Random Graph Notes

written by Ben Budway

ORF 387 – Princeton University – Spring 2024

The purpose of these notes is to organize all of the information given about Random Graphs, along with some extra properties. We include some proofs of the theorems we present for the interested reader. You are not required to know the proofs, but you should be familiar with the statements of the theorems.

We recall some of the basic definitions in graph theory.

Definition 0.1. *A (unweighted, undirected)* ***graph*** *is a pair $G = (V, E)$, where V is the set of vertices, and $E \subseteq V \times V$ is a subset of (unordered) pairs of vertices, called edges.*

Definition 0.2. *A simple graph is a graph with $G = (V, E)$ with no self loops. That is, for any $v \in V, (v, v) \notin E$.*

Unless explicitly stated otherwise, all graphs are undirected. For any $u, v \in V$, we will use the notation $v \sim u$ when $(v, u) \in E$. We will use the notation $\mathcal{G}$ for a random graph.

1 Erdös-Rényi Random graphs

We begin with the simplest random graph model, introduced by Erdös and Rényi [2].

Definition 1.1. *An Erdös-Rényi graph on n vertices with edge density p, denoted $\mathcal{G}_{ER}(n, p)$ is a random (simple) graph with the following properties:*

- *The vertex set is $V = \{1, ..., n\}$*
- *For each pair $u \neq v \in V$, the probability that the edge (u, v) is present in $\mathcal{G}_{ER}(n, p)$ is given by p independently of all other edges.*

Despite its simple form, Erdös-Rényi graphs provide a rich enough structure to display many of the properties found in real world networks. We begin with a study of connectivity.

Theorem 1.2 (Erdös-Rényi 1961)**.**

- *Let $\lambda > 1$. Then, we have that*

$$\mathbb{P}\Big(\mathcal{G}_{ER}(n, \lambda\frac{\log n}{n}) \text{ is connected}\Big) \to 1.$$

- *Instead, if $\lambda < 1$, we have that*

$$\mathbb{P}\Big(\mathcal{G}_{ER}(n, \lambda\frac{\log n}{n}) \text{ is connected}\Big) \to 0.$$

Proof. We start with the proof of the second claim. The idea will be to show that with probability going to 1, there will exist an isolated vertex. Let $\lambda < 1$, and let $p_n = \lambda \frac{\log n}{n}$. Also, let $u \neq v \in \{1, ..., n\}$. Then, using the independence properties of the graph

$$\mathbb{P}\Big(v \text{ is isolated}\Big) = \mathbb{P}\Big(v \not\sim u\Big)^{n-1} = (1-p_n)^{n-1}.$$

Given two vertices $u \neq v \in V$, we can also calculate the probability that they are both isolated. For u and v to both be isolated, v cannot be connected to the $n-1$ other vertices. u can also not be connected to any other vertices and this includes $n-2$ other vertices (we already counted v. So, in total, we need $n-1+n-2 = 2n-3$ edges to not exist. Therefore,

$$\mathbb{P}(u \text{ and } v \text{ are isolated}) = (1-p_n)^{2n-3}.$$

Let X_n be the number of isolated vertices in $\mathcal{G}_{ER}(n, p_n)$. We will use the inequality that if $X \geq 0$ is a nonnegative random variable, then $\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$ (This is called the Paley-Sygmund inequality). Now, by linearity of expectation, $\mathbb{E}[X_n] = n(1-p_n)^{n-1}$. To calculate the expectation of X_n^2, we write X_n as the sum of indicators.

$$X_n = \sum_{i=1}^{n} \mathbf{1}_{\{\text{vertex } i \text{ is isolated}\}}$$

$$\mathbb{E}[X_n^2] = \mathbb{E}\Big[\sum_{i=1}^{n} \mathbf{1}_{\{\text{vertex } i \text{ is isolated}\}} + \sum_{j \neq i} \mathbf{1}_{\{\text{vertices } i \text{ and } j \text{ are isolated}\}}\Big] = n(1-p_n)^{n-1} + (n^2-n)(1-p_n)^{2n-3}.$$

Now we try to calculate the ratio.

$$\mathbb{P}(X_n > 0) \geq \frac{(\mathbb{E}[X_n])^2}{\mathbb{E}[X_n^2]} = \frac{n^2(1-p_n)^{2n-2}}{n(1-p_n)^{n-1} + (n^2-n)(1-p_n)^{2n-3}} = \frac{n(1-p_n)^{n-1}}{1 + (n-1)(1-p_n)^{n-2}}$$

Remember that we want to show that the above goes to 1 as n goes to ∞. For the numerator, we will use the inequality $1 - x \geq e^{-\frac{x}{1-x}}$ for $x < 1$. Therefore,

$$n(1-p_n)^{n-1} \geq n \exp(-(n-1)\frac{p_n}{1-p_n}) = n \exp(-(n-1)p_n)^{\frac{1}{1-p_n}}$$

$$= n \exp(-\lambda \frac{n-1}{n} \log n)^{\frac{1}{1-p_n}} = n \times n^{-\lambda \frac{n-1}{n} \frac{1}{1-p_n}}$$

For the denominator, we use the inequality $1 - x \leq e^{-x}$.

$$(n-1)(1-p_n)^{n-2} \leq n(1-p_n)^{n-2} \leq n \exp(-(n-2)p_n).$$

$$= n \exp(-\lambda \frac{n-2}{n} \log n) = n \times n^{-\lambda \frac{n-2}{n}}.$$

Therefore, we have the inequality

$$\mathbb{P}(X_n > 0) \geq \frac{n^{1-\lambda \frac{n-1}{n} \frac{1}{1-p_n}}}{1 + n^{1-\lambda \frac{n-2}{n}}} = \frac{1}{n^{-(1-\lambda \frac{n-1}{n} \frac{1}{1-p_n})} + n^{\lambda(\frac{n-1}{n} \frac{1}{1-p_n} - \frac{n-2}{n})}}.$$

Now, p_n goes to 0, and $\lambda < 1$. Therefore, there exists an N and an $\epsilon > 0$ so that for all $n \geq N$, $1 - \lambda \frac{n-1}{n} \frac{1}{1-p_n} \geq \epsilon$. Therefore, for $n \geq N$

$$n^{-(1-\lambda \frac{n-1}{n} \frac{1}{1-p_n})} \leq n^{-\epsilon} \to 0.$$

Now, we want to argue

$$n^{\lambda(\frac{n-1}{n} \frac{1}{1-p_n} - \frac{n-2}{n})} \to 1 \tag{1}$$

First, note that

$$\frac{n-1}{n} \frac{1}{1-p_n} - \frac{n-2}{n} = (1 - \frac{1}{n})(\frac{1}{1-p_n} - 1) + \frac{1}{n} \geq 0$$

Therefore, we have

$$n^{\lambda(\frac{n-1}{n}\frac{1}{1-p_n}-\frac{n-2}{n})} \geq 1.$$

Now, for the upper bound, we use the inequality $\frac{1}{1-x} \leq 1+2x$ when $x \leq \frac{1}{2}$. Therefore,

$$(1-\frac{1}{n})(\frac{1}{1-p_n}-1)+\frac{1}{n} \leq 2p_n(1-\frac{1}{n})+\frac{1}{n} \leq 2\lambda\frac{\log n}{n}+\frac{1}{n} \leq 3\frac{\log n}{n}$$

Therefore, we have that

$$\log n^{\lambda(\frac{n-1}{n}\frac{1}{1-p_n}-\frac{n-2}{n})} \leq 3\lambda\frac{(\log n)^2}{n} \to 0.$$

Taking exponentials proves equation (1). This shows that

$$\mathbb{P}(X_n > 0) \to 1.$$

And so, finally, we have that

$$\mathbb{P}(\mathcal{G}_{ER}(n,p_n) \text{ is not connected}) \geq \mathbb{P}(\mathcal{G}_{ER}(n,p_n) \text{ has an isolated vertex}) = \mathbb{P}(X_n > 0) \to 1.$$

To prove the first claim, note that in order for vertices $\{1,...,k\}$ to be disconnected from the rest of the graph, there must be $k(n-k)$ edges that are not present. Therefore,

$$\mathbb{P}(\{1,...,k\} \text{ is disconnected from the rest of the graph}) = (1-p_n)^{k(n-k)}$$

By taking a union bound over all possible subsets of size k, we find

$$\mathbb{P}(\text{ there exists a subset of size k that is disconnected from the rest of the graph}) \leq \binom{n}{k}(1-p_n)^{k(n-k)}$$

If the graph is disconnected, then there must be a subset of size k that is disconnected from the rest of the graph with $k \leq \frac{n}{2}$ (if the disconnected set is bigger then $\frac{n}{2}$, just look at the complement of that set). Therefore, by taking another union bound, we find

$$\mathbb{P}(\mathcal{G}_{ER}(n,p_n) \text{ is disconnected}) \leq \sum_{k=1}^{\frac{n}{2}} \binom{n}{k}(1-p_n)^{k(n-k)} \tag{2}$$

We claim (without proof) that if $p_n = \lambda\frac{\log n}{n}$ with $\lambda > 1$, then equation (2) goes to 0. The interested reader can see [3] Theorem 4.1 for a proof where they prove a slightly more general result. □

The above theorem leaves open the question of what happens when $\lambda = 1$. It turns out that the probability the graph is connected tends to the constant e^{-1}. One can see [3] for a proof of this result. The proof for $\lambda = 1$ is harder than the rest.

Perhaps the most interesting aspect of the Erdös-Rényi graph is the emergence of a *giant component*. Roughly speaking, a giant component is a connected component of a graph containing a nontrivial fraction of the nodes (specifically, a fraction of the nodes not getting smaller as the number of nodes gets large). We collect the results on the size of the largest component in the following theorem.

Theorem 1.3. *Let $\mathcal{G}_{ER}(n,p_n)$ be an Erdös-Rényi graph with edge probability $p_n = \frac{\lambda}{n}$. (Note that Theorem 1 will imply that with high probability, the graph is disconnected). Then, we have the following*

- *If $\lambda > 1$, then a giant component exists with high probability. Let X_n denote the size of the largest component, and Y_n denote the size of the second largest component. Then, for any $\epsilon > 0$, we have the following:*

  $$\lim_{n\to\infty} \mathbb{P}\Big(|\frac{1}{n}X_n - \gamma_\lambda| > \epsilon\Big) = 0$$

  where γ_λ is the unique solution to the equation $1-e^{\lambda\gamma} = \gamma$. Furthermore, for any $\epsilon > 0$

  $$\lim_{n\to\infty} \mathbb{P}\Big(Y_n > \epsilon n\Big) = 0$$

- *If $\lambda < 1$, then there exists a constant C_λ such that*

$$\lim_{n\to\infty} \mathbb{P}\Big(X_n > C_\lambda \log n\Big) = 0$$

The fraction of vertices in the giant component is closely related to branching processes. Consider the following branching process. Start with a single vertex. The first node has Poisson(λ) children. Each child afterwards has Poisson(λ) children independently. It turns out that the probability of this process survives forever is given by the unique solution to the equation $\gamma = 1 - e^{\lambda\gamma}$ when $\lambda > 1$, and is 0 when $\lambda < 1$. So, it turns out that the probability for a randomly selected node to be in the giant component is exactly the survival probability of a branching process with Poisson(λ) as the offspring distribution.

2 Stochastic Block Model

While the Erdös-Rényi Model is very simple, it is not often a good model for real world networks. One of the reasons is that each vertex is identical (in distribution) to every other vertex. One may look for a model in which different subsets of the vertices are treated differently.

Recall that a partition of n is a collection of subsets $\mathcal{C} = \{C_1, ..., C_k\}$ such that all of the C_i are disjoint and $\cup_{i=1}^k C_i = \{1, ..., n\}$.

Definition 2.1. *The Stochastic Block Model on n vertices with communities $\mathcal{C} = \{C_1, ..., C_k\}$ and edge density matrix P, denoted $\mathcal{G}(\mathcal{C}, P)$, is a random graph model defined by the following properties:*

1. *$\mathcal{C} = \{C_1, ..., C_k\}$ is a partition of $\{1, ..., n\}$.*
2. *P is a $k \times k$ symmetric matrix for each each entry $P_{ij} = p_{ij} \in [0, 1]$.*
3. *Between two vertices $u, v \in C_i$ in the same community C_i, an edge exists with probability p_{ii} independently from all other edges.*
4. *Between two vertices in different communities $u \in C_i$, $v \in C_j$, $i \neq j$, an edge exists independently from all other edges with probability p_{ij}.*

Usually, we think about the off-diagonal entries of P to be much smaller than the diagonal entries. In words, this is because we expect there to be more connections within a group than there are connections between different groups.

The most well studied question in the Stochastic Block Model is the question of *community detection*. Heuristically, this means that given a sample from the SBM, we would like to determine the communities (that is, the partition $\mathcal{C}$ in the definition). There are different definitions of community detection, and we won't define any of them. However, it is natural to question how one might detect communities. An interested reader can check the book [1].

Let us consider the simplest case where n is an even integer, and the partition splits the set $\{1, ..., n\}$ perfectly in half. Within communities, an edge is present with probability p. Between communities, an edge is present with probability $q << p$. (Note that if $p = q$, this is exactly an Erdös-Rényi model.

A natural way to detect the communities is to do some sort of clustering. If we use spectral clustering, recovery is possible under the following condition:

$$\min(q, \frac{p-q}{2}) >> \frac{1}{\sqrt{n}}$$

For specific definitions of community detection, one can read chapters 4,5, and 6 of [1].

3 Preferential Attachment (Barabási-Albert)

This is the first random directed graph model we will look at. SBM and Erdös-Rényi are incapable of modeling some phenomena, even with correct choices of edge weights. The Preferential Attachment model was originally formulated as a model for citations in academic papers, but can also be applied to website link counts. These systems display the "rich get richer" phenomena. Papers with higher citations are the most likely to get a new citation.

The Barabási-Albert model is defined by the law of a random graph given by the following generation procedure.

Definition 3.1. *Fix a probability $p \in [0,1]$, and a number of nodes $n \geq 1$. A random graph will be called a Barabási-Albert (or BA) random graph if it can be generated by the following procedure.*

1. *Start with an initial vertex v_1.*
2. *At any next step $2 \leq m \leq n$, pick a vertex v_i with $i < m$ uniformly at random. With probability p, link a new vertex v_m to v_i. With probability $1-p$, attach vertex v_m to the vertex that v_i links to.*

The degree distribution of an Erdös-Rényi Graph looks approximately gaussian (this is essentially a consequence of the central limit theorem). Gaussian distributions yield outliers very rarely. By introducing the "rich-get-richer" dynamic, we expect it to be more likely for there to be vertices with high degree. This is formalized in the following claim.

Claim 3.2. *Let $c = 1 + (1-p)^{-1} \geq 2$. Also, define the degree distribution as $\tilde{p}(k) = \frac{1}{n}|\{v \in \{1,...,n\} : \textit{in-degree}(v) = k\}|$. Then,*

$$\tilde{p}(k) \sim \frac{1}{k^c}$$

The above means that $\lim_{k\to\infty} \tilde{p}(k)k^c = 1$.

These tails are called "power-law" tails with parameter c. This is a much slower decay to 0 than gaussian tails.Real world networks often display power-law tails, specifically with $c = 3$. Thus, $p = \frac{1}{2}$ is a very common choice when generating a BA model.

4 Configuration Models

The configuration model is a way to generate a random graph with an arbitrary degree distribution.

Definition 4.1. *Fix a number of vertices n and a degree sequence $(k_1,...,k_n)$. Note that whatever graph we produce must have $\frac{1}{2}\sum_{i=1}^n k_i$ edges (so we must have that $\sum_{i=1}^n k_i$ is an even number). To generate a graph according to the configuration model, do the following:*

1. *For each vertex $i = 1,...,n$ add k_i "half-edges" (these are also called "stubs").*
2. *Choose two (distinct) half-edges uniformly at random, and connect them to form an edge. Then, among the $\sum_{i=1}^n k_i - 2$ remaining half-edges, choose another pair uniformly at random and connect them.*
3. *If the resulting graph is simple (that is, does not have self loops, and has no multi-edges), then the resulting graph is said to be from the configuration model.*

One concern with the configuration model is simulating it. One may be concerned that for large graphs and common degree distributions, the probability that the resulting graph is simple goes to 0. This turns

out not to be the case, and one only expects to have to go through steps 1 and 2 a finite number of times (independent of n) to get a simple graph.

Configuration models are very helpful for testing various conjectures about graphs. Specifcally, one may want to know whether a certain property of a particular graph is "special" to that graph or if it only relies on the degree sequence. To test this, one can just generate configuration models with the same degree sequence and look for the emergence of said property.

5 Watts-Strogatz

We start with the 1-dimensional Watts-Strogatz Model

Definition 5.1. *The Watts Strogatz Model on n vertices with parameters $\beta \in [0,1]$ and k neighbors (in one dimension) is an undirected random graph generated by the following procedure:*

1. *First, start with a regular ring lattice. That is an undirected graph on n vertices where each vertex i is connected to vertex $i+1$ if $i \leq n$, and vertex n is connected to vertex 1.*
2. *Next, for each vertex i, connect i to all of the vertices j whose graph distance is less than or equal to $\frac{k}{2}$. The resulting graph is called Ring(n,k).*
3. *Finally, for every edge (i,j), we perform the following procedure. With probability $1-\beta$, we keep the edge as it is. With probability β, we choose either i or j with probability $\frac{1}{2}$, and choose a random vertex to connect either i or j to. All randomness is independent from the other randomness.*

One can also have this model in multiple dimensions. In two dimensions, you start with an $n \times n$ square grid in two dimensions. Then, you connect the vertices a (graph distance) less than or equal to $\frac{k}{2}$.

The Watts-Strogatz model is useful for its "small-world" properties. Given two nodes, there exists a relatively short path between them with high probability. This is easy to argue heuristically. Since the graph starts with vertices connected to the other close vertices, highly connected clusters will maintain after the random reassignment. The random reassignment will then allow us to traverse between these highly connected clusters in a short time.

One deficiency of the Watts-Strogatz model is that it is not obvious how vertices would "communicate" with each other. Imagine each vertex is a person. One vertex needs to send a letter to another far away person. All they know is who they are connected to, and the location of the far away person. The *decentralized search* is performed by you choosing your neighbor that is closest to the far away person. This ends up taking about $\sqrt{n}$ steps. This contrasts with the real world phenomena of "six-degrees of separation".

We refine thw Watts-Strogatz model in the following definition.

Definition 5.2. *The refined Watts-Strogatz Model on n vertices with parameters $\beta \in [0,1]$ and k neighbors (in one dimension) is an undirected random graph generated by the following procedure:*

1. *Perform steps 1 and 2 from the original procedure to get Ring(n,k)*
2. *For every edge (i,j), we perform the following procedure. With probability $1-\beta$, we keep the edge as it is. With probability β, we choose either i or j with probability $\frac{1}{2}$, say i. Then, we sample another vertex v with probability proportional to $\frac{1}{d(i,v)^q}$ where q is a parameter and d denotes graph distance.*

The optimal q for the above model is $q=1$. In general, the optimal q is equal to the dimension of the graph. Under the refined Watts-Strogatz model, we have the following (slightly imprecise) theorem:

Theorem 5.3. *Suppose we have a refined Watts-Strogatz model with. Assume the following:*

1. *The model is on n nodes.*
2. *We have chosen q optimally.*
3. *k is small compared to n.*

Then, the expected steps it takes decentralized search to get from one vertex to a target vertex is of order $\log(n)^2$.

References

[1] Emmanuel Abbe. Community detection and stochastic block models, 2023.

[2] P Erdös and A Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[3] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2015.