

Lecturer: David Aldous *Scribe: Andrej Bogdanov*

In this lecture we investigate the size of the giant component in the Erdős-Renyi random graph model with edge probability $p \approx 1/n$. In this model, the presence of each edge in the graph is decided by an independent coin flip with success probability p. We show that for p slightly bigger than $1/n$, the giant component has size of order $n^{2/3}$. There are several proofs of this result, and we will opt for an intuitive, back-of-envelope heuristic argument. This argument has the advantage of showing off some sophisticated concepts from probability like the Central Limit Theorem and Brownian motion.

At the heart of this argument is a random process that, starting from an arbitrary vertex v, exposes the neighbors of v, their neighbors, and so on, until the whole component of v is revealed. The sizes of the components are found by analyzing the dynamics of this process. For the Erdős-Renyi model, the relevant statistics can be computed exactly, giving a detailed picture of the behavior of component sizes near $p = 1/n$.

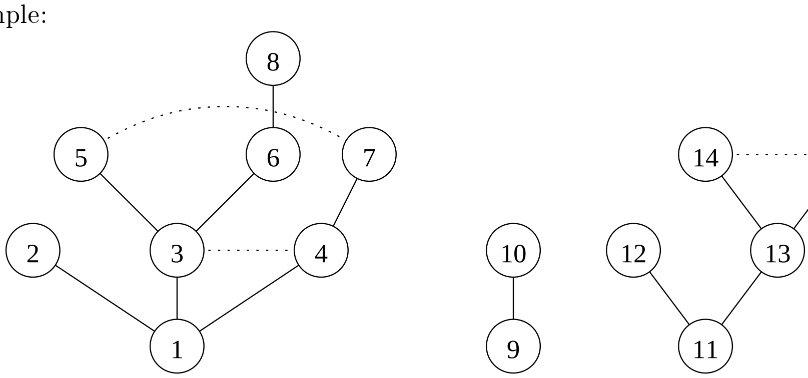
7.1 The breadth first spanning forest of a graph

Let G be a (non-random) graph on vertex set $[n]$. The *breadth first spanning forest* of G is the spanning forest generated by the breadth first search algorithm:

Until all vertices of G have been visited,

1. Pick a vertex v that has not been visited yet. Put v in a queue Q.
2. While Q is nonempty, pull a vertex v from the head of Q, draw edges to all its neighbors that have not been previously visited, and put these children at the tail of Q.

Here is an example:



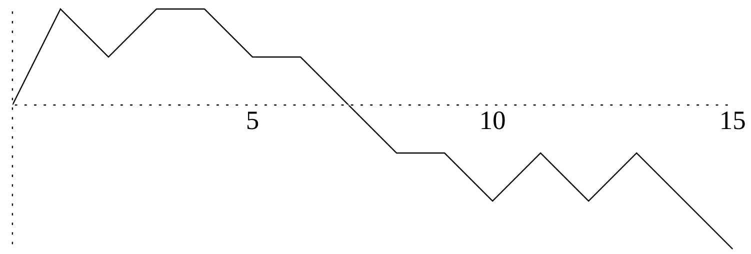
A nice property of this construction is that all the edges of G which fail to be included in the breadth first spanning forest connect descendants of the same generation. For the random graph $\mathcal{G}(n, p \approx 1/n)$ we expect to see few such cross edges.

We will need a few more notions related to the breadth first search algorithm. Suppose we interrupt the algorithm before it terminates. Say a vertex v has been *visited* if it was in the queue Q at some point before, or at the time of the interruption. Say v has been *processed* if v has been in Q at some point before the interruption, but is not there at the time of the interruption. For instance, suppose we have just finished putting vertex 4 from our example in Q, and we interrupt the algorithm. At this point, the algorithm has visited the vertices $1, 2, 3, 4, 5, 6$, but has processed only vertices $1, 2$ and 3.

From here on, we will assume that the vertices of G are labeled in their order of visitation in the breadth first spanning forest F, just like in the picture. Given this labeling, we define a (deterministic) walk $w(0), w(1), \ldots, w(n)$ by the formula

$$w(i) = w(i-1) + (\text{number of children of } i \text{ in } F) - 1,$$

with $w(0) = 0$. For example, for the above graph the walk will look like this:



Notice that the size of the first component is the smallest i such that $w(i) = -1$. More generally, the last visited vertex of the kth component is the smallest i such that $w(i) = -k$. This is true for any graph, and we will use it to study the sizes of components in the random graph $\mathcal{G}(n, p)$.

7.2 A little bit of probability

In the graph $\mathcal{G}(n, p)$, the walk $w(0), w(1), \ldots, w(n)$ is a random process. The key parameter of the process is the distribution of the increments $w(i) - w(i-1)$ conditioned on the past $w(i-1), \ldots, w(0)$. Let D_i denote the number of all vertices that have been visited but not processed (excluding i itself) when the algorithm reaches vertex i. At this point, the probability of an edge between i and any vertex that has not been visited is independent of the past. Since the set of candidate neighbors for i is exactly the set of non-visited vertices, the distribution of $w(i) - w(i-1)$ is binomial with $n - i - D_i$ samples and success probability p.

We are interested in the limiting behavior of the process $w(0), w(1), \ldots$ as $n \to \infty$. If, for the moment, we ignore the contribution of the D_i, the differences $w(i) - w(i-1)$ are independent random variables. If we keep i small compared to n, these are also almost identically distributed. Therefore, at least to a first approximation, we can model $w(i)$ as a sum of independent, identically distributed random variables. We can now appeal to a celebrated statement of probability theory that describes the limiting behavior of such sums:

Theorem 7.1 (The Central Limit Theorem) *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean 0 and variance 1. Fix $t > 0$ and let $\overline{X}_t^{(m)} = (X_1 + \ldots + X_{\lfloor mt \rfloor})/\sqrt{m}$. As $m \to \infty$, $\overline{X}_t^{(m)}$ converges in distribution to a normal mean 0, variance t random variable $\overline{X}_t^{(\infty)}$.*

If we now let the "time" t vary, we can think of $\overline{X}_t^{(\infty)}$ as a continuous collection of random variables. Another celebrated theorem says that this collection is a continuous random process known as *Brownian motion*.

Unfortunately, we cannot apply the Central Limit Theorem to our analysis, as our random variables $w(i) - w(i-1)$ are not quite independent. It would be nice if there were some form of the Theorem that allows dependencies between the X_i. Unlike independence, of which there is only one kind, statistical dependencies come in many varieties. As Tolstoy might have said, "All independent variables are alike, all dependent ones are dependent in their own way." However, there is a version of the Central Limit Theorem that covers exactly our form of dependence. Roughly, this version of the theorem says:

Theorem 7.2 (The CLT and Brownian motion for martingales) *Let $X_1, X_2, \ldots$ be random variables such that for all i, $\mathrm{E}[X_i|X_1,\ldots,X_{i-1}] \approx 0$ and $\mathrm{Var}[X_i|X_1,\ldots,X_{i-1}] \approx 1$. Fix $t > 0$ and let $\overline{X}_t^{(m)} = (X_1 + \ldots + X_{\lfloor mt \rfloor})/\sqrt{m}$. As $m \to \infty$, $\overline{X}_t^{(m)}$ converges in distribution to a normal mean 0, variance t random variable $\overline{X}_t^{(\infty)}$. Moreover, the collection $(\overline{X}_t^{(\infty)})_{t \geq 0}$ describes Brownian motion.*

7.3 Dynamics of the walk on the Erdős-Renyi random graph

We now apply these observations to the random graph $\mathcal{G}(n,p)$. With a bit of hindsight, we set $p = \frac{1}{n} + \frac{\lambda}{n^{4/3}}$. It turns out that this is the appropriate scaling for which, as we vary λ from $-\infty$ to ∞, we will observe the emergence of a component of size $O(n^{2/3})$ in the random graph.

In our computation, we will ignore the contribution of the D_i, which can be shown negligible. We have

$$\mathrm{E}[w(i) - w(i-1)|w(0),\ldots,w(i-1)] \approx \mathrm{E}[\mathrm{Binom}(n-i,p) - 1] = \frac{\lambda}{n^{1/3}} + \frac{i}{n} + \frac{\lambda i}{n^{4/3}}$$

and

$$\mathrm{Var}[w(i) - w(i-1)|w(0),\ldots,w(i-1)] \approx (n-i)p(1-p) \approx 1.$$

We now apply the Central Limit Theorem for martingales to the sequence $X_i = w(i) - w(i-1) - \mathrm{E}[w(i) - w(i-1)|w(0),\ldots,w(i-1)]$. For $m = n^{2/3}$, this gives

$$X_{mt} = w(mt) - \sum_{i=1}^{m} t\left(\frac{\lambda}{n^{1/3}} + \frac{i}{n} + \frac{\lambda i}{n^{4/3}}\right) = \lambda t n^{1/3} - \frac{1}{2}t^2 n^{1/3} + o(n^{1/3})$$

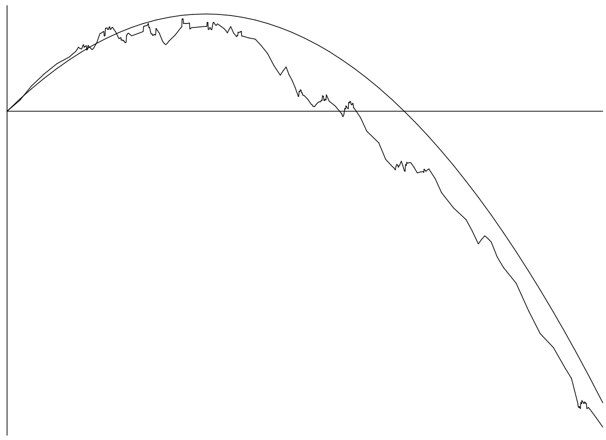
so that, as $n \to \infty$,

$$\frac{w(tn^{2/3}) - (\lambda t - t^2/2)n^{1/3}}{n^{1/3}} \xrightarrow{\mathrm{d}} \overline{X}_t^{(\infty)}.$$

Rearranging terms, we obtain

$$\frac{w(tn^{2/3})}{n^{1/3}} \xrightarrow{\mathrm{d}} \lambda t - t^2/2 + \overline{X}_t^{(\infty)}.$$

This is Brownian motion superimposed on the parabola $\lambda t - t^2/2$. For $\lambda > 0$, the evolution of $w(tn^{2/3})/n^{1/3}$ as a function of t will look like this:



The size of the kth component is $tn^{2/3}$, where t is the smallest value at which the curve dips below the line $y = -kn^{-1/3}$. For $\lambda > 0$, the curve might dip under zero a few times before it "takes off", at which point the giant component in $\mathcal{G}(n,p)$ begins to form. This component will keep growing roughly until the time at which the limiting process intersects the t axis. In expectation, this happens when $t = 2\lambda$, so that the giant component has size $\approx 2\lambda n^{2/3}$. After this point, the process follows the drop of the parabola, and we do not expect to see any more large components. For $\lambda < 0$, the function $\lambda t - t^2$ is decreasing over the whole range of t, and no giant component will appear.