Typical Distance in Erdős–Rényi Binomial Random Graph and Lattice Random Graph Model: A Simulation Study

Sayak Chatterjee

A joint project with Aditya Ghosh
under the supervision of
Prof. Antar Bandyopadhyay

June 18, 2020

Overview

Introducing Typical and Average Distance

- Let us consider the Erdős–Rényi binomial random graph model and denote it by $\mathrm{ER}(n, p)$. We take a realisation $G \sim \mathrm{ER}(n, p)$. Now, the object of our study is the typical distance which is defined by the graph distance of two randomly selected vertices in G. We will denote the typical distance by H_n.

Introducing Typical and Average Distance

- Let us consider the Erdős–Rényi binomial random graph model and denote it by $\mathrm{ER}(n, p)$. We take a realisation $G \sim \mathrm{ER}(n, p)$. Now, the object of our study is the typical distance which is defined by the graph distance of two randomly selected vertices in G. We will denote the typical distance by H_n.
- The average distance in G is the average of all graph distances $d(u, v)$ where u and v are two vertices of G belonging to the same connected component of G. So the average distance is just $\mathrm{E}(H_n | H_n \text{ is finite})$. We have the following result.

Introducing Typical and Average Distance

- Let us consider the Erdős–Rényi binomial random graph model and denote it by $\mathrm{ER}(n, p)$. We take a realisation $G \sim \mathrm{ER}(n, p)$. Now, the object of our study is the typical distance which is defined by the graph distance of two randomly selected vertices in G. We will denote the typical distance by H_n.
- The average distance in G is the average of all graph distances $d(u, v)$ where u and v are two vertices of G belonging to the same connected component of G. So the average distance is just $\mathrm{E}(H_n | H_n \text{ is finite})$. We have the following result.

Theorem (Chung and Lu., 2002)

If $np \geq c > 1$ for some constant c, then the average distance in $G \sim ER(n, p)$ is almost surely $(1 + o(1))(\log n / \log np)$ provided $(\log n / \log np) \to \infty$ as $n \to \infty$.

Introducing Typical and Average Distance

- Let us consider the Erdős–Rényi binomial random graph model and denote it by $\mathrm{ER}(n, p)$. We take a realisation $G \sim \mathrm{ER}(n, p)$. Now, the object of our study is the typical distance which is defined by the graph distance of two randomly selected vertices in G. We will denote the typical distance by H_n.
- The average distance in G is the average of all graph distances $d(u, v)$ where u and v are two vertices of G belonging to the same connected component of G. So the average distance is just $\mathrm{E}(H_n | H_n \text{ is finite})$. We have the following result.

Theorem (Chung and Lu., 2002)

If $np \geq c > 1$ for some constant c, then the average distance in $G \sim ER(n, p)$ is almost surely $(1 + o(1))(\log n / \log np)$ provided $(\log n / \log np) \to \infty$ as $n \to \infty$.

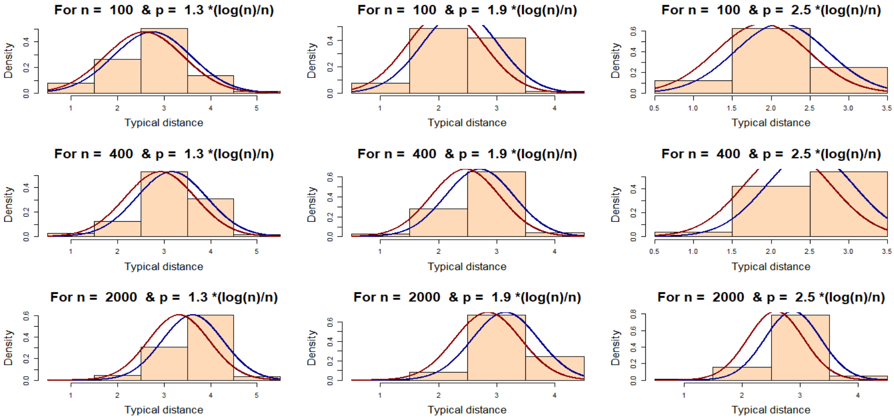
- It is well known that $G \sim \mathrm{ER(n,p)}$ is asymptotically almost surely connected if $p = c * (\log n/n)$ where $c > 1$. Here we have considered that specific case (connectivity regime).

Connectivity Regime: Histograms

We simulated random graphs from connectivity regime 1000 times varying $c = 1.1,\ 1.3,\ \ldots,\ 2.5$ and $n = 20, 60, 100, 150, 250, 400, 675, 1000, 2000$ so that $\log n$ increased more or less linearly. We considered the typical distances for each simulated graph when it was finite. Here we present some of the histograms obtained.

Connectivity Regime: Histograms

We simulated random graphs from connectivity regime 1000 times varying $c = 1.1, 1.3, \dots, 2.5$ and $n = 20, 60, 100, 150, 250, 400, 675, 1000, 2000$ so that $\log n$ increased more or less linearly. We considered the typical distances for each simulated graph when it was finite. Here we present some of the histograms obtained.



Connectivity Regime: The o(1) Term

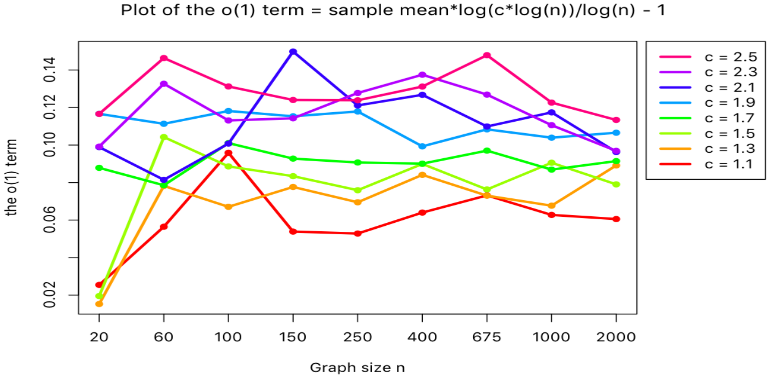
We observe that in the histograms, the red curve (with $\log n / \log np$ mean) is always behind the blue curve (with the sample mean). So, we take a closer look at the $o(1)$ term referred in the theorem. To do that, we plot the following quantity against n:

$$\frac{\text{sample mean}}{\log(n)/\log(np)} - 1.$$

Connectivity Regime: The o(1) Term

We observe that in the histograms, the red curve (with $\log n / \log np$ mean) is always behind the blue curve (with the sample mean). So, we take a closer look at the $o(1)$ term referred in the theorem. To do that, we plot the following quantity against n:

$$\frac{\text{sample mean}}{\log(n)/\log(np)} - 1.$$

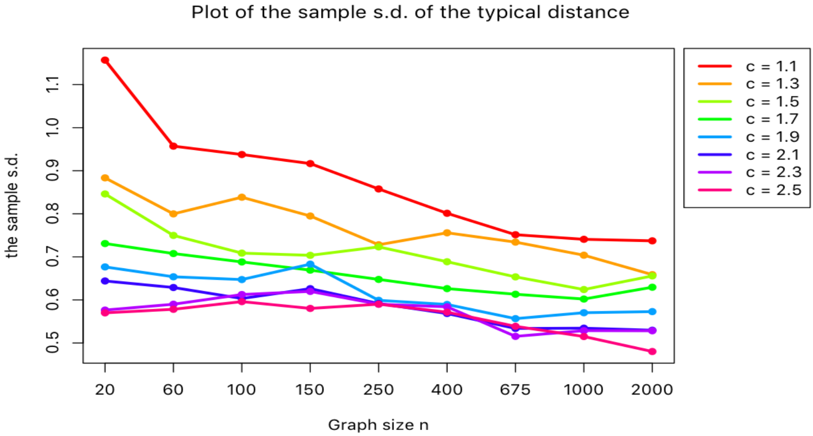


Connectivity Regime: The Sample S.D.

Here we present the plot of the sample standard deviations of typical distance (when it is finite) for different values of c and n.

Connectivity Regime: The Sample S.D.

Here we present the plot of the sample standard deviations of typical distance (when it is finite) for different values of c and n.



We can observe an overall decreasing pattern in the sample s.d.

Connectivity Regime: Testing Normality and Symmetry

- Before performing any tests of normality, first we standardized the samples using sample mean and sample s.d. and looked at the Q-Q plot. To break ties, we jittered the data by adding random noise from normal distribution with zero mean and small variance.

Connectivity Regime: Testing Normality and Symmetry

- Before performing any tests of normality, first we standardized the samples using sample mean and sample s.d. and looked at the Q-Q plot. To break ties, we jittered the data by adding random noise from normal distribution with zero mean and small variance.
- Then we plotted the standardized histograms and performed Pearson's χ^2 goodness of fit test, Kolmogorov-Smirnov test and Shapiro-Wilk test for normality on standardized data. The p-values we got were very close to zero except for few cases for Kolmogorov-Smirnov test. The histograms were also discrete in nature and did not come any close to the normal distribution.

Connectivity Regime: Testing Normality and Symmetry

- Before performing any tests of normality, first we standardized the samples using sample mean and sample s.d. and looked at the Q-Q plot. To break ties, we jittered the data by adding random noise from normal distribution with zero mean and small variance.
- Then we plotted the standardized histograms and performed Pearson's χ^2 goodness of fit test, Kolmogorov-Smirnov test and Shapiro-Wilk test for normality on standardized data. The p-values we got were very close to zero except for few cases for Kolmogorov-Smirnov test. The histograms were also discrete in nature and did not come any close to the normal distribution.
- After normality got rejected in most of the times, we went for testing symmetry by performing the Randles-Fligner-Policello-Wolfe (RFPW) non-parametric test of symmetry. When we tabulated the p-values, we found zero p-values lying around the main diagonal of the table. There were some very high p-values scattered in the table. For example, when $n = 250$, $c = 2.5$, we got a p-value of 0.8907. Looking closely, we found H_n taking only 3 values namely 1, 2 and 3 where frequency of 1 was negligible compared to 2 and 3.

Definition and Illustration

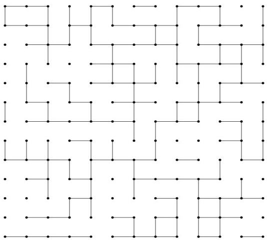
- Consider all the lattice points (x, y) (with integer coordinates) on the Cartesian plane such that $-n \leq x \leq n$ and $-n \leq y \leq n$. We consider these points as vertices and join pairs of vertices with an edge that are unit distance apart. We will get a grid like structure. Let us call this graph as the complete lattice.

Definition and Illustration

- Consider all the lattice points (x, y) (with integer coordinates) on the Cartesian plane such that $-n \leq x \leq n$ and $-n \leq y \leq n$. We consider these points as vertices and join pairs of vertices with an edge that are unit distance apart. We will get a grid like structure. Let us call this graph as the complete lattice.
- For each edge in the complete lattice, we draw a $\text{Bernoulli}(p)$ random variable. If it is 1, we keep the edge and delete it otherwise. In this way we can generate a random graph. We denote this model of random graph as $\text{Lat}(n, p)$.

Definition and Illustration

- Consider all the lattice points (x, y) (with integer coordinates) on the Cartesian plane such that $-n \leq x \leq n$ and $-n \leq y \leq n$. We consider these points as vertices and join pairs of vertices with an edge that are unit distance apart. We will get a grid like structure. Let us call this graph as the complete lattice.
- For each edge in the complete lattice, we draw a $\text{Bernoulli}(p)$ random variable. If it is 1, we keep the edge and delete it otherwise. In this way we can generate a random graph. We denote this model of random graph as $\text{Lat}(n, p)$.
- The graph below is a simulation from $\text{Lat}(6, 0.5)$ model.

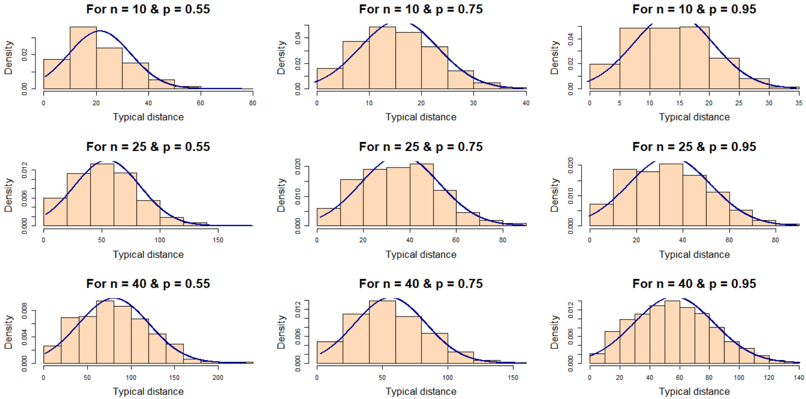


The Lat(n,p) Model: Histograms

It is known that the connectivity threshold for $\text{Lat}(n, p)$ is $p = 0.5$. So we generated from $\text{Lat}(n, p)$ 1000 times for each pair (n, p) varying $n = 5, 10, \ldots, 40$ and $p = 0.55, 0.60, \ldots, 0.95$. We considered the typical distance in each simulated graph when it was finite. Here we are presenting some of the histograms obtained.

The Lat(n,p) Model: Histograms

It is known that the connectivity threshold for $\mathrm{Lat}(n,p)$ is $p = 0.5$. So we generated from $\mathrm{Lat}(n,p)$ 1000 times for each pair (n,p) varying $n = 5, 10, \ldots, 40$ and $p = 0.55, 0.60, \ldots, 0.95$. We considered the typical distance in each simulated graph when it was finite. Here we are presenting some of the histograms obtained.



The Lat(n,p) Model: Testing Normality and Symmetry

- Again, first we standardized the data using sample mean and s.d. and observed the Q-Q plot. After that, we performed the same tests for normality as before.

The Lat(n,p) Model: Testing Normality and Symmetry

- Again, first we standardized the data using sample mean and s.d. and observed the Q-Q plot. After that, we performed the same tests for normality as before.
- After tabulating the p-values for the Kolmogorov-Smirnov test, we observed some high p-values in the lower left corner of the table than the other parts which corresponded to small values of p and large values of n.

The Lat(n,p) Model: Testing Normality and Symmetry

- Again, first we standardized the data using sample mean and s.d. and observed the Q-Q plot. After that, we performed the same tests for normality as before.
- After tabulating the p-values for the Kolmogorov-Smirnov test, we observed some high p-values in the lower left corner of the table than the other parts which corresponded to small values of p and large values of n.
- For Pearson's χ^2 goodness of fit test, we got almost all of the p-values very close to zero. But for that test too, the p-values in the lower left corner were comparatively larger.

The Lat(n,p) Model: Testing Normality and Symmetry

- Again, first we standardized the data using sample mean and s.d. and observed the Q-Q plot. After that, we performed the same tests for normality as before.
- After tabulating the p-values for the Kolmogorov-Smirnov test, we observed some high p-values in the lower left corner of the table than the other parts which corresponded to small values of p and large values of n.
- For Pearson's χ^2 goodness of fit test, we got almost all of the p-values very close to zero. But for that test too, the p-values in the lower left corner were comparatively larger.
- When we performed the non-parametric test of symmetry (RFPW), we saw the same pattern again; high p-values accumulating at the lower left corner of the table.

The Lat(n,p) Model: Testing Normality and Symmetry

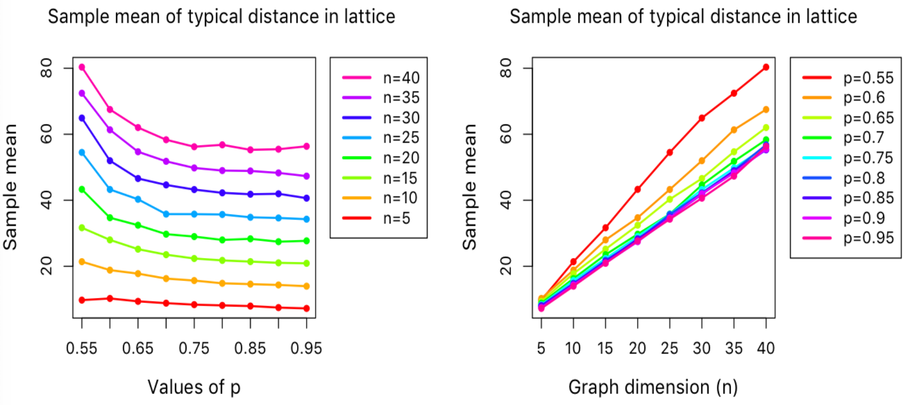
- Again, first we standardized the data using sample mean and s.d. and observed the Q-Q plot. After that, we performed the same tests for normality as before.
- After tabulating the p-values for the Kolmogorov-Smirnov test, we observed some high p-values in the lower left corner of the table than the other parts which corresponded to small values of p and large values of n.
- For Pearson's χ^2 goodness of fit test, we got almost all of the p-values very close to zero. But for that test too, the p-values in the lower left corner were comparatively larger.
- When we performed the non-parametric test of symmetry (RFPW), we saw the same pattern again; high p-values accumulating at the lower left corner of the table.
- A possible reason maybe: if we fix p and increase n, the mode of the histogram shifts rightward and the left tail becomes more and more visible, making the distribution more symmetric. So for fixed p we should have n large enough to get the symmetry.

The Lat(n,p) Model: Sample Mean

Let us take a closer look to the sample means of typical distances (when it is finite) for different choices of n and p. The plots are given below.

The Lat(n,p) Model: Sample Mean

Let us take a closer look to the sample means of typical distances (when it is finite) for different choices of n and p. The plots are given below.



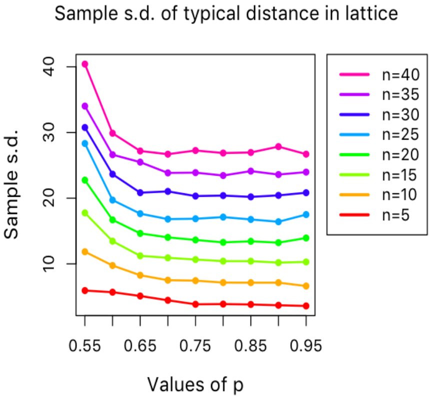
We observe that the sample mean grows more or less linearly with n. If we take p close to 1, the ratio of sample mean to n approaches a constant close to $4/3$.

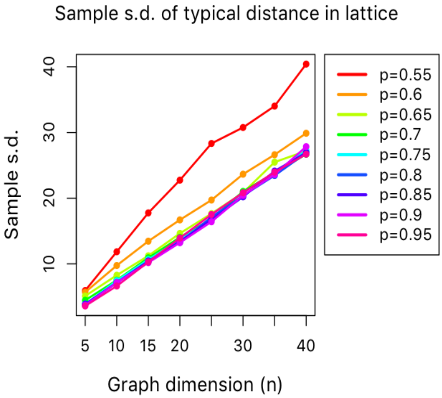
The Lat(n,p) Model: Sample S.D.

Now we will look at the sample s.d. of typical distance for different values of n and p. The plots are given below.

The Lat(n,p) Model: Sample S.D.

Now we will look at the sample s.d. of typical distance for different values of n and p. The plots are given below.





Here the sample s.d. also seems to grow linearly with n. And if we take p close to 1, the ratio of sample s.d. to n approaches a constant close to $2/3$ at a faster rate.

The Lat(n,p) Model: Observations

- Consider the complete lattice graph from $\text{Lat}(n, 1)$. We select two vertices from this graph with replacement. Let the coordinates be $(X_1,\ Y_1)$ and $(X_2,\ Y_2)$.

The Lat(n,p) Model: Observations

- Consider the complete lattice graph from $\mathrm{Lat}(n, 1)$. We select two vertices from this graph with replacement. Let the coordinates be $(X_1,\ Y_1)$ and $(X_2,\ Y_2)$.
- Then $X_1,\ X_2,\ Y_1,\ Y_2$ are i.i.d. random variables uniformly distributed over the set $\{-n,\ -n+1,\ \ldots,\ -1,\ 0,\ 1,\ \ldots,\ n-1,\ n\}$. And $H_n = |X_1 - X_2| + |Y_1 - Y_2|$.

The Lat(n,p) Model: Observations

- Consider the complete lattice graph from $\mathrm{Lat}(n, 1)$. We select two vertices from this graph with replacement. Let the coordinates be $(X_1,\ Y_1)$ and $(X_2,\ Y_2)$.
- Then $X_1,\ X_2,\ Y_1,\ Y_2$ are i.i.d. random variables uniformly distributed over the set $\{-n,\ -n+1,\ \ldots,\ -1,\ 0,\ 1,\ \ldots,\ n-1,\ n\}$. And $H_n = |X_1 - X_2| + |Y_1 - Y_2|$.
- $$\mathbb{P}(|X_1 - X_2| = k) = \begin{cases} \frac{1}{2n+1} & \text{when } k = 0. \\ \frac{2(2n+1-k)}{(2n+1)^2} & \text{when } k \in \{1,\ 2,\ \ldots,\ 2n\}. \end{cases}$$

The Lat(n,p) Model: Observations

- Consider the complete lattice graph from $\text{Lat}(n, 1)$. We select two vertices from this graph with replacement. Let the coordinates be $(X_1,\ Y_1)$ and $(X_2,\ Y_2)$.
- Then $X_1,\ X_2,\ Y_1,\ Y_2$ are i.i.d. random variables uniformly distributed over the set $\{-n,\ -n+1,\ \ldots,\ -1,\ 0,\ 1,\ \ldots,\ n-1,\ n\}$. And $H_n = |X_1 - X_2| + |Y_1 - Y_2|$.
- $$\mathbb{P}(|X_1 - X_2| = k) = \begin{cases} \frac{1}{2n+1} & \text{when } k = 0. \\ \frac{2(2n+1-k)}{(2n+1)^2} & \text{when } k \in \{1,\ 2,\ \ldots,\ 2n\}. \end{cases}$$
- Using this, we get $\mathrm{E}(H_n)/n \to 4/3$ and $\mathrm{Var}(H_n)/n^2 \to 4/9$ for $p = 1$. So this explains what we saw in the above two plots.

Concluding Remarks

- For the connectivity regime in Erdős–Rényi binomial random graph we saw a discrete nature in the distribution of H_n even after increasing n which was far from being normal. The limiting distribution of H_n requires further investigation (including studying the s.d.).

Concluding Remarks

- For the connectivity regime in Erdős–Rényi binomial random graph we saw a discrete nature in the distribution of H_n even after increasing n which was far from being normal. The limiting distribution of H_n requires further investigation (including studying the s.d.).
- As p increased through sparse super-critical regime and connectivity regime, the $o(1)$ term also increased gradually. Some more work is needed to unveil the relationship between p and the $o(1)$ term.

Concluding Remarks

- For the connectivity regime in Erdős–Rényi binomial random graph we saw a discrete nature in the distribution of H_n even after increasing n which was far from being normal. The limiting distribution of H_n requires further investigation (including studying the s.d.).
- As p increased through sparse super-critical regime and connectivity regime, the $o(1)$ term also increased gradually. Some more work is needed to unveil the relationship between p and the $o(1)$ term.
- For $\text{Lat}(n, p)$, our simulations tried to reveal the nature of H_n when $n \to \infty$ and $p \to 1$. The normality tests were rejected in most of the times. Nothing could be said from here about H_n if we fix $p \in (1/2, 1)$ and let $n \to \infty$. The limiting distribution of H_n for fixed p is yet to be studied.

Concluding Remarks

- For the connectivity regime in Erdős–Rényi binomial random graph we saw a discrete nature in the distribution of H_n even after increasing n which was far from being normal. The limiting distribution of H_n requires further investigation (including studying the s.d.).
- As p increased through sparse super-critical regime and connectivity regime, the $o(1)$ term also increased gradually. Some more work is needed to unveil the relationship between p and the $o(1)$ term.
- For $\text{Lat}(n, p)$, our simulations tried to reveal the nature of H_n when $n \to \infty$ and $p \to 1$. The normality tests were rejected in most of the times. Nothing could be said from here about H_n if we fix $p \in (1/2, 1)$ and let $n \to \infty$. The limiting distribution of H_n for fixed p is yet to be studied.
- We have only worked for square lattices. A similar study could be done for typical distances in triangular or hexagonal lattices or even higher dimensional lattice structures.

References

- Chung, F., & Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25), 15879-15882.
- Van Der Hofstad, R. (2009). Random graphs and complex networks. *Available on http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*, 11.
- *Link for Github Repository:* https://github.com/ghoshadi/random-graphs/

Thank You!