

Erdős-Renyi random graphs basics

Nathanaël Berestycki

U.B.C. - class on percolation

We take n vertices and a number $p = p(n)$ with $0 < p < 1$. Let $G(n, p(n))$ be the graph such that there is an edge between any two given vertices i and j with probability $p(n)$, and the existence of an edge between two vertices are independent events for different pairs of vertices. We denote by P_p the law of this random variable on the space of all graphs with n vertices. (The graph $G(n, p(n))$ is now known as an Erdős-Renyi random graph). The question asked by Erdős and Renyi was the following: what can be said about the properties of the graph for various scaling of the function $p(n)$?

1 Connectivity

Our first example concerns connectivity of the random graph. Let A denote the event that the graph is connected.

Theorem 1. Erdős and Renyi (1960). *Let $\varepsilon > 0$. If $p = (1 - \varepsilon) \log n/n$ then $P_p(A) \rightarrow 0$, and if $p = (1 + \varepsilon) \log n/n$ then $P_p(A) \rightarrow 1$.*

We won't prove this result, but it is easy to see that connectivity cannot occur for p smaller than $(1 - \varepsilon) \log n/n$, since simple calculations show that some vertex is still isolated. In fact the random graph becomes connected when the last isolated vertex connects to something else. In order to get a more exciting phase transition, closer to the percolation one, we need to look at different scalings of p .

2 The giant component

The result announced concerns the emergence of a giant cluster that encompasses a positive fraction of all vertices much before the graph becomes completely connected. More precisely, the size of the largest connected component has a phase transition at $p = c/n$ for $c = 1$.

Theorem 2. (Erdős-Renyi, 1959). Suppose $p(n) = c/n$ for $c > 0$. As $n \rightarrow \infty$,

Regime	Size of largest component
$c < 1$ (subcritical regime)	$\frac{3}{(1-c)^2} \log n$
$c = 1$ (critical regime)	$O(n^{2/3})$
$c > 1$ (supercritical regime)	Largest = $\theta(c)n$ Second = $O(\log n)$.

Here, $\theta(c)$ is the probability of survival for a branching process whose offspring distribution is a Poisson random variable with mean c .

This has the precise meaning that the ratio of the size of the largest component to $3(1-c)^{-2} \log n$ (when $c < 1$), or to $\theta(c)n$ (when $c > 1$), converges to 1 in probability. When $c = 1$ there is a non-degenerate limit in distribution, although a satisfactory description of the limiting distribution has only recently been given by Aldous (1997).

2.1 Proof: The branching process approximation

The proof of Theorem 2 relies on a limit theorem for the size of the connected component containing a distinguished vertex, which can be identified as the total progeny of a branching process with offspring distribution a Poisson random variable with mean c (abbreviated into $PGW(c)$, for Poisson Galton-Watson process). This limit theorem will also be very useful in the understanding of the random transposition random walk. Recall that $(Z_i, i \geq 0)$ is a $PGW(c)$ process if it is a Markov chain on the nonnegative integers \mathbf{N} , with transition probabilities $p(k, \cdot) = \mu^{*k}(\cdot)$ where μ is the Poisson(c) distribution and μ^{*k} denotes k -fold convolution of μ with itself. Let $Z = \sum_{i=0}^{\infty} Z_i$ be the total progeny of this process when started with $Z_0 = 1$ individual.

Proposition 1. Let C be the cluster that contains vertex 1. Let $c > 0$ and $p = c/n$. Then as $n \rightarrow \infty$

$$P_p(|C| = k) \rightarrow P(Z = k)$$

Proof. The number of children of vertex 1, $Z_n^1 = |Y_1|$ has distribution Binomial $(n-1, c/n)$ since it has $n-1$ potential neighbors, each being an actual neighbor with probability c/n . This converges to a Poisson(c) limit by elementary computations with Stirling's formula. We can proceed one generation further: given that 1 is connected to exactly n_1 other vertices, each of the n_1 children is connected to a Binomial $(n-1-n_1, c/n)$ number of individuals. This also converges to Poisson(c) distribution. However this time we need to be a little

careful, since some of these children in the second generation could be the same for different individuals in the first. An easy calculation shows that these can be neglected asymptotically. Proceeding like this with further generations gives the proof. \square

Armed with this Proposition, we now finish the sketch of the proof of Theorem 2. When $c < 1$, the expected number of offsprings is < 1 so the branching process dies out. Roughly, there are n clusters of finite size Z with exponential tails, so assuming independence of the size of different clusters gives that the largest cluster is of order $O(\log n)$.

When $c > 1$, there is a probability $1 - \theta(c)$ that $PGW(c)$ dies out. This means that for roughly $(1 - \theta(c))n$ vertices, their cluster dies very quickly. For the remaining $\theta(c)n$ vertices, their cluster grows quite large, but this observation is not enough to understand the existence of the giant component: it remains to see why there is only one such large component.

2.2 Uniqueness of the giant cluster

We need to see why all vertices whose cluster is quite large are actually part of the same cluster. In the Proposition above, we proved finite-dimensional distribution convergence toward those of a $PGW(c)$. However a more sophisticated argument (a consequence of the birthday problem) shows that this approximation is reasonable at least until $n^{1/2}\omega(n)$ have been exposed in this growing procedure, where $\omega(n) \rightarrow \infty$ slowly enough. Let x and y be two vertices whose clusters are larger than $n^{1/2}\omega(n)$, and suppose that by the time each has size $n^{1/2}\omega(n)$, they haven't intersected yet. We show that it is likely that they will intersect at the next generation. Indeed, there are roughly $n^{1/2}\omega(n)$ vertices in the last generation of each cluster, which means that there are potentially $N = \binom{n^{1/2}\omega(n)}{2}$ edges between these two sets. Therefore, the probability that there is no edge between the two sets of vertices is smaller than

$$(1 - c/n)^N \leq \exp(-\omega(n)^2) \rightarrow 0$$

which explains the uniqueness of the giant component. For more on this approach (and real proofs), see the recent notes of Durrett (2005).

3 Exact asymptotics for $|C|$

Let C be the cluster containing a distinguished vertex, say 1. Let $p = c/n$.

Theorem 3. *For $k \geq 1$, $P(|C| = k) \rightarrow \frac{1}{c} \frac{k^{k-1}}{k!} (ce^{-c})^k$. This is known as the Borel distribution.*

Since we have already identified that $|C| \rightarrow_d Z$, this means that

Corollary 1. *Let Z be the total progeny of a $PGW(c)$. Then Z has a Borel (c) distribution.*

3.1 Proof: combinatorics

We start by doing a different problem and count the expected number $N(k)$ of clusters of a given size k in the random graph. Using the previous results we may simplify the problem and only count the number of trees of size k . Each given tree on k given vertices has probability

$$\begin{aligned} & \left(\frac{c}{n}\right)^{k-1} \left(1 - \frac{c}{n}\right)^{\binom{k}{2} - (k-1) + k(n-k)} \\ \sim & n^{-(k-1)} \frac{1}{c} (ce^{-c})^k \end{aligned}$$

to be one of the connected components of the random graphs. By Cayley's (1889) formula, there are k^{k-2} ways to draw a tree on k given vertices, so, since $\binom{n}{k} \sim n^k/k!$ for fixed k , the expected number of clusters of size k is, approximately,

$$E(N(k)) \sim n \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k$$

Now, observe that by exchangeability,

$$P(|C| = k) = E\left(\frac{kN(k)}{n}\right)$$

which gives the result.

4 A closer look at the critical regime

In (1997) David Aldous gave a very precise description of the component sizes of the random graph at and near the point of its phase transition. Remember that the critical value of p is $p = 1/n$. Aldous takes

$$p = \frac{1}{n} + \lambda n^{-4/3} = \frac{1}{n}(1 + \lambda n^{-1/3})$$

for $\lambda \in \mathbf{R}$, that is $c = 1 + \lambda n^{-1/3}$.

To state his result, we call $K_1^n \geq K_2^n \geq \dots$ the ordered component sizes of $G(n, p)$. Let $(B_s, s \geq 0)$ be a standard one-dimensional Brownian motion. Call X the process such that

$$X_t = B_t + \int_0^t (\lambda - s) ds = B_t + \lambda t - t^2/2$$

X_t is a Brownian motion with a parabolic drift. Let \underline{X} be the process obtained by reflecting X above its infimum:

$$\underline{X}_t = X_t - \inf_{0 \leq s \leq t} X_s$$

If X was Brownian motion then an old result of Paul Lévy says that \underline{X} would have the same distribution as the absolute value $|B_t|$ of a Brownian motion, but here \underline{X} is a bit more complicated because of the parabolic drift.

Having made these definitions, we may state Aldous' result:

Theorem 4. Fix $\lambda \in \mathbf{R}$. As $n \rightarrow \infty$

$$\{n^{-2/3}K_j^n; j \geq 1\} \rightarrow_d \{L_j; j \geq 1\}$$

where L_j are the ranked lengths of the excursions of \underline{X} away from 0.

Here we don't worry too much about the sense of the convergence in distribution, but if you are curious, to know the exact meaning of this statement we need to put a metric on the space of nonincreasing nonnegative sequences, which here is the ℓ^2 metric.

The sketch of proof given here is inspired by Durrett's notes (2005), but they follow quite closely Aldous' original arguments. The idea is to expose the vertices of a cluster one at a time instead of generation per generation when we did the branching process approximation. We proceed in discrete time and start at vertex 1. We split the vertices into three categories. We will have a set of removed sites R_t , corresponding to vertices whose connections have been entirely explored by time t . Thus initially $R_0 = \emptyset$. Then we will have a set of active sites A_t corresponding to sites who have discovered by time t but their connections have not been completely explored (hence initially $A_0 = \{1\}$). Finally, there will be the set of unexplored sites U_t , consisting of all the other vertices in the graph. We explore the cluster one vertex at a time, this vertex being in the active set. If it has neighbours that were previously never discovered before, we add them to the active sites and remove the one being explored from the list of active sites. This process is called *breadth-first walk* by computer scientists. Note that since we explore one vertex at a time we always have that $|U_t| = t$. When we are exploring the last vertex of a cluster we take the smallest integer in U_t and then begin exploring its cluster.

Note that as long as we are exploring a cluster, $|A_t| > 0$ and $A_t = 0$ only at the beginning and at the end of this exploration, so that the size of the cluster is exactly the length of the excursion of A_t away from 0. In fact to be consistent with ourselves we will artificially subtract 1 to $|A_t|$ each time we finish the exploration of a cluster. We call a_t the resulting process. Then the sizes of the clusters are the lengths of the excursions of a_t away from its infimum.

Hence the result will be proved (... or at least plausible!) if we show that after speeding time by a factor of $n^{2/3}$ (to scale the size of the components) and the increments of a_t by a factor of $n^{1/3}$ in order to have a random walk scaling,

$$x_s = n^{-1/3}a_{\lfloor sn^{2/3} \rfloor} \rightarrow_d X_s$$

On the other hand, note that the increments of a_t are given by the number of neighbors of the active site being explored. Since at time t it has $n - t - a_t$ potential new neighbors, each with probability p ,

$$E(\Delta a_t | \mathcal{F}_t) = -1 + (n - t - a_t)p$$

while

$$\text{var}(\Delta a_t | \mathcal{F}_t) = (n - t - a_t)p(1 - p)$$

After taking $p = (1 + \lambda n^{-1/3})/n$ and speeding up time by a factor of $n^{2/3}$ and space by $n^{-1/3}$:

$$\begin{aligned} E(\Delta x_s | \mathcal{F}_{\lfloor sn^{2/3} \rfloor}) &= (\lambda - s)n^{-2/3} + o(n^{-2/3}) \\ &\sim (\lambda - s)ds \end{aligned}$$

since now the time-increments are $ds = n^{-2/3}$. On the other hand

$$\text{var}(\Delta x_s | \mathcal{F}_{\lfloor sn^{2/3} \rfloor}) \sim n^{-2/3} = ds$$

In other words, asymptotically

$$M_t = x_t - \int_0^t (\lambda - s)ds$$

is a martingale and

$$M_t^2 - \int_0^t ds = M_t^2 - t$$

is another martingale. Hence asymptotically, x_s is continuous and satisfies the martingale problem¹ associated with the parabolic drift and diffusion coefficient equal to 1, which is the process X_t . By uniqueness in law of the solution to the martingale problem, we conclude that

$$x_t \rightarrow_d X_t$$

as $n \rightarrow \infty$, which finishes the proof of Aldous' result. \square

In fact, Aldous' results are stronger than what is stated here. In particular, a simple modification of the arguments also yield results about the corresponding complexity of the component and also allows to discuss the much more complex problem of the dynamics of the components as λ evolves from $-\infty$ to $+\infty$. To do this he introduced a process called the multiplicative coalescent.

¹If you need to refresh your memory on martingale problems, here is a quick reminder. Recall that a process continuous process X is said to satisfy the martingale problem associated with the functions f and σ if

$$M_t = X_t - \int_0^t f(s, X_s)ds$$

is a martingale and

$$M_t^2 - \int_0^t \sigma^2(s, X_s)ds$$

is another martingale. It can be shown that if X satisfies the (f, σ) -martingale problem and that f and σ are nice enough, then X is in fact a solution to the stochastic differential equation:

$$dX_t = \sigma(t, X_t)dW_t + f(t, X_t)dt$$