

Connected components in Erdős-Rényi random graphs

Karen Gunderson *

26 November – 3 December 2013

These notes contain the material presented in class on the component structure of $G_{n,p}$. The particular focus is when is $G_{n,p}$ likely to have just one component, and in the case that $G_{n,p}$ is not connected, how big are the largest components likely to be? There are many resources on these topics, of varying degrees of difficulty (e.g. [1, 2, 4, 7]).

1 Connectedness

The first question we shall address is, for which values of p is $G_{n,p}$ either very likely to be connected, or very likely to be disconnected. One of the simplest reasons for a graph to be disconnected is that it contains an isolated vertex. In this section, it is shown that, in fact, isolated vertices are the main obstruction to connectedness in $G_{n,p}$.

For any $k \geq 1$, let X_k denote the number of connected components of k vertices in $G_{n,p}$. Then, X_1 is the number of isolated vertices in $G_{n,p}$ and

$$\mathbb{E}(X_1) = n(1-p)^{n-1}.$$

Set $\mu_1 = \mu_1(n, p) = n(1-p)^{n-1}$. As a rough estimate, $\mu_1 \approx ne^{-pn} = e^{\log n - pn}$ and so when p is much larger than $\frac{\log n}{n}$, one might expect very few isolated vertices, on average, and when p is much smaller than $\frac{\log n}{n}$, one might expect many isolated vertices, on average.

To make this rough estimate precise, let $\alpha(n)$ be a function such that $\alpha(n) \rightarrow \infty$ and $\alpha(n) \leq \log \log n$. Note that, in much of the literature, $\omega(n)$ is used to denote a function growing arbitrarily slowly and tending to infinity, but α is used here to avoid confusion with a previous use of ω .

First consider

$$p = p(n) = \frac{\log n + \alpha(n)}{n} = \frac{\log n}{n} (1 + o(1)).$$

For these values of p ,

$$\mu_1 = n(1-p)^{n-1} \leq ne^{-p(n-1)} = ne^{-\log n - \alpha(n)} e^p \leq e^{-\alpha(n)} = o(1).$$

On the other hand, for

$$p = p(n) = \frac{\log n - \alpha(n)}{n} = \frac{\log n}{n} (1 - o(1)),$$

then the expected number of isolated vertices is

$$\mu_1 = n(1-p)^{n-1} \geq n(1-p)^n \geq ne^{-(p+\alpha(n))n} = e^{\alpha(n)} e^{-p^2 n} \rightarrow \infty.$$

The final expression tends to infinity since $p^2 n = o(1)$.

Neither of these estimates for μ_1 is sufficient, on its own, to determine whether or not $G_{n,p(n)}$ is likely to be connected or disconnected. In the case that $p(n) < \frac{\log n}{n}$, the second moment

*karen.gunderson@bristol.ac.uk

method shall be used to show that with high probability, there is an isolated vertex and hence $G_{n,p}$ is disconnected. Conversely, for the case $p(n) > \frac{\log n}{n}$, it is shown below that the expected number of connected components with fewer than n vertices is small and hence that $G_{n,p}$ is likely to be connected. These are each proved separately.

Proposition 1. Let $\alpha(n) \geq 0$ be such that $\alpha(n) \rightarrow \infty$ and $\alpha(n) \leq \log \log n$. Let $p(n) = \frac{\log n - \alpha(n)}{n}$, then $G_{n,p(n)}$ is disconnected with high probability.

Proof. Chebyshev's inequality is used to show that $\mathbb{P}(X_1 = 0) = 1 - o(1)$. In order to compute $\text{Var}(X_1)$, consider the random variable $X_1(X_1 - 1)$, which is precisely the number of ordered pairs of distinct vertices (v, w) so that both v and w are isolated. Thus,

$$\begin{aligned} \mathbb{E}(X_1(X_1 - 1)) &= n(n-1)(1-p)^{2n-3} \\ &\leq n^2(1-p)^{2n-2} \frac{1}{1-p} \\ &= \frac{\mu_1^2}{1-p} \\ &\leq \mu_1^2 + 1. \end{aligned} \quad (\text{for } n \text{ large enough})$$

One can bound the variance of X_1 by

$$\begin{aligned} \text{Var}(X_1) &= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 \\ &= \mathbb{E}(X_1(X_1 - 1)) + \mathbb{E}(X_1) - \mathbb{E}(X_1)^2 \\ &\leq \mu_1^2 + 1 + \mu_1 - \mu_1^2 = \mu_1 + 1. \end{aligned}$$

By Chebyshev's inequality, and using $\mu_1 \rightarrow \infty$,

$$\mathbb{P}(X_1 = 0) \leq \frac{\text{Var}(X_1)}{\mathbb{E}(X_1)^2} = \frac{\mu_1 + 1}{\mu_1^2} = o(1).$$

Therefore, with high probability, there are isolated vertices in $G_{n,p}$ and hence $G_{n,p}$ is disconnected. $\square$

To show that for p slightly larger than $\frac{\log n}{n}$, $G_{n,p}$ is connected, components of nearly all sizes are considered. In order to compute the expected number of components of size k , one can use the fact that every connected component contains a spanning tree and there are exactly k^{k-2} different labelled trees on k vertices. Since every tree on k vertices has exactly $k-1$ edges,

$$\begin{aligned} \mathbb{E}(X_k) &= \sum_{|V|=k} \mathbb{P}(V \text{ connected}) (1-p)^{k(n-k)} \\ &\leq \sum_{|V|=k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} \\ &= \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} \\ &\leq \left(\frac{ne}{k}\right)^k k^{k-2} p^{k-1} (1-p)^{k(n-k)} \\ &\leq n^k e^k p^{k-1} e^{-pk(n-k)}. \end{aligned}$$

Proposition 2. Let $\alpha(n) \geq 0$ be such that $\alpha(n) \rightarrow \infty$ and $\alpha(n) \leq \log \log n$. Let $p(n) = \frac{\log n + \alpha(n)}{n}$, then $G_{n,p(n)}$ is connected with high probability.

Proof. Note that, if a graph G is disconnected, then there is at least one connected component on at most $n/2$ vertices. Therefore, to show that with high probability, $G_{n,p}$ is connected, it suffices to show that $\mathbb{E}\left(\sum_{k=1}^{n/2} X_k\right) = o(1)$. This sum is broken into a few parts that are estimated separately. Throughout the proof, it is assumed that n is large enough so that $p \leq (2 \log n)/n$.

Note that, as was shown previously, $\mathbb{E}(X_1) = \mu_1 = o(1)$.

Consider now the case $k \leq \log \log n$. Using $p > (\log n)/n$

$$\begin{aligned} \mathbb{E}(X_k) &\leq n^k e^k p^{k-1} e^{-pk(n-k)} \\ &\leq n^k e^k \left(\frac{2 \log n}{n}\right)^{k-1} e^{-k \log n + pk^2} \\ &\leq n^k e^k \left(\frac{2 \log n}{n}\right)^{k-1} n^{-k} e \quad (\text{since } pk^2 = o(1)) \\ &= e^2 \left(\frac{2e \log n}{n}\right)^{k-1}. \end{aligned}$$

Thus, summing over $k \leq \log \log n$,

$$\sum_{k=2}^{\log \log n} \mathbb{E}(X_k) \leq \sum_{k=2}^{\log \log n} e^2 \left(\frac{2e \log n}{n}\right)^{k-1} \leq e^2 \frac{\frac{2e \log n}{n}}{1 - \frac{2e \log n}{n}} = o(1).$$

For $k > \log \log n$, we use the fact that $n - k \geq n/2$ so that

$$\begin{aligned} \mathbb{E}(X_k) &\leq n^k e^k p^{k-1} e^{-pk(n-k)} \\ &\leq n^k e^k \left(\frac{2 \log n}{n}\right)^{k-1} e^{-pkn/2} \\ &\leq ne^k (2 \log n)^{k-1} e^{-k/2 \log n} \\ &= en^{1/2} \left(\frac{2e \log n}{n^{1/2}}\right)^{k-1}, \end{aligned}$$

from which, one can show that

$$\sum_{k=\log \log n+1}^{n/2} \mathbb{E}(X_k) \leq \frac{n}{2} n^{1/2} \left(\frac{2e \log n}{n^{1/2}}\right)^{\log \log n} = o(1).$$

Thus, $\sum_{k=1}^{n/2} \mathbb{E}(X_k) = o(1)$ and so with high probability, there are no components of $G_{n,p}$ with fewer than $n/2$ vertices. Thus, with high probability $G_{n,p}$ is connected. $\square$

Exercise 1. Recall that a *spanning tree* of a graph G on n vertices is a subgraph of G that is a tree (connected and acyclic) that contains all vertices.

1. For any n, p , compute the expected number of spanning trees in $G_{n,p}$. (*Hint: use Cayley's formula for the number of trees*).
2. Find a function $p(n)$ so that with high probability $G_{n,p(n)}$ contains an isolated vertex and the expected number of spanning trees tends to infinity.

2 Review of branching processes

Branching processes are a random model of infinite trees that can be used to describe a breadth-first search in a graph random. A few basic facts about branching processes are given here.

Definition 3. Let ξ be a probability distribution on the non-negative integers. A probability distribution on the set of rooted ordered trees (finite and infinite) called a *branching process* is defined using ξ as follows: Start with a single vertex, called the root. Every vertex has a random number of ‘children’ each with distribution ξ , independently of all others. Vertices are joined to each of their child vertices by an edge and the random rooted tree is denoted by T_ξ .

For each $n \geq 0$, let Z_n denote the number of vertices in T_ξ at distance n from the root. Then $Z_0 = 1$ and for every $n \geq 0$, let $\xi_1, \xi_2, \dots, \xi_{Z_n}$ be independent copies of ξ , then

$$Z_{n+1} \sim \xi_1 + \xi_2 + \dots + \xi_{Z_n}.$$

Exercise 2. Show that the sequence $(Z_n)_{n=0}^\infty$ is a Markov Chain on $\mathbb{N} = \{0, 1, 2, \dots\}$.

One question of interest is when the tree T_ξ is finite or infinite. Note that if $\xi \geq 1$, then with probability 1, the tree T_ξ is infinite. On the other hand, if $\mathbb{P}(\xi = 0) > 0$, then $\mathbb{P}(T_\xi \text{ is finite}) > 0$ (*Exercise: prove this*).

In general if there exists n such that $Z_n = 0$, then the process is said to *die out* or *become extinct*. If, instead, for all $n \geq 0$, $Z_n > 0$, the process is said to *survive*.

Definition 4. Let for every $k \geq 0$, set $p_k = \mathbb{P}(\xi = k)$ and define $\mu = \mathbb{E}(\xi) = \sum_{k=0}^\infty k p_k \in [0, \infty]$.

Lemma 5. For every $n \geq 0$, $\mathbb{E}(Z_n) = \mu^n$.

Proof. The proof proceeds by induction. The base case holds since $\mathbb{E}(Z_0) = 1$.

Assume for some $n \geq 0$ that $\mathbb{E}(Z_n) = \mu^n$.

$$\begin{aligned} \mathbb{E}(Z_{n+1}) &= \mathbb{E}(\mathbb{E}(Z_{n+1} | Z_n)) \\ &= \sum \mathbb{P}(Z_n = k) \mathbb{E}(Z_{n+1} | Z_n = k) \\ &= \sum \mathbb{P}(Z_n = k) \mathbb{E}(\xi_1 + \dots + \xi_k | Z_n = k) \\ &= \sum \mathbb{P}(Z_n = k) \mathbb{E}(\xi_1 + \dots + \xi_k) = \sum \mathbb{P}(Z_n = k) k \mathbb{E}(\xi_i) \\ &= \mathbb{E}(Z_n) \mu = \mu^{n+1}. \end{aligned}$$

□

Corollary 6. If $\mu < 1$ then $\mathbb{P}(T_\xi \text{ survives}) = 0$.

Proof. For every $n \geq 0$,

$$\begin{aligned} \mathbb{P}(T_\xi \text{ survives}) &\leq \mathbb{P}(Z_n > 0) = \mathbb{E}(1_{\{Z_n > 0\}}) \\ &\leq \mathbb{E}(Z_n) && (\text{since } Z_n \in \mathbb{Z}^+) \\ &= \mu^n. \end{aligned}$$

Since n was arbitrary and $\mu^n \rightarrow 0$, $\mathbb{P}(T_\xi \text{ survives}) = 0$.

□

One can show that if $\mu \geq 1$, the $\mathbb{P}(T_\xi \text{ survives}) > 0$ directly, but the proof can also be obtained by looking at the analytic properties of the generating function for ξ .

Recall that for a random variable $\xi \in \mathbb{N}$ where for every k , $p_k = \mathbb{P}(\xi = k)$, the generating function for ξ is defined as

$$f(x) = \mathbb{E}(x^\xi) = \sum_{k=0}^\infty p_k x^k.$$

Note that $f(x)$ is an analytic function for all $x \in \mathbb{C}$ with $|x| < 1$.

Example 7. If ξ is given by a Binomial distribution: $p_k = \binom{n}{k} p^k (1-p)^{n-k}$, then $f(x) = \sum \binom{n}{k} p^k (1-p)^{n-k} x^k = (1-p + px)^n$.

Example 8. Let $\lambda > 0$ and let $\xi \sim \text{Po}(\lambda)$. The generating function for ξ is $f(x) = \sum_k \frac{e^{-\lambda} \lambda^k}{k!} x^k = e^{-\lambda(1-x)}$.

Some basic properties of the generating function of a distribution ξ are summarized in the following lemma.

Lemma 9. 1. $f(x)$ and all derivatives exist and are non-negative for $x \in [0, 1]$.

2. $f(x)$ is continuous, increasing, convex in $[0, 1]$.

3. $f(0) = p_0, f(1) = 1$.

4. $\mathbb{E}\xi = \begin{cases} f'(1) & \text{if left derivative } f'(1) \text{ exists} \\ \infty & \text{otherwise} \end{cases}$

5. $\text{Var}(\xi) = \begin{cases} f''(1) + f'(1)(1 - f'(1)) & \text{if left second derivative } f''(1) \text{ exists} \\ \infty & \text{otherwise} \end{cases}$

Proof. The function $f(x) = \sum p_k x^k$ is analytic in $|x| < 1$ and so all derivatives are also analytic and obtained by differentiating term by term in $|x| < 1$. Since all coefficients are positive, the derivatives of f are non-negative for x in $[0, 1]$. This proves part 1. The remaining parts can be deduced from part 1, taking one-sided limits where necessary. $\square$

Using iterations, the function f can also be used to encode the distribution of the n -th level of a branching process T_ξ . For every n , define

$$f_{Z_n}(x) \stackrel{\text{def}}{=} \mathbb{E}(x^{Z_n}).$$

Lemma 10. Let $f(x)$ be the generating function for ξ . For every $n \geq 1$, $f_{Z_n}(x) = f(f(\cdots f(x) \cdots))$, where this is the n -fold iterate of $f(x)$.

Proof. The proof is by induction on n . $f_{Z_0}(x) = \mathbb{E}(x^{Z_0}) = \mathbb{E}(x^1) = x$.

$$\begin{aligned} f_{Z_{n+1}}(x) &= \mathbb{E}(x^{Z_{n+1}}) = \mathbb{E}(\mathbb{E}(x^{Z_{n+1}} | Z_n)) = \sum \mathbb{P}(Z_n = k) \mathbb{E}(x^{\xi_1 + \cdots + \xi_k}) \\ &= \sum \mathbb{P}(Z_n = k) \mathbb{E}(x^{\xi_1}) \mathbb{E}(x^{\xi_2}) \cdots \mathbb{E}(x^{\xi_k}) = \sum \mathbb{P}(Z_n = k) f(x)^k = f_{Z_n}(f(x)) \end{aligned}$$

$\square$

Definition 11. For any distribution ξ , denote the *extinction probability* by $p_e = p_e(\xi) = \mathbb{P}(T_\xi \text{ is finite})$.

Set $\rho = \rho(\xi) = 1 - p_e$, called the *survival probability*.

Theorem 12. Let ξ be a distribution with generating function $f(x)$. The extinction probability $p_e \in [0, 1]$ is the smallest solution of the equation $f(x) = x$.

Proof. Let p be smallest solution to $f(x) = x$. To see that p_e is a solution to $f(x) = x$, note that

$$\begin{aligned} p_e &= \mathbb{P}(T_\xi \text{ goes extinct}) = \mathbb{P}(Z_n = 0 \text{ for some } n) \\ &= \mathbb{P}(\cup_n \{Z_n = 0\}) = \lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) = \lim_{n \rightarrow \infty} f^{(n)}(0) \end{aligned}$$

Thus, since the function f is continuous,

$$f(p_e) = f(\lim_{n \rightarrow \infty} f^{(n)}(0)) = \lim_{n \rightarrow \infty} f(f^{(n)}(0)) = \lim_{n \rightarrow \infty} f^{(n+1)}(0) = p_e$$

To show that $p_e = p$, note that since $0 \leq p$ then $f(0) \leq f(p) = p$ since f is increasing. So, $f^{(n)}(0) \leq p$. Thus $p_e = \lim_{n \rightarrow \infty} f^{(n)}(0) \leq p$ and by minimality of p , $p_e = p$. $\square$

Corollary 13. If $\mathbb{E}\xi \leq 1$, and ξ is not identically 1, then $\mathbb{P}(\text{extinction}) = 1$. If $\mathbb{E}\xi > 1$, then $\mathbb{P}(\text{extinction}) < 1$.

Proof. If $\mathbb{E}\xi > 1$ then $f'(1) > 1$ and so there exists $h > 0$ such that $\frac{f(1)-f(1-h)}{h} > 1$. That is, $f(1-h) < 1-h$. Apply the intermediate value theorem to $f(x) - x$ to obtain $x \in [0, 1-h]$ such that $f(x) = x$. So, $p_e \leq 1-h < 1$.

Conversely, if $p_e < 1$, then since $f(p_e) = p_e$ and $f(1) = 1$, the convexity of f implies that for every $x \in [p_e, 1]$, $f(x) \leq x$. Thus for $h > 0$ (small enough), $\frac{f(1)-f(1-h)}{h} \geq 1$ and so $\mu = f'(1) \geq 1$.

Two cases: either $\exists x_0 \in (p_e, 1)$ such that $f(x_0) < x_0$ and then for $x \in [x_0, 1]$,

$$f(x) \leq 1 - \left(\frac{1-f(x_0)}{1-x_0} \right) (1-x)$$

for $x \in [x_0, 1]$. Therefore, $\frac{f(1)-f(1-h)}{h} \geq \frac{1-f(x_0)}{1-x_0} > 1$ as $h \rightarrow 0$. Thus $f'(1) > 1$.

Otherwise $f(x) = x$ for all $x \in (p_e, 1)$. Then $f(x) = x$ everywhere and then $\xi = 1$. $\square$

3 The giant component

One of the, perhaps, most startling results given by Erdős and Rényi in their 1959 and 1960 papers on random graphs [5, 6] was that for $p = \lambda/n$, there is a sudden phase transition in the size of the largest connected component of $G_{n,p}$, as λ varies. When $\lambda < 1$, all components are relatively small, $O(\log n)$, and on the other hand, when $\lambda > 1$, there is component on $\Theta(n)$ vertices, with the next largest component having size $O(\log n)$. In this section, we shall use branching processes to explore the component structure of a random graph and deduce a form of Erdős and Rényi's result.

One of the useful facts that shall be needed from probability is that

$$\text{Bin}(n, \lambda/n) \xrightarrow{p} \text{Po}(\lambda),$$

meaning that as n tends to infinity, binomial random variables with parameters n and λ/n converge, in probability to a Poisson random variable with mean λ . Indeed, for any fixed $c > 0$,

$$\text{Bin}(n - c, \lambda/n) \xrightarrow{p} \text{Po}(\lambda).$$

Thus, for small values of k , a breadth-first search in $G_{n,\lambda/n}$ in a component on k vertices 'looks like' a branching process with offspring distribution $\text{Po}(\lambda)$ that has exactly k vertices.

For any vertex v , let C_v denote the connected component of v , in $G_{n,\lambda/n}$. One can explore the component of C_v vertex by vertex starting with $R_0 = B_0 = \{v_0\}$. For $k \geq 0$, R_k is the set of vertices 'reached' at time k and B_k is the set of 'boundary' vertices at time k (those that have been reached, but have unexplored neighbourhoods). For each $k \geq 0$, pick $w \in B_k$ and let $R_{k+1} = N(w) \cap V \setminus (R_0 \cup \dots \cup R_k)$: the newly discovered vertices reached when the neighbourhood of w is explored. Define $B_{k+1} = B_k \setminus \{w\} \cup R_{k+1}$. This exploration stops when $B_k = 0$ and at this time, we will have constructed a spanning tree of the component of v . Note that at each time, conditioned on the values of $|R_0|, |R_1|, \dots, |R_k|$,

$$|R_{k+1}| \sim \text{Bi}(n - |R_0| - |R_1| - \dots - |R_k|, p).$$

Since the distributions of number of newly reached vertices is not always the same, this exploration is not a branching process. However, it can be closely approximated by an appropriately chosen branching process.

Use T_λ for a branching process with offspring distribution $\text{Po}(\lambda)$. For every k , set $\rho_k(\lambda) = \mathbb{P}(|T_\lambda| = k)$.

Lemma 14. Let λ, k be fixed constants. For every vertex v of $G_{n,\lambda/n}$,

$$\mathbb{P}(|C_v| = k) \rightarrow \rho_k(\lambda).$$

Proof. Fix a rooted, ordered tree T^* with k vertices and then

$$\mathbb{P}(\text{spanning-tree rooted at } v \simeq T^*) \rightarrow \mathbb{P}(T_\lambda = T^*).$$

The proof is completed by summing over finitely many trees on k vertices. $\square$

For every k , let $N_k(G)$ denoted the number of vertices in a graph G in components of order k . Recall that $X_k(G)$ was used for the number of components with k vertices so that $N_k(G) = kX_k(G)$.

Use N_k for $N_k(G_{n,\lambda/n})$.

Corollary 15. For every fixed k, λ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}(N_k) = \rho_k(\lambda).$$

Proof.

$$\mathbb{E}(N_k) = \mathbb{E}(\#v \text{ s.t. } |C_v| = k) = \sum_{v \in V} \mathbb{P}(|C_v| = k) = n\mathbb{P}(|C_v| = k).$$

$\square$

With a little more careful counting, one can show the following.

Lemma 16. For λ, k fixed,

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \mathbb{E}(N_k^2) = \rho_k(\lambda)^2.$$

For a random variables $(X_n)_{n \geq 1}$ and $x^* \in \mathbb{R}$, the notation $X_n \xrightarrow{p} x^*$ shall be used to mean that $(X_n)_{n \geq 1}$ converges *in probability* to x^* . That is, for every $\varepsilon > 0$, there is an N_ε so that for $n \geq N_\varepsilon$,

$$\mathbb{P}(|X_n - x^*| > \varepsilon) < \varepsilon.$$

Using Chebyshev's inequality, Lemmas 15 and 16 show that for every k ,

$$\frac{N_k}{n} \xrightarrow{p} \rho_k(\lambda).$$

Define the following quantities:

$$\begin{aligned} N_{\leq k}(G) &= N_1(G) + \cdots + N_k(G), \\ N_{> k}(G) &= \# \text{ vertices in components with } > k \text{ vertices,} \\ \rho_{\leq k}(\lambda) &= \mathbb{P}(|T_\lambda| \leq k). \end{aligned}$$

Corollary 17. For k, λ fixed

1. $\frac{1}{n} N_{\leq k} \xrightarrow{p} \rho_{\leq k}(\lambda)$, and
2. $\frac{1}{n} N_{> k} \xrightarrow{p} 1 - \rho_{\leq k}(\lambda)$.

Proof. For the first, note that k is a fixed finite number, so

$$\frac{1}{n} N_{\leq k} = \frac{1}{n} \sum_{i=1}^k N_i \xrightarrow{p} \sum_{i=1}^k \rho_i(\lambda) = \rho_{\leq k}(\lambda).$$

For the other, note that

$$\frac{1}{n} N_{> k} = 1 - \frac{1}{n} N_{\leq k} \xrightarrow{p} 1 - \rho_{\leq k}(\lambda).$$

$\square$

Using some general properties of sequences of random variables, this result extends to the following.

Theorem 18. Let λ be constant. There is a function $\alpha(n) \rightarrow \infty$ slowly enough so that

1. $\frac{1}{n}N_{\leq \alpha(n)}(G_{n,\lambda/n}) \xrightarrow{p} 1 - \rho(\lambda)$, and
2. $\frac{1}{n}N_{> \alpha(n)}(G_{n,\lambda/n}) \xrightarrow{p} \rho(\lambda)$.

Proof. Use the fact that $\sum_k \rho_k(\lambda) = \mathbb{P}(T_\lambda \text{ is finite}) = 1 - \rho(\lambda)$. □

From this theorem, we know that a $\rho(\lambda)$ fraction of vertices in $G_{n,\lambda/n}$ are in ‘big’ components (of size at least $\alpha(n)$) and the rest are all in ‘small’ components (of size at most $\alpha(n)$). What remains is to show that there are not many big components - there is only a single ‘giant’ component. As notation, let $L_1(G)$ denote the number of vertices in the largest component of a graph G and $L_2(G)$ denote the number of vertices in the second largest component.

Theorem 19 (Erdős, Rényi [5, 6]). Let λ be constant, then

1. $L_1(G_{n,\lambda/n}) = \rho(\lambda)n + o_p(n)$, and
2. $L_2(G_{n,\lambda/n}) = o_p(n)$.

Proof. First we shall show that there is a single giant component that covers most of the vertices in big components.

Assume that $\lambda > 1$ so that $\rho(\lambda) > 0$ (otherwise there is nothing to prove) and fix $\varepsilon \in (0, \rho(\lambda))$. Since the function ρ is continuous, let λ' be such that $\rho(\lambda') \geq \rho(\lambda) - \varepsilon/3$. Using a technique sometimes known as ‘sprinkling’, the random graph $G_{n,\lambda/n}$ is constructed as a union of two independent graphs: $G_1(n, \lambda'/n)$ and $G_2(n, \delta)$, in which case $\delta \geq \frac{\lambda - \lambda'}{n}$.

Exercise 3. Check that $G_{n,\lambda/n}$ can be defined as $G_1(n, \lambda'/n) \cup G_2(n, \delta)$.

As before, $\frac{1}{n}N_{> \alpha(n)}(G_1) \xrightarrow{p} \rho(\lambda')$. Define a set of vertices in ‘big components’,

$$B = \{v \mid \text{components of } v \text{ in } G_1 \text{ with } > \alpha(n) \text{ vxs}\}.$$

For n sufficiently large, with probability at least $1 - \varepsilon$,

$$\frac{1}{n}|B| \geq \rho(\lambda') - \varepsilon/3 \geq \rho(\lambda) - 2\varepsilon/3.$$

The goal is to show that B is connected. So far, the edges arising from G_2 have been ignored. They are now used to show that, even if B is disconnected in G_1 , the subgraph in B becomes connected when the G_2 -vertices are included. Note that if $|L_1(G)| \leq (\rho(\lambda) - \varepsilon)n$, then

$$|L_1(G)| \leq (\rho(\lambda) - \varepsilon)n \leq |B| - \varepsilon n/3.$$

Call a partition of $B = B_1 \cup B_2$ a *bad partition* if $|B_1|, |B_2| \geq \varepsilon n/6$, there are no G_1 -edges between B_1 and B_2 , and there are no G_2 -edges between the two sets.

If $|L_1(G)| \leq |B| - \varepsilon n/3$, then there exists a partition satisfying the first two conditions and, given (B_1, B_2) , for some constant $c > 0$,

$$\mathbb{P}(\text{no } G_2 \text{ edges b/w } B_1, B_2) = (1 - \delta)^{|B_1||B_2|} \leq \exp(-\delta n^2 \varepsilon^2 / 36) = \exp(-cn).$$

Thus, since there are at most $n/\alpha(n)$ different components in B ,

$$\begin{aligned} \mathbb{P}(L_1(G_{n,\lambda/n}) \leq |B| - \varepsilon n/3) &\leq \mathbb{P}(\exists \text{ a bad partition}) \\ &\leq 2^{n/\alpha(n)} e^{-cn} = o(1). \end{aligned}$$

The final inequality uses the fact that $\alpha(n) \rightarrow \infty$. Thus, for every $\varepsilon > 0$,

$$\mathbb{P}(L_1 \geq (\rho(\lambda) - \varepsilon)n) \geq 1 - \varepsilon.$$

It remains to show that the size of the second largest component is significantly smaller (at least in the case $\lambda > 1$).

Let n be large enough so that $\mathbb{P}(N_{>\alpha(n)} \leq (\rho(\lambda) + \varepsilon/3)n) \geq 1 - \varepsilon$ and $\alpha(n) \leq \varepsilon n/3$.

Deterministically (not just with high probability),

$$L_1 \leq \max\{\alpha(n), N_{>\alpha(n)}\} \leq \alpha(n) + N_{>\alpha(n)}$$

and $L_1 + L_2 \leq N_{>\alpha(n)} + 2\alpha(n)$. Thus, with probability at least $1 - \varepsilon$,

$$\begin{aligned} L_1 &\leq \alpha(n) + N_{>\alpha(n)} \leq (\rho(\lambda) + \varepsilon)n, \\ L_1 + L_2 &\leq N_{>\alpha(n)} + 2\alpha(n) \leq (\rho(\lambda) + \varepsilon)n. \end{aligned}$$

From the previous, with probability at least $1 - \varepsilon$, $L_1 \geq (\rho(\lambda) - \varepsilon)n$ and so with probability at least $1 - 2\varepsilon$, $L_2 \leq 2\varepsilon n$.

As $\varepsilon > 0$ was arbitrary, $L_2 = o_p(n)$. □

3.1 A more careful look at the sub-critical case

In the case where $\lambda < 1$, one can use a slightly different component exploration, introduced by Karp, to get a more precise bound on the size of the largest component in $G_{n,\lambda/n}$.

The exploration is defined as follows. Given $v_0 \in V$, set $R_0 = B_0 = \{v_0\}$. The sets R_n are the ‘reached’ vertices, and the sets B_n are the ‘boundary’ vertices - those with unexplored neighbourhoods. At each step $n \geq 0$, choose a vertex $v \in B_n$ and set $X_{n+1} = |N_{V \setminus R_n}(v)|$ and update

$$\begin{aligned} B_{n+1} &= B_n \setminus \{v\} \cup N_{V \setminus R_n}(v) \\ R_{n+1} &= R_n \cup N_{V \setminus R_n}(v). \end{aligned}$$

Consider the distribution of the random variables $(X_n)_{n \geq 1}$: $X_1 \sim \text{Bi}(n-1, p)$ and for every $n \geq 2$, conditioned on the values of $X_1, \dots, X_n$,

$$X_{n+1} \sim \text{Bi}(n - (X_1 + X_2 + \dots + X_n), p).$$

With inequalities in terms of stochastic domination, for every n ,

$$X_n \leq \text{Bi}(n, p).$$

Furthermore, one can check that for every time $t \geq 1$, $|R_t| = t + 1$ and

$$|B_t| = 1 + \sum_{i=1}^t (X_i - 1) = \left(\sum_{i=1}^t X_i \right) - (t - 1).$$

The above process stops when there are no boundary vertices. At this point, we have found all vertices in the component of v_0 : C_{v_0} . Thus,

$$|C_{v_0}| = 1 + \min\left\{t \mid \sum_{i=1}^t X_i \leq t - 1\right\}.$$

The stochastic domination of the variables X_i and the characterization of C_{v_0} in terms of these variables can be used to give bounds on the size of the largest component.

Proposition 20. Let $\lambda < 1$ be fixed, then with high probability $L_1(G_{n,\lambda/n}) = O(\log n)$.

Proof. Using the vertex exploration described above, for any $t \geq 1$,

$$\begin{aligned} \mathbb{P}(|C_{v_0}| \geq t + 2) &\leq \mathbb{P}\left(\sum_{i=1}^t X_i \geq t\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^t \text{Bi}(n, \lambda/n) \geq t\right) \\ &\leq \mathbb{P}(\text{Bi}(nt, \lambda/n) \geq t). \end{aligned}$$

Using an inequality about tails for Binomial random variables due to Janson, since $\lambda t < t$,

$$\begin{aligned} \mathbb{P}(\text{Bi}(nt, \lambda/n) \geq t) &\leq \exp\left(-\frac{(1-\lambda)^2 t^2}{2(\lambda t + (1-\lambda)t/3)}\right) \\ &\leq \exp\left(-\frac{(1-\lambda)^2 t}{2}\right). \end{aligned}$$

Setting $t = \frac{3 \log n}{(1-\lambda)^2}$, we see that

$$\mathbb{P}(|C_{v_0}| \geq t + 2) \leq \exp\left(-\frac{3}{2} \log n\right) = n^{-3/2}$$

and hence

$$\mathbb{P}(\exists \text{ cmpt with } \geq t + 2 \text{ vxs}) \leq nn^{-3/2} = n^{-1/2} = o(1).$$

Thus, with high probability,

$$L_1(G_{n,\lambda/n}) \leq \frac{3 \log n}{(1-\lambda)^2} + 2 = O(\log n).$$

□

4 Fixed degree distributions

4.1 The configuration model

The configuration model was introduced by Bollobás to study random regular graphs.

In general, given a sequence $d_1, d_2, \dots, d_n$ with the condition that $\sum_{i=1}^n d_i$ is even, a random (multi-)graph with degree sequence $d_1, d_2, \dots, d_n$ can be defined as follows.

For each $i \leq n$, let $S_i = \{(i, j) \mid 1 \leq j \leq d_i\}$. A random graph is generated by choosing, uniformly at random, a matching of the set $\cup S_i$ and then identifying edges: $xy \in E$ iff there are j_x, j_y so that $\{(x, j_x), (y, j_y)\}$ is in the matching. The set S_i can be thought of as the ‘edge stubs’ emanating from vertex i and a edge stubs joined uniformly at random. Note that in general, such a procedure can produce loops and multiple edges. One can either work with the resulting multigraph, or condition on the graph being simple.

Exercise 4. Show that, conditioned on the graph being simple, the configuration model produces a graph with degree sequence $(d_1, d_2, \dots, d_n)$ uniformly at random.

In the case of r -regular graphs, one can generate a random r -regular multigraph using the configuration model with $d_1 = d_2 = \dots = d_n = r$ (on the condition that their sum is even). Bollobás showed that in this random model, the random multi-graph G_r satisfies $\mathbb{P}(G_r \text{ is simple}) > 0$. This allows one to move between a random multi-graph and the random r -regular graphs: if a property holds with high probability in the random multi-graph, the property also holds with high probability in the random simple graph.

4.2 Exploration by branching process

It is known that for $r \geq 3$, with high probability, a random r -regular graph is connected (see, e.g. [1]). In fact, Robinson and Wormald [10] showed that a random regular graph will have, with high probability, a Hamilton cycle!

Molloy and Reed [8, 9] looked at sequences of random graphs with fixed degree distribution and found a critical point for the appearance of a ‘giant’ component in well-behaved sequences of graphs.

Definition 21. A sequence of integer valued functions $\mathcal{D} = \{d_i\}_{i \geq 0}$ is called an *asymptotic degree sequence* iff

1. for all $i \geq n$, $d_i(n) = 0$, and
2. for all n , $\sum_{i \geq 0} d_i(n) = n$.

An asymptotic degree sequence is called *feasible* iff for every n , there exists at least one graph, G , on n vertices so that for every i , G contains $d_i(n)$ vertices of degree i .

Such a sequence is called *smooth* iff there exist a sequence of constants $\{\lambda_i\}$ so that

$$\lim_{n \rightarrow \infty} d_i(n)/n = \lambda_i.$$

Given such a (feasible, smooth) sequence $\mathcal{D}$, one can consider a sequence of random graphs, $G_{n,\mathcal{D}}$, given by the configuration model and the degree distribution given by $\mathcal{D}$. That is, for every n , a random graph is chosen, according to the configuration model so that for every i , there are $d_i(n)$ vertices of degree i . When n is large, such a graph will have, for every i , approximately $\lambda_i n$ vertices of degree i .

Assuming some other technical conditions, Molloy and Reed [8, 9] showed that if $0 < \sum_{i \geq 1} i(i-2)\lambda_i < \infty$, then there exists $c_1, c_2 > 0$, depending only on $\mathcal{D}$ so that with high probability, $G_{n,\mathcal{D}}$ has exactly one component with at least $c_1 n$ vertices and all other components have order at most $c_2 \log n$. On the other hand, if $\sum_{i \geq 1} i(i-2)\lambda_i < 0$, then with high probability, every component is of order $o(n)$.

How does this compare to the results regarding branching processes from the previous section? Set $\mu = \sum_i i\lambda_i$, the (asymptotic) average degree of a vertex in the sequence of random graphs. Then, $\sum_{i \geq 1} i(i-2)\lambda_i > 0$ iff

$$\sum_{i \geq 0} \frac{i(i-1)\lambda_i}{\mu} > 1.$$

Consider a component exploration from a fixed vertex v , as was done in the Erdős-Rényi random graph. When n is large and v is chosen at random, the number of neighbours of v is going to be k with probability approximately λ_k . From there, one can explore the un-matched edge stubs of v . Since it is edge stubs that are matched up initially (and not vertices), a vertex w of degree i has probability of having one of its edge stubs matched to an edge stub of v proportional to approximately $i\lambda_i$. If such a vertex of degree i is chosen, it will have only $i-1$ neighbours in future exploration of the graph.

Thus, a branching process that is slightly different from what was seen previously can be used to describe the component exploration. For the first level (Z_1) of the branching process, the offspring distribution is ξ_0 given by $\mathbb{P}(\xi_0 = k) = \lambda_k$ and for all $n \geq 2$, the n -th level is given from the $(n-1)$ -st with an offspring distribution ξ_1 so that

$$\mathbb{P}(\xi_1 = i-1) = \frac{i\lambda_i}{\mu}.$$

Note that, as we have seen previously, the branching process T_{ξ_1} survives with positive probability if

$$\mathbb{E}(\xi_1) = \sum_i \frac{i(i-1)\lambda_i}{\mu} > 1.$$

One can check that, if T_{ξ_1} survives with positive probability, then so does a random tree with offspring distribution ξ_0 for the first ‘generation’ and distribution ξ_1 for all further generations.

There are a lot of details to fill in for a proof that this branching process is, indeed, a good model for the component exploration in the random graphs with fixed degree distribution. In addition, the ‘sprinkling’ argument used for $G_{n,p}$, does not translate in a straightforward way to graphs with a fixed degree distribution. Some details can be found, for example, in Durrett [4].

References

- [1] Béla Bollobás, *Random graphs*, Second edition, Cambridge Studies in Advanced Mathematics, **73**, Cambridge University Press, Cambridge, 2001.
- [2] Béla Bollobás, Oliver Riordan, Random graphs and branching processes, in *Handbook of large-scale random networks*, pp 15–115, *Bolyai Soc. Math. Stud.*, **18**, Springer, Berlin, 2009.
- [3] Reinhard Diestel, *Graph theory*, Fourth edition, Graduate Texts in Mathematics, **173**, Springer, Heidelberg, 2010.
- [4] Rick Durrett, *Random graph dynamics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010.
- [5] P. Erdős, A. Rényi, On random graphs. I, *Publ. Math. Debrecen* **6** (1959), 290–297.
- [6] P. Erdős, A. Rényi, On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 17–61.
- [7] Svante Janson, Tomasz Łuczak, Andrzej Ruciński, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.
- [8] Michael Molloy, Bruce Reed, A critical point for random graphs with a given degree sequence, in *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, “Random Graphs ’93” (Poznań, 1993)*, *Random Structures Algorithms* **6** (1995), no. 2-3, 161–179.
- [9] Michael Molloy, Bruce Reed, The size of the giant component of a random graph with a given degree sequence, *Combin. Probab. Comput.* **7** (1998), no. 3, 295–305.
- [10] R. W. Robinson, N. C. Wormald, Almost all regular graphs are Hamiltonian, *Random Structures Algorithms* **5** (1994), no. 2, 363–374.