Basic Properties of Erdös Renýi Model

Ravi Niure

Seminar "Theoretical Topics in Data Science"

Supervisor: León Bohn

RWTH Aachen University

Email: *ravi.niure@mail.rwth-aachen.de*

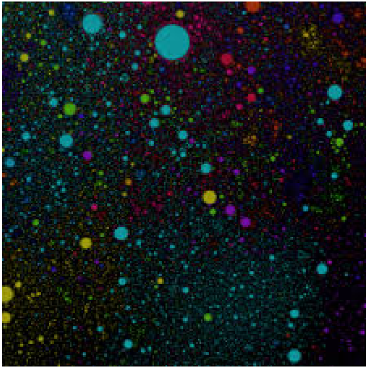
February 5, 2021

Overview

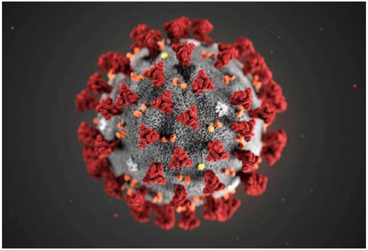
Motivation

Networks are everywhere!

- COVID-19, World Wide Web, The Internet, Transportation networks, Gene interactions, Social networks, scientific collaborations, etc.



(a) Map of Internet
Source: https://internet-map.net



(b) SARS-CoV-2 virus

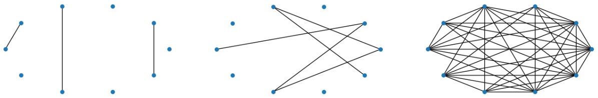
Study of Networks

- **Mathematical representation:** Graphs with several *nodes* or *vertices* and *edges* or *links*
- Real networks are large and complex.
- Complete description of such networks and graphs is almost impossible.
- Statistical approach: link between microscopic properties and macroscopic phenomena!

This leads to **Random Graphs!**

Random Graphs

- A **random graph** is a result of **random process**.
- **Random process:** Connection between two elements is a random event determined by the sum of a very large number of unpredictable events. [2]
- A simple random graph model is **Erdös-Rényi Model.**



(a) p=0.05 (b) p = 0.1 (c) p = 0.9

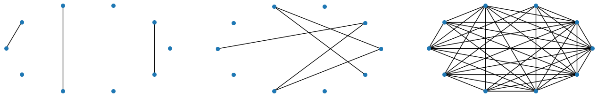
Figure: Graphs with N = 10 with different edge probability

Definition

Erdös-Rényi Model

Definition

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).[3]



(a) p=0.05 (b) p = 0.1 (c) p = 0.9

Figure: Graphs with N = 10 with different edge probability

Erdös-Rényi Model cont'd

- **Potential neighbours**

$$n - 1 \cong n \text{ for } n \to \infty$$

- **Edge Probability** as a function of n

$$p = \frac{d}{n}, \text{ where } d \text{ is some constant in many cases.}$$

- **Expected Degree**

$$E[degree] = (n-1)\frac{d}{n} \approx n\frac{d}{n} = d$$

Additional Concepts

Linearity of Expectation

Theorem (Linearity of expectation)

For any random variables $X_1, X_2, \ldots, X_n$ *and constants* $\alpha_1, \alpha_2, \ldots, \alpha_n$*,*

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

Markov's Inequality

Theorem (Markov's inequality)

Let x be a non-negative random variable. Then for $\alpha > 0$,

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

If $\alpha = 1$, then,

$$\text{Prob}(x \geq 1) \leq \mathrm{E}(x).$$

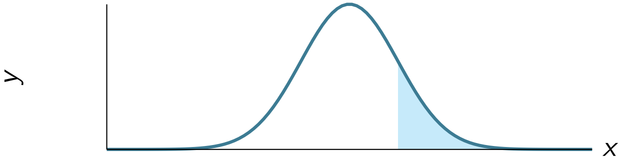


Figure: Markov's Inequality

Theorem (Chebychev's inequality)

Let x be a random variable. Then for $c > 0$,

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

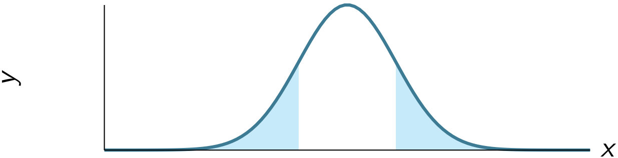


Figure: Chebychev's Inequality

Degree Distribution

Degree Distribution

- The degree distribution of $G(n,p)$ is **binomial**.

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

$$\approx \binom{n}{k} p^k (1-p)^{n-k}$$

- As $n \to \infty$, the **binomial** distribution approaches **poisson** distribution

$$\lim_{n\to\infty} \binom{n}{k} \left(\frac{d}{n}\right)^k \left(1-\frac{d}{n}\right)^{n-k} = \frac{n^k}{k!}\frac{d^k}{n^k} e^{-d} = \frac{d^k}{k!} e^{-d}$$

for $p = \frac{d}{n}$ and constant d.

Degree Distribution in Real Networks

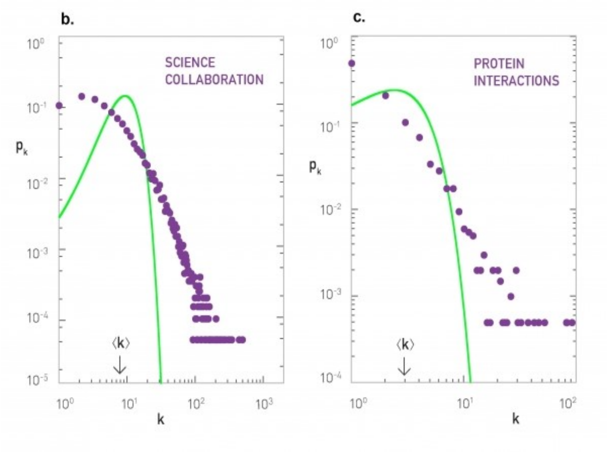


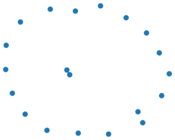
Figure: Degree Distribution in various networks and Poisson approximation [1]

Limitations of ER Model

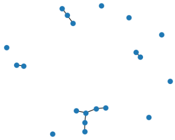
- Degree distribution is concentrated around expected degree.
- Independent edge formation doesn't reflect real networks.

Phase Transitions

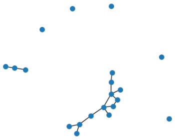
Phase Transitions



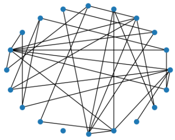
(a) p=0.01



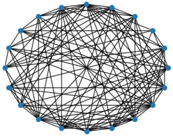
(b) $p = 0.05$



(c) $p = 0.06$



(d) $p = 0.15$



(e) $p = 0.55$

Figure: Graphs with $N = 20$

Threshold

Definition

For any monotonically increasing property, if there exists a function $p(n)$ such that:

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

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

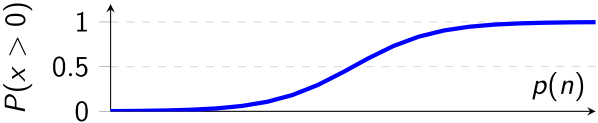


Figure: Threshold

Sharp Threshold

Definition

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

- for $c < 1$, the graph $G(n, cp(n))$ almost surely does not have the property.
- for $c > 1$, the graph $G(n, cp(n))$ almost surely has the property.

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

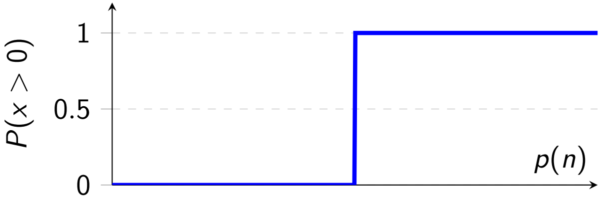


Figure: Sharp Threshold

Moment Methods

Theorem (First Moment Method)

Let $x \in \mathbb{Z}_{\geq 0}$ be a random variable representing the number of occurrence of a certain property. Then, as $\mathrm{E}(x)$ approaches 0, with n approaching infinity, the probability of the existence of the property tends to 0.

Theorem (Second Moment Method)

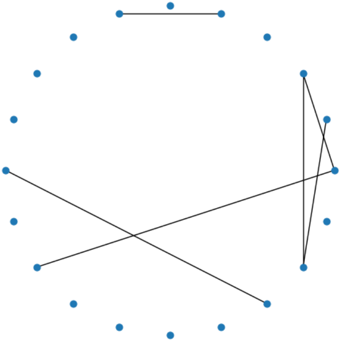
Let $x \in \mathbb{Z}_{\geq 0}$ be a random variable representing the number of occurrence of a certain property, and $E(x) > 0$. If $Var(x) = o(\mathrm{E}^2(x))$, then x is almost surely greater than zero.

Corollary

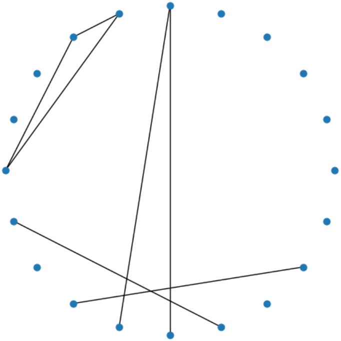
If $\mathrm{E}(x^2) \leq \mathrm{E}^2(x)(1 + o(1))$, then x is almost surely greater than zero.

Existence of Triangles

Triangles



(a) Graph without a triangle



(b) Graph with a triangle

Expectation calculation

Lemma

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

Proof:

Let x be the number of triangles. Then x is given by,
$x = \sum_{ijk} \triangle_{ijk}$, where $\triangle_{ijk}$ is an indicator variable.

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

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

Introvert vs Extrovert Rooms

Figure: Introverts and Extroverts
Source: https://www.asapscience.com/

Variance Estimation

Lemma

Let x be the number of triangles in $G(n, \frac{d}{n})$, then $Var(x) \leq E(x) + o(1)$.

Proof:

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

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

Expanding the squares of sum,

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

Proof - cont'd

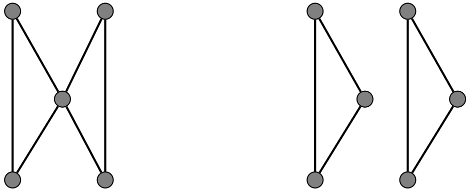


Figure: Triangles share 1 or less vertex

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

Note: The independence of the triangles and linearity of expectation!

Proof - cont'd

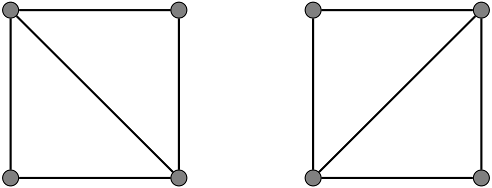


Figure: Triangles share 2 vertices

Note: We can't make independence assumption with the triangles. However, we can reformulate the problem as "rectangle with a diagonal" as opposed to two triangles.

Proof - cont'd

let y be the number of rectangles with diagonal.

$$\text{Prob}\left(\boxslash_{abcd}\right) = \left(\frac{d}{n}\right)^4 \left(\frac{d}{n} + \frac{d}{n}\right) - \left(\frac{d}{n}\right)^6 \approx 2\left(\frac{d}{n}\right)^5$$

The expected number of such rectangles with either diagonal becomes,

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

As n approaches ∞, this expectation will approach zero, i.e $o(1)$.

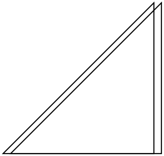


Figure: Two triangles sharing 3 vertices

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 $\mathrm{E}(x)$.

Proof - cont'd

Thus, putting all the parts together,

$$\mathrm{E}(x^2) = \mathrm{E}\left(\sum_{\substack{i,j,k \\ i',j',k'}} \triangle_{ijk}\triangle_{i',j',k'}\right) \leq \mathrm{E}^2(x) + \mathrm{E}(x) + o(1)$$

which implies,

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

Threshold for Existence of Triangles

Corollary (Threshold for triangles)

The threshold for the existence of triangles in $\mathsf{G}(n, \frac{d}{n})$ *is* $p(n) = \frac{1}{n}$.

Proof. For x to be zero, x would have to differ from its expected value by at least the expected value. Thus,

$$\text{Prob}(x = 0) \leq \text{Prob}\left(|x - E(x)| \geq \mathrm{E}(x)\right) \leq \frac{\mathrm{Var}(x)}{\mathrm{E}^2(x)} \leq \frac{\mathrm{E}(x) + o(1)}{\mathrm{E}^2(x)}$$

$$\leq \frac{6}{d^3} + o(1)$$

For $d > \sqrt[3]{6} \cong 1.8$, $\text{Prob}(x = 0) < 1$
For $d < 1.8, \mathrm{E}(x) = \frac{d^3}{6} < 1$
Therefore, $p = \frac{d}{n} \approx \frac{1}{n}$ is the threshold.

Simulation with networkx

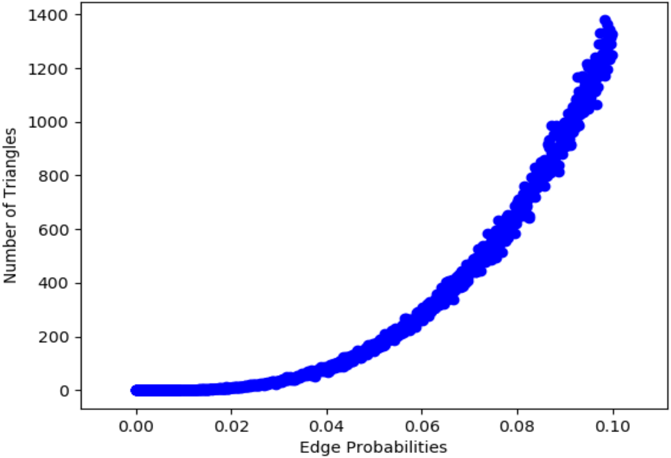


Figure: Number of Triangles when N = 200

Threshold for existence of triangles when N = 200

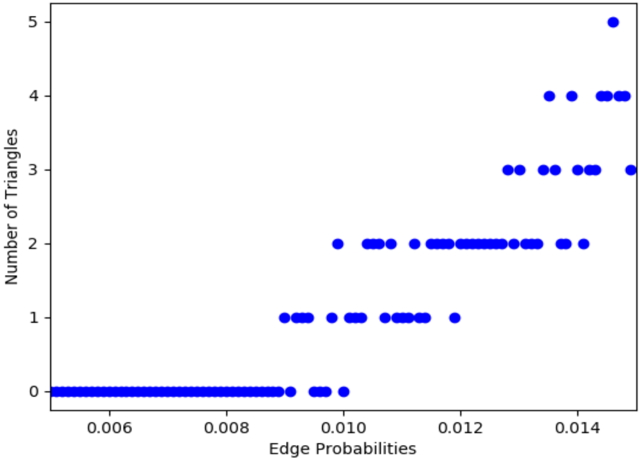
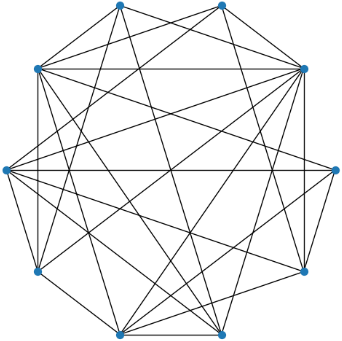


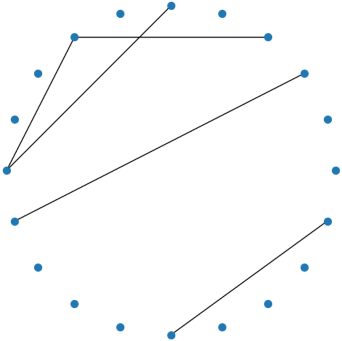
Figure: Number of Triangles when N = 200

Threshold for Diameter of two

Different diameter graphs



(a) Graph with diameter 2



(b) Graph with undefined diameter

Definition of diameter two

Definition

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.

Good Pair Bad Pair!

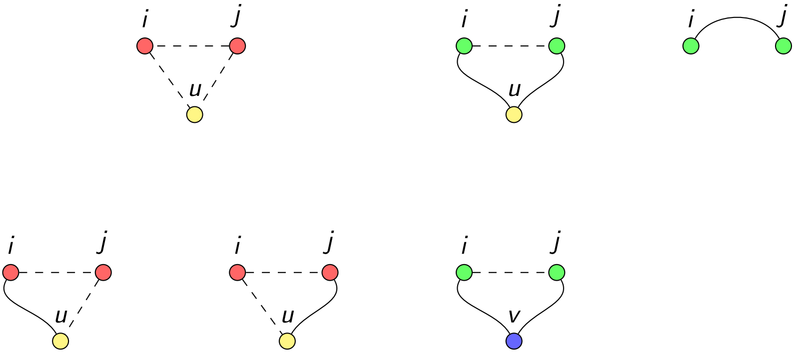


Figure: Good Pair Bad Pair

Threshold for diameter two

Theorem (Threshold for diameter two)

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

- $\frac{1}{\sqrt{n}}$: threshold for finding a common neighbor
- $\sqrt{\ln n}$: ensures every pair has a common neighbor

Proof - First moment argument

Number of bad pair of vertices $(x) = \sum_{i<j} I_{ij}$

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

Let $p = \sqrt{c}\sqrt{\frac{lnn}{n}}$ and for large n,

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

Expectation

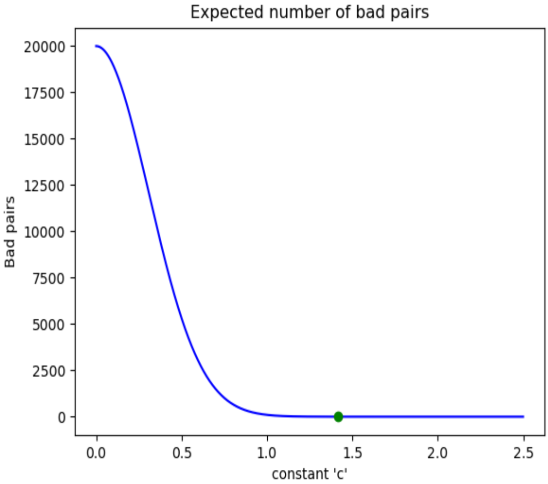


Figure: Expected bad pairs

For $c < \sqrt{2}$, Second moment argument

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

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

The above sum can be partitioned into:

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

Second moment argument - cont'd

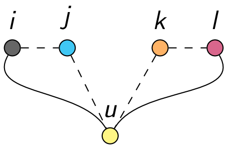


Figure: Four distinct vertices

For the first part,

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

Across all $\{i, j, k, l\}$,

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

Second moment argument - cont'd

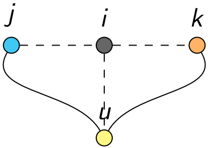


Figure: Three distinct vertices

For the second part of the sum,

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

Second moment argument - cont'd

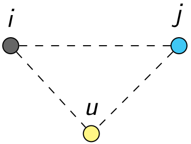


Figure: Three distinct vertices

For the third summation,

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

Combining all three parts,

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

Second moment argument - cont'd

For $c < \sqrt{2}$,

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

This gives,

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

Thus,

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

As such, from second moment argument, a bad pair almost surely exists and therefore there is no graph with diameter ≤ 2.

Simulation with networkx

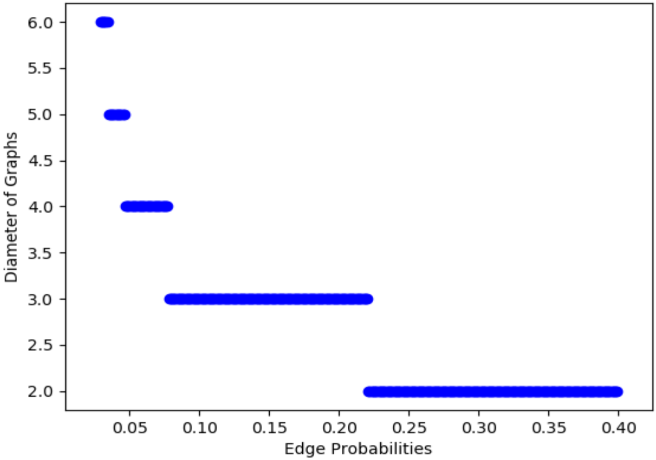


Figure: Diameter of Graphs with different probability when $N = 200$

Conclusion

Summary

- Random Graphs provide essential frameworks for studying large and complex networks.
- Erdös-Rényi model is the earliest and simple random graph model.
- **Limitations:** Thin tail, independent edge formation, etc
- Concentration inequalities provide expectation and variance relationship; used to argue for/against existence of properties.
- Triangles are essential to compute clustering coefficients.
  - Expected number of triangles, independent of nodes, is $\frac{d^3}{6}$.
  - The sharp threshold for existence is $\frac{1}{n}$.
- The threshold for two degrees of separation is $\sqrt{2\frac{\ln n}{n}}$.

References

Albert Barabási. *Random Networks Network Science*. URL: http://networksciencebook.com/chapter/3 (visited on 02/05/2021).

Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. 1st. New York, USA: Cambridge University Press, 2008. ISBN: 9780521879507.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. pre-publication. Cambridge, UK: Cambridge University Press, 2020. ISBN: 9781108485067.