Basic Properties of Erdös-Rényi model

Seminar "Theoretical Topics in Data Science"

Ravi Niure

February 12, 2021

RWTH Aachen University

Graphs are mathematical representations of networks in terms of nodes, corresponding to network elements, and edges, representing the relationship among any two nodes. Many networks in nature are large and complex. This complexity can be attributed to the spontaneous nature of the relation between any two nodes determined by many unpredictable circumstances [1]. This stochastic nature is modeled with random graphs, the topic of this paper. More specifically, this paper deals with the statistical properties of a certain kind of random graph the Erdös-Rényi model. This paper will explore, in detail, the two properties of random graphs: the existence of triangles and threshold for a graph with diameter two. The paper will conclude that the existence of triangles has a somewhat gentle transition at the threshold, $\frac{1}{n}$, as compared to the property of diameter two which has a sharp threshold at $\sqrt{\frac{2\ln n}{n}}$.

1 Erdös-Rényi Model

Definition 1. The $G(n,p)$ model is a graph-valued random variable with two parameters, $n \in \mathbb{N}$, the total number of vertices and $p \in [0,1]$, the probability of edge formation between any pair of vertices (v,w) [2].

The edge formation between any pair is statistically independent of all other edges. For large graphs, $n \to \infty$, $n-1$, the number of potential neighbours for any vertex v can be approximated as n, the total number of vertices. Edge probability, p, is usually the function of n such that $p = \frac{d}{n}$ for some constant d. d can be interpreted as an expected degree of a vertex.

Certain global properties of the graph emerge even with statistically independent edge formation. The emergence of these properties has a threshold w.r.t p. These properties, specifically existence of triangles and two degrees of seperation, and their thresholds are the subjects of study in this paper.

2 Additional Notations

In addition to the definition of the Erdös-Rényi Model, this paper will use some theorems which are summarized below.

Theorem 2. *(Markov's inequality) Let x be a non-negative random variable. Then for $\alpha > 0$,*

$$\text{Prob}(x \geq \alpha) \leq \frac{\text{E}(x)}{\alpha}.$$

Theorem 3. *(Chebychev's inequality) Let x be a random variable. Then for $c > 0$,*

$$\text{Prob}\left(|x - \text{E}(x)| \geq c\right) \leq \frac{\text{Var}(x)}{c^2}.$$

Theorem 4. *(Linearity of expectation) For any random variables $X_1, X_2, \ldots, X_n$ and constants $\alpha_1, \alpha_2, \ldots, \alpha_n$,*

$$\text{E}\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \text{E}\left(\alpha_i X_i\right).$$

3 Degree Distribution

The degree of each vertex is the sum of n Bernoulli independent random variables corresponding to the presence or absence of an edge. The result is a binomial distribution.

$$\text{Prob}(\text{vertex has degree } k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \approx \binom{n}{k} p^k (1-p)^{n-k}.$$

The binomial distribution falls off exponentially as one moves away from the mean. In other words, the degree distribution for a vertex of random graph $\text{G}(n,p)$ is tightly concentrated around its expected value.

Theorem 5. *(Degree distribution) Let v be a vertex of the random graph* $\text{G}(n,p)$*. Let α be a real number in* $(0, \sqrt{np})$*.*

$$\text{Prob}\left(|np - deg(v)| \geq \alpha\sqrt{np}\right) \leq 3e^{\frac{-\alpha^2}{8}}.$$

The limitations with $G(n,p)$ model is that in a real world the degree distribution of a vertex is not as tightly concentrated around its expected value. Real world graphs, in fact, usually have a degree distribution similar to power law that has a "heavy tail".

4 Phase Transitions

Many monotonically increasing properties go through structural changes as the edge probability passes some threshold value. Monotonically increasing property refers to those which do not get affected with the addition or subtraction of an edge.

Definition 6. For any monotonically increasing property, if there exists $p(n)$ such that:

- For any $p_1(n) \in o\left(p(n)\right)$, the graph $\mathrm{G}(n, p_1(n))$ almost surely does not have the property.
- For any $p_2(n) \in \omega\left(p(n)\right)$, the graph $\mathrm{G}(n, p_2(n))$ almost surely has the property.

Then a phase transition occurs and $p(n)$ is the *threshold.*

Definition 7. For threshold of $p(n)$, if there exists a constant c such that:

- For $c < 1$, the graph $\mathrm{G}(n, cp_1(n))$ almost surely does not have the property.
- For $c > 1$, the graph $\mathrm{G}(n, cp_2(n))$ almost surely has the property.

Then $p(n)$ is a *sharp threshold.*

To establish phase transitions, $x(n)$ is often used to count the occurence of an item in a graph. It should be noted that $x(n)$ is random and non-negative and hence expectation and variance can be calculated. The following two theorems, which are based on Markov's and Chebychev's inequality, are used to establish the existence or absence of the item, and hence the threshold of the property. For convenience, we will write $x(n)$ as simply x.

Theorem 8. *(First moment method) Let x be a non-negative random variable and* $\mathrm{Prob}\left(x \geq 1\right) \leq \mathrm{E}\left(x\right)$*. Then, as* $\mathrm{E}\left(x\right) \to 0$ *with* $n \to \infty$*,*

$$\mathrm{Prob}\left(x > 0\right) \to 0.$$

When the expected value of x tends towards infinity, there is no guarantee that a graph picked at random will have the item. The possibility of small fraction of graphs containing large number of items and remaining containing none is there. In such cases, we can use the second moment argument to precisely argue that a random sample will definitely contain the property.

Theorem 9. *(Second moment method) Let x be a random variable with* $\mathrm{E}\left(x\right) > 0$*. If* $\mathrm{Var}(x) = o(\mathrm{E}^2(x))$*, then x is almost surely greater than zero.*

Proof. If $\mathrm{E}(x) > 0$, then for x to be less than or equal to zero, it must differ from its expected value by at least its expected value. Thus,

$$\mathrm{Prob}(x \leq 0) \leq \mathrm{Prob}(|x - \mathrm{E}(x)| \geq \mathrm{E}(x))$$

By Chebyshev's Inequality,

$$\mathrm{Prob}(|x - \mathrm{E}(x)| \geq \mathrm{E}(x)) \leq \frac{\mathrm{Var}(x)}{\mathrm{E}^2(x)}$$

Thus, $\mathrm{Prob}(x \leq 0)$ goes to zero if $\mathrm{Var}(x)$ is $o(\mathrm{E}^2(x))$. □

5 Existence of Triangles

Lemma 10. *The expected number of triangles in a random graph* $\mathrm{G}\left(n, \frac{d}{n}\right)$ *is* $\frac{d^3}{6}$.

Proof. Let $\triangle_{ijk}$ be the indicator variable for the triangle with vertices i, j, and k being present. Additionally, let x be the number of triangles. Then,

$$x = \left(\sum_{ijk} \triangle_{ijk}\right)$$

The expectation calculation results in expectation of the sum of indicator variables. With linearity of expectation, the summation can be moved outside,

$$\mathrm{E}(x) = \mathrm{E}\left(\sum_{ijk} \triangle_{ijk}\right) = \sum_{ijk} \mathrm{E}\left(\triangle_{ijk}\right) = \binom{n}{3}\left(\frac{d}{n}\right)^3$$

$$= \frac{n!}{3!(n-3)!}\left(\frac{d}{n}\right)^3 = \frac{n(n-1)(n-2)(n-3)!}{3!(n-3)!}\left(\frac{d}{n}\right)^3 \approx \frac{n^3}{3!}\left(\frac{d}{n}\right)^3 = \frac{d^3}{6}.$$

□

However, if $\frac{1}{n}$ of the graphs have $\frac{d^3}{6}n$ triangles and the remaining graphs have no triangles, then as $n \to \infty$, the expecation would remain the same but the probability that a graph selected at random would have a triangle would go to zero resulting in high variance. Thus, we need to establish a relationship between expectation and variance of x using second moment method.

Lemma 11. *Let x be the number of triangles in* $\mathrm{G}(n, p)$, *then* $\mathrm{Var}(x) \leq \mathrm{E}(x) + o(1)$.

Proof. Lets start by calculating $\mathrm{E}(x^2)$. We can write x as

$$x = \left(\sum_{ijk} \triangle_{ijk}\right)$$

Expanding the square term,

$$\mathrm{E}(x^2) = \mathrm{E}\left(\sum_{i,j,k} \triangle_{ijk}\right)^2 = \mathrm{E}\left(\sum_{\substack{i,j,k \\ i',j',k'}} \triangle_{ijk}\triangle_{i',j',k'}\right).$$

Splitting the summation above leads to three parts. Part 1 contains set of i, j, k and i', j', k' that share at most one vertex as shown in the figure below. In this case, the $\triangle_{ijk}$ and $\triangle_{i',j',k'}$ are independent leading to

$$\mathrm{E}\left(\sum \triangle_{ijk}\triangle_{i',j',k'}\right) = \sum \mathrm{E}(\triangle_{ijk})\,\mathrm{E}(\triangle_{i'j'k'}) \leq \left(\sum_{ijk} \mathrm{E}(\triangle_{ijk})\right)\left(\sum_{i'j'k'} \mathrm{E}(\triangle_{i'j'k'})\right) = \mathrm{E}^2(x).$$

For part 2, i, j, k and i', j', k' share two vertices with one edge. Independence can not be assumed in this scenario. However, the problem can be reformulated as a "rectangle with a diagonal" problem as opposed to two triangles. As such, let $\boxslash_{abcd}$ be the indicator variable for the rectangle with vertices a, b, c, and d and either diagonal is present. Additionally, let y be the number of rectangles with diagonal.

$$\mathrm{Prob}\left(\boxslash_{abcd}\right) = \mathrm{Prob}\left(\square_{abcd}\right) \cdot \left(\mathrm{Prob}\left(edge_{ac}\right) + \mathrm{Prob}\left(edge_{bd}\right)\right) = \left(\frac{d}{n}\right)^4\left(\frac{d}{n} + \frac{d}{n}\right) = 2\left(\frac{d}{n}\right)^5.$$

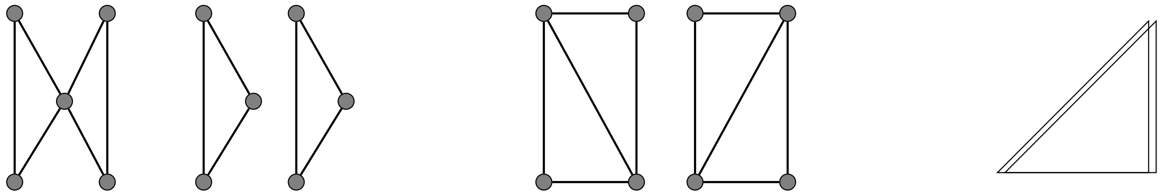

Figure 1: The triangles in Part 1, Part 2, and Part 3 of the second moment argument for existence of triangles in $\mathrm{G}(n, \frac{d}{n})$.

There are $\binom{n}{4}$ subsets with 4 vertices. As such, the expected number of such rectangles with either diagonal becomes,

$$\mathrm{E}(y) = \mathrm{E}\left(\sum_{abcd} \boxslash_{abcd}\right) = \sum_{abcd} (\mathrm{E}\boxslash_{abcd}) = 2\binom{n}{4}\left(\frac{d}{n}\right)^5 \approx \frac{d^5}{6n}$$

As n goes to ∞, the expected value of y approaches zero. In other words, there are so few rectangles and triangles in the graph, the probability of two triangles sharing an edge is extremely unlikely.

For part 3, i, j, k and i', j', k' are the same sets. The contribution of this part of the summation to $\mathrm{E}(x^2)$ is $\mathrm{E}(x)$. Thus, putting all the parts together,

$$\mathrm{E}(x^2) \leq \mathrm{E}^2(x) + \mathrm{E}(x) + o(1)$$

which implies,

$$\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}^2(x) \leq \mathrm{E}(x) + o(1)$$

□

Corollary 12. *(Threshold for triangles) The existence of triangles has a threshold at* $p(n) = \frac{1}{n}$.

Proof. Lemma 3. states that for number of triangles to be zero, it must differ from its expected value by at least its expected value. Thus,

$$\mathrm{Prob}(x = 0) \leq \mathrm{Prob}\left(|x - E(x)| \geq \mathrm{E}(x)\right)$$

By Chebychev inequality,

$$\mathrm{Prob}(x = 0) \leq \frac{\mathrm{Var}(x)}{\mathrm{E}^2(x)} \leq \frac{\mathrm{E}(x) + o(1)}{\mathrm{E}^2(x)} \leq \frac{6}{d^3} + o(1).$$

Thus, for $d > \sqrt[3]{6} \cong 1.8$, $\mathrm{Prob}(x = 0) < 1$ and $\mathrm{G}(n, p)$ has a triangle with non-zero probability. For $d < 1.8, \mathrm{E}(x) = \frac{d^3}{6} < 1$ and there are not enough edges in the graph for there to be a triangle. □

6 Graph with diameter two

Definition 13. A graph has diameter two if and only if for each pair of vertices i and j, either there is an edge between them or there is another vertex k to which both i and j have an edge.

To arrive at the threshold for the property, the number of "bad" pairs i and j that fail to satisfy the definition are counted and are analyzed.

Theorem 14 (Threshold for diameter two). *The property that* $\mathrm{G}(n,p)$ *has diameter two has a sharp threshold at* $p = \sqrt{2}\sqrt{\frac{\ln n}{n}}$.

Proof. Assuming G has a diameter greater than two, let $I_{ij} \in \{0,1\}$ indicate bad pair of vertices i and j. Then the number of bad pair of vertices x is equal to $\sum_{i<j} I_{ij}$. A graph has diameter two if and only if it has no bad pair, i.e, $x = 0$. Thus, if $\lim_{n\to\infty} \mathrm{E}(x) = 0$, then for large n, almost surely, a graph has no bad pair and hence has the diameter at most two.

Since p is the probability of an edge being present between i and j, and p^2 is the probability for two edges from i and j to some arbitrary vertex,

$$\mathrm{Prob}\left(I_{ij} = 1\right) = \mathrm{Prob}(i \text{ and } j \text{ are not adjacent}) \cdot \mathrm{Prob}(\text{no other vertex is adjacent to } i \text{ and } j)$$

$$= (1-p)\left(1-p^2\right)^{n-2}$$

There are $\binom{n}{2}$ pairs of vertices and thus,

$$\mathrm{E}(x) = \binom{n}{2}(1-p)\left(1-p^2\right)^{n-2}$$

Let $p = \sqrt{2}\sqrt{\frac{\ln n}{n}}$ and for large n,

$$\mathrm{E}(x) \cong \frac{n^2}{2}\left(1 - c\sqrt{\frac{\ln n}{n}}\right)\left(1 - c^2\frac{\ln n}{n}\right)^n \cong \frac{n^2}{2}e^{-c^2 \ln n} \cong \frac{n^2}{2}\left(e^{\ln n}\right)^{-c^2} \cong \frac{1}{2}(n)^{2-c^2}.$$

For $c > \sqrt{2}, \mathrm{E}(x)$ is $o(1)$. The first moment method says that for $p = c\sqrt{\frac{\ln n}{n}}$ with $c > \sqrt{2}, \mathrm{G}(n,p)$ has no bad pair and has diameter at most two. For $c < \sqrt{2}, \mathrm{E}(x)$ is $\omega(1)$. The second moment method will be used to prove that the graph definitely has a bad pair and hence has a diameter greater than two.

$$\mathrm{E}(x^2) = \mathrm{E}\left(\sum_{i<j} I_{ij}\right)^2 = \mathrm{E}\left(\sum_{i<j} I_{ij} \sum_{k<l} I_{kl}\right) = \mathrm{E}\left(\sum_{\substack{i<j\\k<l}} I_{ij}I_{kl}\right) = \sum_{\substack{i<j\\k<l}} \mathrm{E}\left(I_{ij}I_{kl}\right).$$

Depending on number of distinct indices, we can partition the above sum into three parts.

$$\mathrm{E}(x^2) = \sum_{\substack{i<j\\k<l\\a=4}} \mathrm{E}\left(I_{ij}I_{kl}\right) + \sum_{\substack{\{i,j,k\}\\i<j\\a=3}} \mathrm{E}\left(I_{ij}I_{ik}\right) + \sum_{\substack{i<j\\a=2}} \mathrm{E}\left(I_{ij}^2\right).$$

For the first part where $a = 4$, all i, j, k, and l are distinct. If $I_{jk}I_{kl} = 1$, then both pairs (i,j) and (k,l) are bad. For both pairs to be bad, for any vertex $u \notin \{i,j,k,l\}$ possible edge between u and others do not exist. The probability of such scenario is $(1-p^2)^2$. Considering all $n-4$ vertices not in $\{i,j,k,l\}$, the events are independent. Thus,

$$\mathrm{E}(I_{ij}I_{kl}) \leq \left(1-p^2\right)^{2(n-4)} \leq \left(1 - c^2\frac{\ln n}{n}\right)^{2n}(1+o(1)) \leq n^{-2c^2}(1+o(1))$$

and the first sum is,

$$\sum_{\substack{i<j\\k<l}} \mathrm{E}\left(I_{ij}I_{kl}\right) \leq \frac{1}{4}n^{4-2c^2}(1+o(1))$$

where, the $\frac{1}{4}$ comes from the fact that only a fourth of the 4-tuples (i, j, k, l) satisfy $i < j$ and $k < l$.

For the second summation, if $I_{ij}I_{ik} = 1$, then for every vertex $u \notin \{i, j, k\}$, either there is no edge between i and u or there is an edge (i, u) and both edges (j, u) and (k, u) are absent. The probability of this event for one u is,

$$1 - p + p(1 - p)^2 = 1 - 2p^2 + p^3 \approx 1 - 2p^2.$$

Thus, for all such u and substituting p,

$$\left(1 - \frac{2c^2 \ln n}{n}\right)^{n-3} \cong e^{-2c^2 \ln n} = n^{-2c^2},$$

For all distinct triples,

$$\sum_{\substack{\{i,j,k\}\\ i<j}} \mathrm{E}\left(I_{ij}I_{ik}\right) = n^{3-2c^2}.$$

For the third summation, since $I_{ij} \in \{0, 1\}, \mathrm{E}(I_{ij}^2) = \mathrm{E}(I_{ij})$. Thus,

$$\sum_{ij} \mathrm{E}(I_{ij}^2) = \mathrm{E}(x).$$

Combining all three parts,

$$\mathrm{E}(x^2) \le \frac{1}{4}n^{4-2c^2} + n^{3-2c^2} + n^{2-c^2}$$

While also, $\mathrm{E}(x) \cong \frac{1}{2}n^{2-c^2}$. From the above expression, it can be observed that for $c < \sqrt{2}, \mathrm{E}(x^2) \le \mathrm{E}^2(x)\,(1 + o(1))$. This gives, $\mathrm{E}(x^2) - \mathrm{E}^2(x) \le \mathrm{E}^2(x)o(1)$. Since, $\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}^2(x)$,

$$\mathrm{Var}(x) \le \mathrm{E}^2(x)o(1) \implies \mathrm{Var}(x) = o(\mathrm{E}^2(x)).$$

As such, according to *Theorem 9*, there is at least one bad pair of vertices resulting in diameter greater than two. Therefore, the property has *sharp threshold* at $p = \sqrt{2}\sqrt{\frac{\ln n}{n}}$. □

7 Conclusion

In conclusion, although the Erdös-Rényi model or the $\mathrm{G}(n, p)$ model assumes independent edge formation and has a binomial degree distribution, not observed in many real-world graphs, it is one of the most influential and mathematically important random graph models. Certain global properties appear and disappear with different edge probability in the $G(n, p)$ model even with the independent edge formation assumption. Properties such as the existence of triangles and the two degrees of separation appear for the edge probabilities higher than the thresholds at $\frac{1}{n}$ and $\sqrt{\frac{2\ln n}{n}}$ respectively. These thresholds, not only for the properties discussed here but also for additional properties such as the disappearance of isolated vertices, can be established with the help of first moment and second-moment methods which are based on concentration inequalities.

References

[1] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks.* Cambridge University Press, New York, USA, 1st edition, 2008.

[2] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science.* Cambridge University Press, Cambridge, UK, pre-publication edition, 2020.