This is a project under Prof. Antar Bandyopadhyay
for the course Random Graphs

Studying the Structure of the Local Neighbourhood of a Randomly Selected Vertex of a Large but Sparse Erdős-Rényi Binomial Random Graph

Soham Das

June 2, 2020

Abstract

The main goal of this project is to statistically study local neighbourhood of a randomly selected vertex of a large but sparse Erdős-Rényi random graph. Given any d, the project tries to identify the statistical properties of the d-depth neighbourhood for large but sparse Erdős-Rényi random graphs.

1 Definitions

Erdős-Rényi random graph: If a graph has n vertices and probability of joining any two vertices by an edge is p. The set of all such graphs is noted as $G(n,p)$. Any graph that belongs to this set is called Erdős-Rényi binomial random graph.

Sparse graph: A graph is called sparse if the ratio of its total number of edges to the total number of possible edges tends to zero when the total number of vertices tends to infinity. In other words, if a graph has n many vertices and it has edges of order $o(n^2)$ then the graph is called sparse.

Large graph: When total number pf vertices in a graph i.e., n is very large.

Sparse Erdős-Rényi random graph: If the parameter p of the set ER random graph $G(n,p)$ is of the form $p = c/n$ then the random graph is called sparse, where c is any positive constant and n is the total number of vertices.

Galton-Watson process: A Galton–Watson process is a stochastic process $\{X_n\}$ which evolves according to the recurrence formula
$X_0 = 1$ and $X_{n+1} = \sum_{j=1}^{X_n} \xi_j^{(n)}$, where $\{\xi_j^{(n)} : n, j \in \mathbb{N}\}$ is a set of independent and identically-distributed natural number-valued random variables.

This project is about identifying the property of a d-neighbourhood in a large but sparse Erdős-Rényi random graph statistically(through simulation). I have simulated random graphs from $G(n,p)$ with $p = c/n$ where c is some positive constant, fixed a vertex randomly, and studied the properties of its d-depth neighbourhood in those graphs.

2 Simulations and observations

2.1 Probability of the d-neighbourhood being a tree

Let be the probability of the d-neighbourhood being tree be p_{tree}. p_{tree} depends on c, d and n simultaneously. In this section, I have fixed $d = 4$. Then I have simulated 20 graphs g_i, $i = 1, 2, \cdots, 20$ from $G(n, p)$, chosen a vertex v_i randomly from g_i and named its d-neighbourhood as ngh_i. An estimate of p_{tree} is $\widehat{p}_{tree} = \frac{1}{20}\sum_{i=1}^{20} I(ngh_i \text{ is a tree})$, $I(\cdot)$ is indicator function. In this way, I have simulated $\widehat{p}_{tree}$ 100 times and plotted the histogram for different values of c and n, $c = 0.5, 1, 1.5, 2; n = 1000, 3000, 6000, 10000, 30000, 60000$.

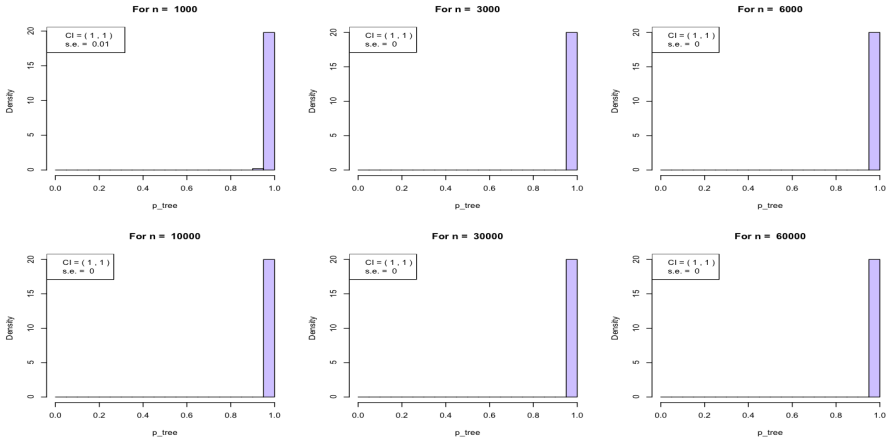

Figure 1: For $c = 0.5$

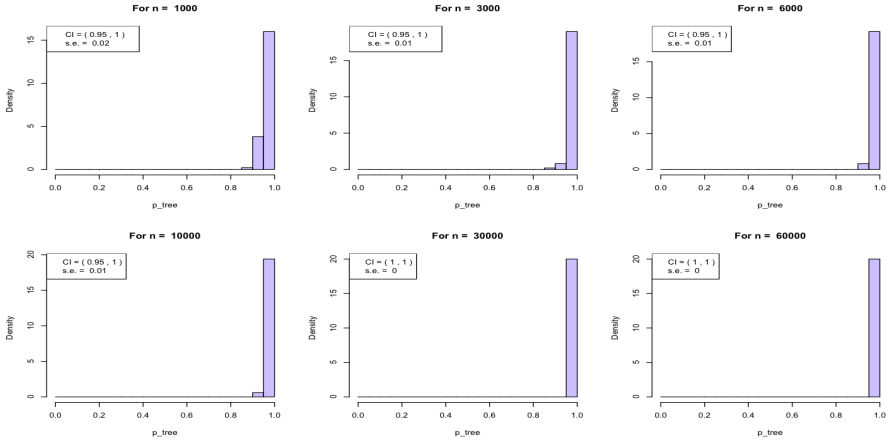

Figure 2: For $c = 1$

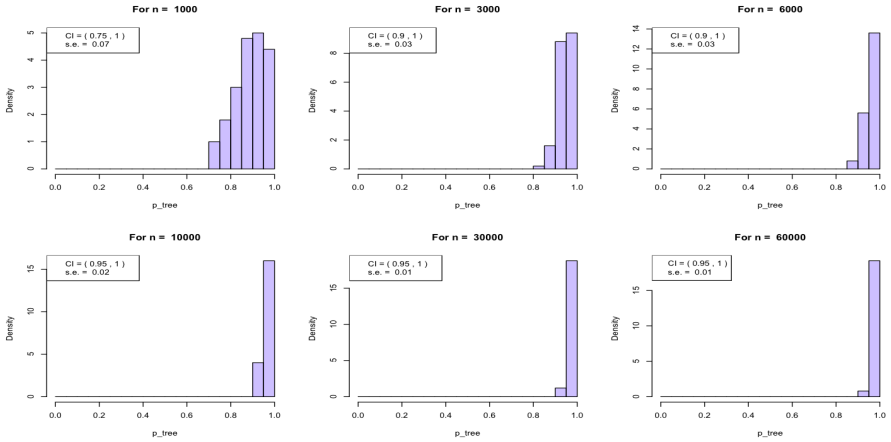

Figure 3: For $c = 1.5$

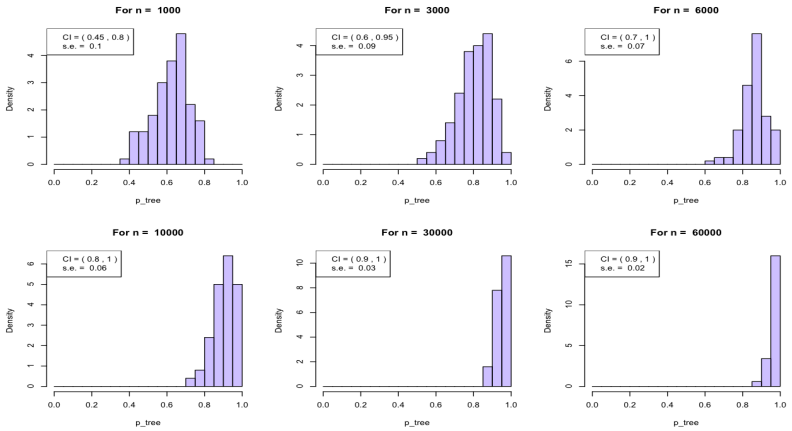

Figure 4: For $c = 2$

Observation:

- We see that for each of c, as n increases, standard error of $\hat{p}_{tree}$ decreases, confidence interval shrinks and $\hat{p}_{tree}$ becomes closer to 1. So, we can say that for a fixed d and sufficiently large n, the d-neighbourhood becomes a tree.
- Also if we increase c, the above property holds true but it is visible only for larger and larger n as c increases. This is obvious because probability of having an edge between two vertices i.e., c/n increases with c.

2.2 Expected number of cycles in d-neighbourhood

In the same spirit, I simulated the expected number of cycles in d-neighbourhood. I counted the total no. of cycles for 500 random graphs from $G(n, p)$ and took their average for d fixed at 4.
Note: Here I have counted only triangles, quadrilaterals and pentagons. Total no. of cycles should not be much different from this count as these three constitute most of the cycles.

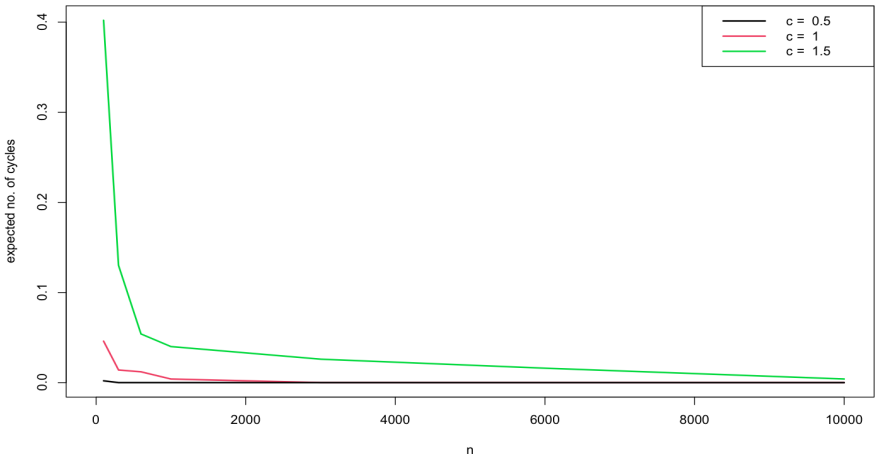

Figure 5: Expected no. of cycles for different c

Observation:

- It follows from the last observation as a corollary that this expectation value goes to zero as n tends to infinity for any fixed c.

2.3 Probability of the d-neighbourhood being a tree(contd.)

In section 3.1, I estimated p_{tree} for different values of c fixing $d = 4$. Now I estimated p_{tree} for different values of d fixing $c = 3$. And plotted the histograms of $\widehat{p}_{tree}$ as in 3.1, for $d = 3, 4, 5; n = 1000, 3000, 6000, 10000, 50000, 100000$.

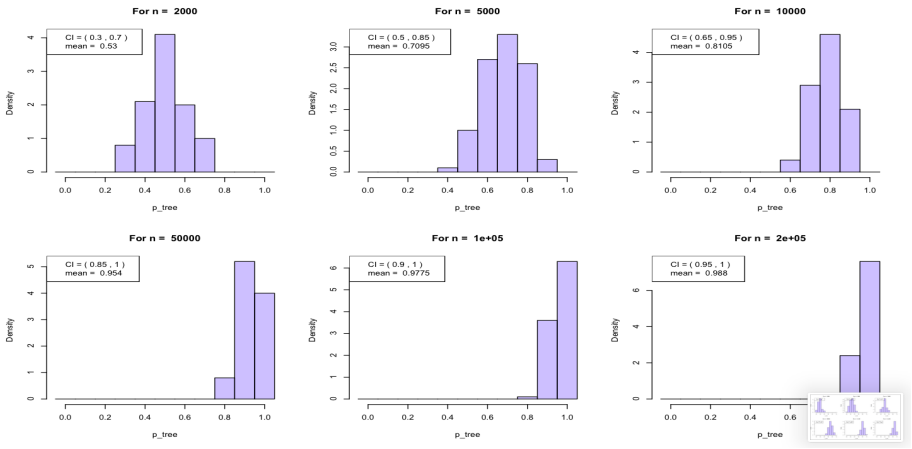

Figure 6: For $d = 3$

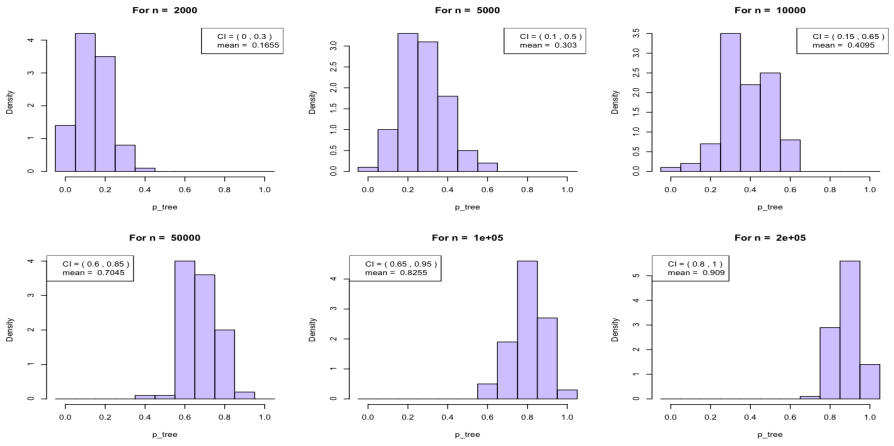

Figure 7: For $d = 4$

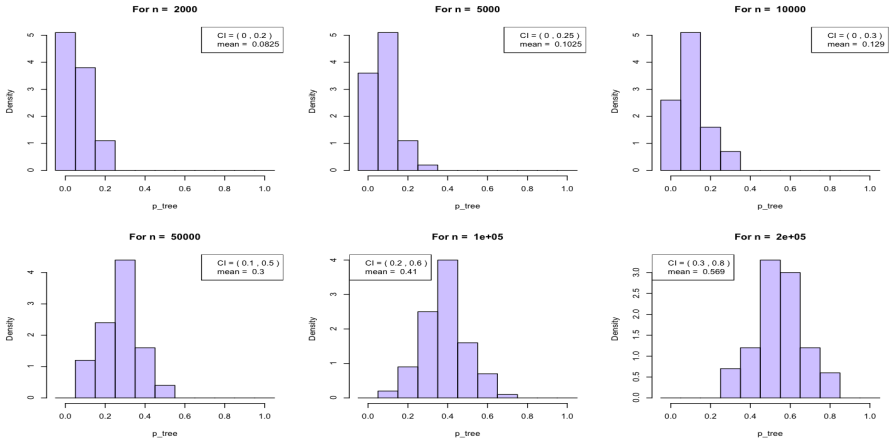

Figure 8: For $d = 5$

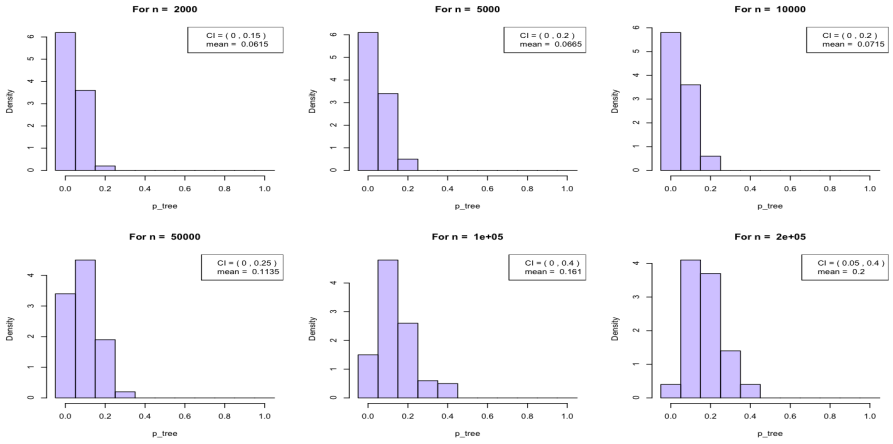

Figure 9: For $d = 6$

Observation:

- We can see that $\hat{p}_{tree}$ becomes close to 1 as n increases as before for any fixed c. Also this property is visible for larger values of n as d increases. The distribution of $\widehat{p}_{tree}$ moves towards left and mean of $\widehat{p}_{tree}$ decreases drastically as d increases keeping n fixed. This is evident from the fact that for larger d we are considering larger subgraph with more vertices and more chance of finding a cycle.
- From here we understand that the probability of d-neighbourhood being a tree, i.e., p_{tree} decreases with increase in c or d and increases with increase in n.

In this project we concentrate on the properties of d-neighbourhood of a random vertex v for large and sparse graph. We can observe the opposing effect of c and d against n for the d-neighbourhood being a tree. That is why from now on we fix d at a standard value, say 4, so that the d-neighbourhood is a tree and observe its properties for different c and n. And sufficiently large n is taken for simulation.
Also from here on, we assume that for any random vertex we choose from a graph in $G(n.p)$, with $p = c/n$, its d-neighbourhood is a tree.

2.4 Degree of a random vertex

In this section we will investigate the distribution of degree of a randomly chosen vertex in d-neighbourhood. I have chosen a vertex v randomly from a graph in $G(n, p)$. Let its degree be Deg_0.
A random vertex v can be connected with $n-1$ many other vertices each with probability $p = c/n$. This implies the degree $Deg_0 \sim \text{Poi}(c)$.
To check the Poisson distribution through simulation, I have formed histograms of 1000 simulated values of Deg_0 for different c and n, $c = 0.5, 1, 1.5, 2, 3, 5$;
$n = 1000, 3000, 6000, 10000, 30000, 60000$. Also the p-values of Kolmogorov-Smirnov test, Chi-squared test and t-test for each of the histograms are also mentioned in the following figure.

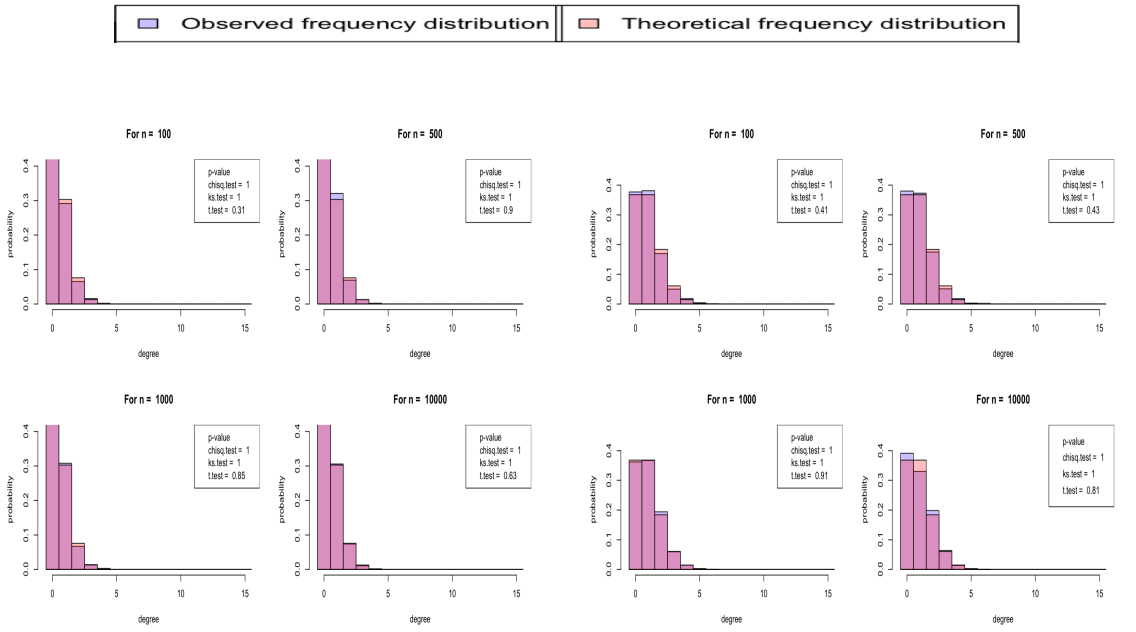

Figure 10: For $c = 0.5$

Figure 11: For $c = 1$

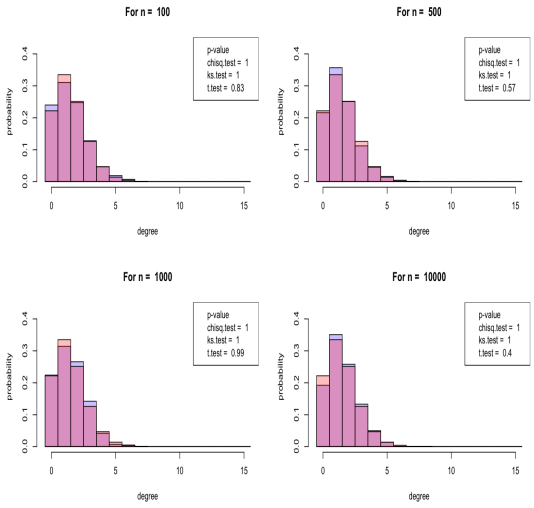

Figure 12: For $c = 1.5$

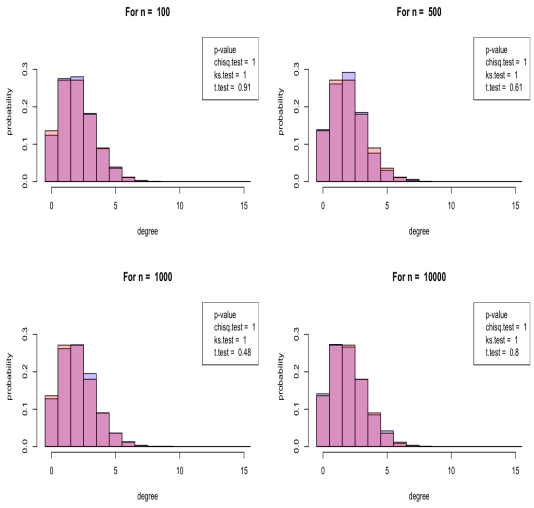

Figure 13: For $c = 2$

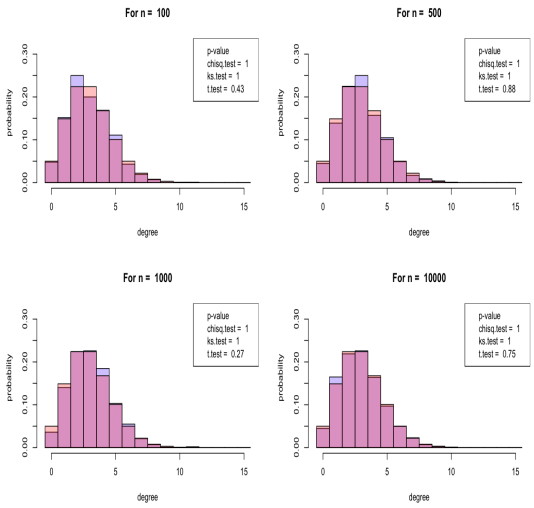

Figure 14: For $c = 3$

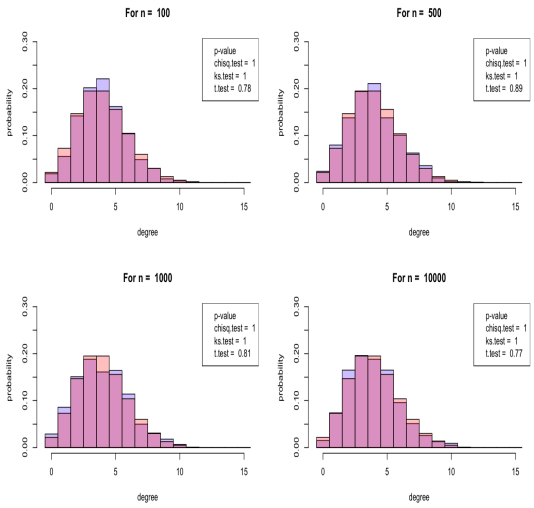

Figure 15: For $c = 4$

Observation:

- It is clear from the histograms that Deg_0 indeed follows Poi(c) as expected and the p-values of each of the histograms also validate that.

2.5 Degree of a random vertex at t-depth form a random vertex

Similarly, I also formed a histogram of 1000 simulated values of degree of a random vertex at t-depth from initially selected random vertex v, for different values of t, $t = 1, 2, 3$. Let the degree of a vertex at t-depth be Deg_t.
I have only shown the histograms for $t = 1$.

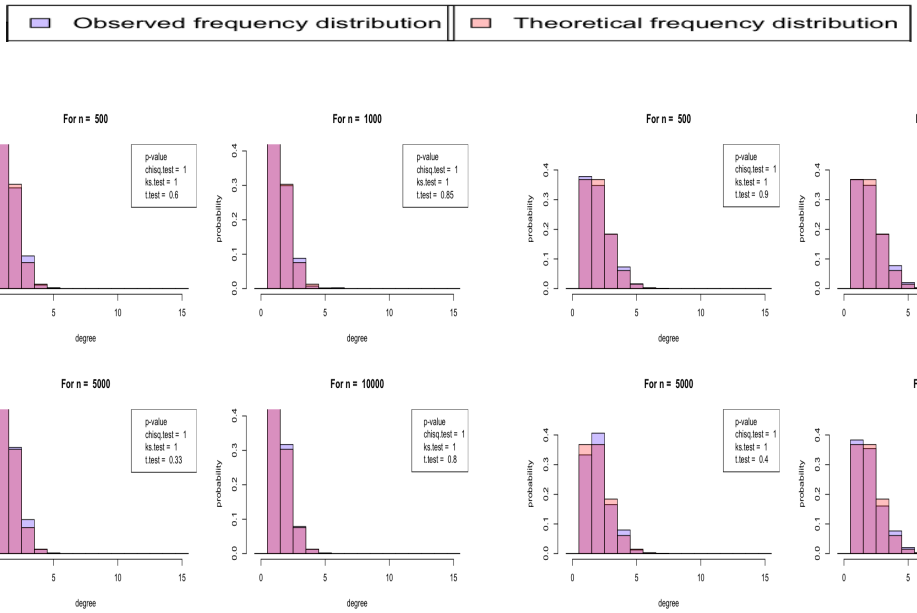

Figure 16: For $c = 0.5$

Figure 17: For $c = 1$

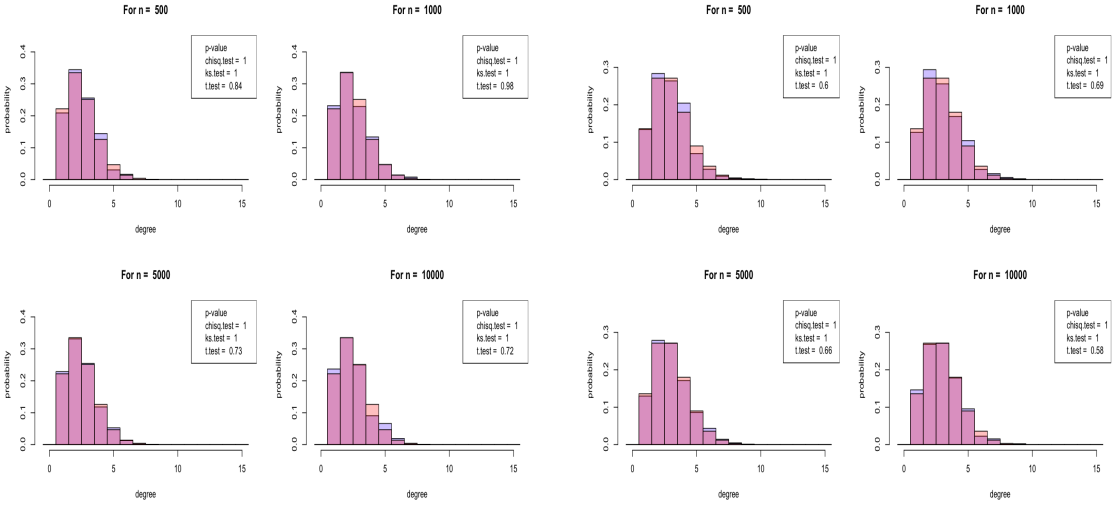

Figure 18: For $c = 1.5$

Figure 19: For $c = 2$

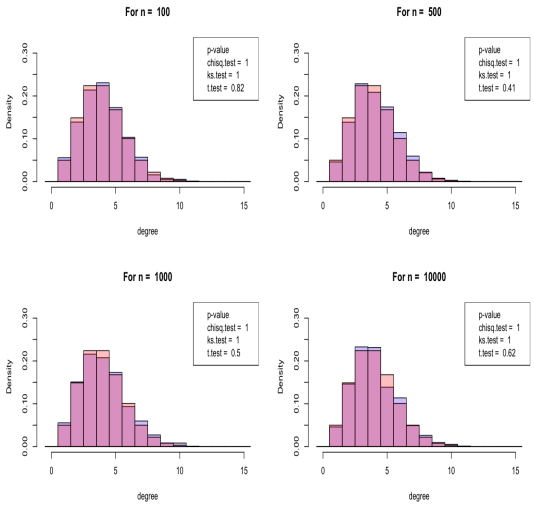

Figure 20: For $c = 3$

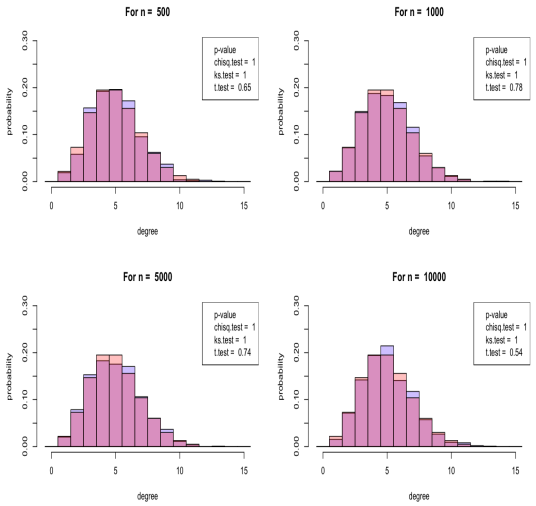

Figure 21: For $c = 4$

Observation:

- We observe that, $Deg_1 \sim 1 + \text{Poi}(c)$ and also for $t = 2, 3$, $Deg_t \sim 1 + \text{Poi}(c)$. We have easy explanation to it. Let us say the initially selected random vertex is v. Given that a vertex is at depth t, we know it is not connected to the vertex v and any vertex at depth $1, 2, \cdots, t-2$ from v in the subgraph. Let the set of such vertices be S. A vertex at depth t can be connected with any vertex which is not in S with probability $p = c/n$. Though the cardinality of S may vary, its expected value is almost negligible to large n. Also it is clear that this vertex is definitely connected to a vertex at depth $t-1$. Hence, its degree follows distribution $1 + \text{Poi}(c)$.

- I also noted the p-values of Kolmogorov-Smirnov test, Chi-squared test and t-test for each of the histograms in the figures. As we see the p-values are pretty high.

2.6 Independence of degrees of two random vertices in d-neighbourhood

Next we check the independence of the degrees of vertices in the d-neighbourhood subgraph. I have collected 100 pairs of degrees of two randomly selected vertices from d-neighbourhood of random vertex v in a graph from $G(n,p)$. Then I calculated Kendall and Pearson correlation coefficient and p-value of Chi-squared test of independence of these pairs for different values of c and n.

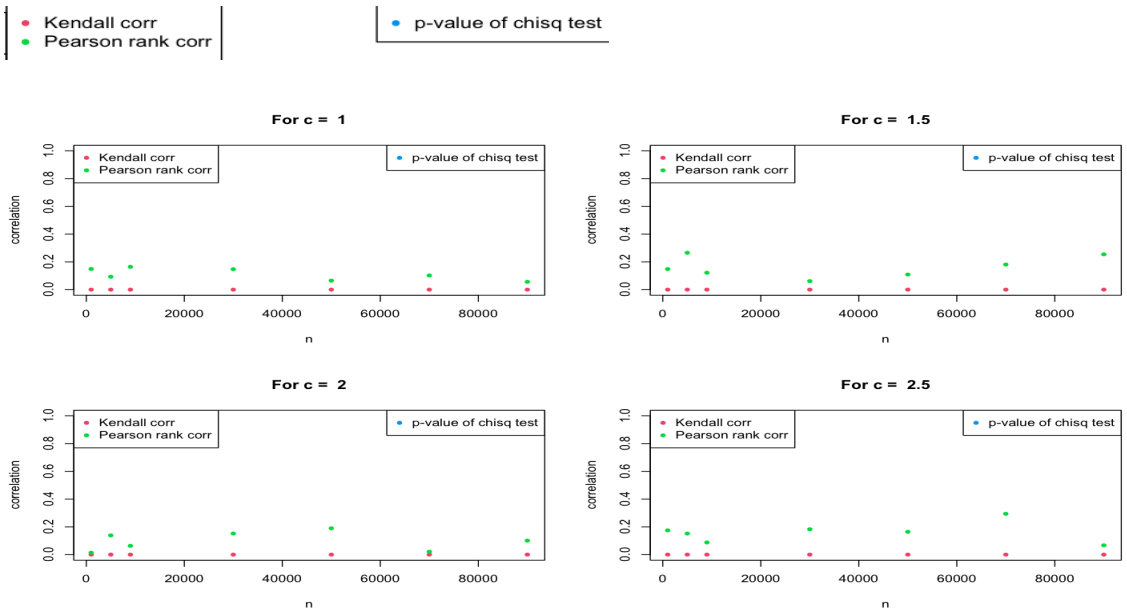

Figure 22: Kendall and Pearson correlation coefficient

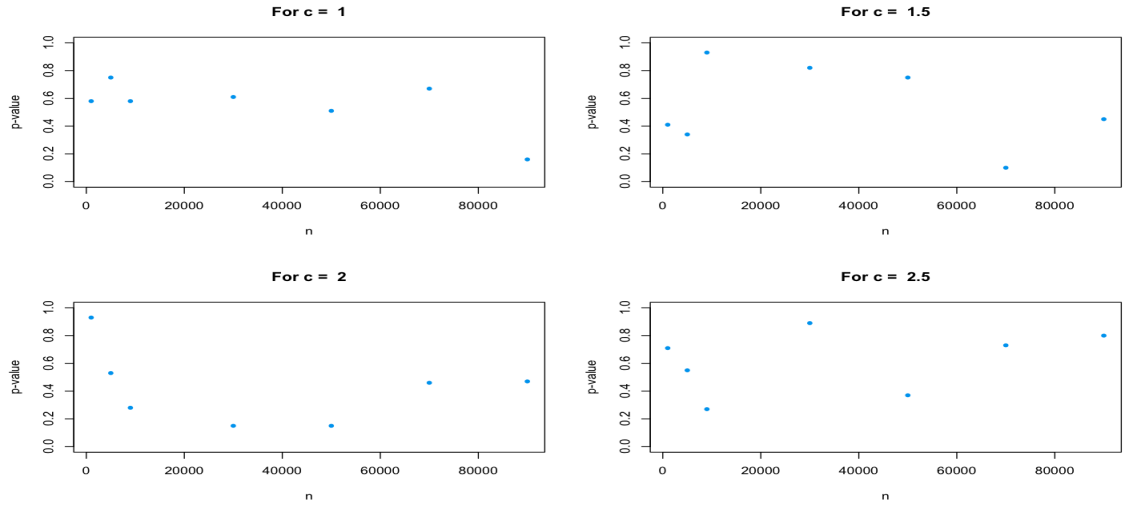

Figure 23: p-value of Chi-square test of independence

Observation:

- As we see, correlation between degrees of any two randomly selected vertices is very low in both Pearson and Kendall method and p-values for chi-square independence test are greater than 0.05. Hence, we can infer that the degrees of the vertices in d-neighbourhood are independent.

Theorem. Let X and Y be two random variables such that $X \sim Poi(\mu)$ and $Y \sim Poi(\lambda)$ then $X+Y \sim Poi(\mu+\lambda)$ if X and Y are independent but not vice versa.

Proof. Let characteristic functions of X, Y and $X+Y$ are $\phi_X = exp[\mu(e^{it}-1)]$, $\phi_Y = exp[\lambda(e^{it}-1)]$ and ϕ_{X+Y} respectively. If X and Y are independent, then $\phi_{X+Y} = \phi_X \phi_Y = exp[(\mu+\lambda)(e^{it}-1)]$ which says $X+Y \sim \text{Poi}(\mu+\lambda)$.

For the other part we have to produce a counter example. Lets start with the given joint distribution table of X and Y.

Total		Y							
		q_0	q_1	q_2	q_3	q_4	q_5	$\cdots$	
X	p_0	p_0q_0	p_0q_1	p_0q_2	$p_0q_3-\delta$	$p_0q_4+\delta$	p_0q_5	$\cdots$	
	p_1	p_1q_0	p_1q_1	p_1q_2	$p_1q_3+\delta$	$p_1q_4-\delta$	p_1q_5	$\cdots$	
	p_2	p_2q_0	$p_2q_1+\delta$	$p_2q_2-\delta$	p_2q_3	p_2q_4	p_2q_5	$\cdots$	
	p_3	p_3q_0	$p_3q_1-\delta$	$p_3q_2+\delta$	p_3q_3	p_3q_4	p_3q_5	$\cdots$	
	p_4	p_4q_0	p_4q_1	p_4q_2	p_4q_3	p_4q_4	p_4q_5	$\cdots$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	

Table 1: Joint distribution table

In table 1, $p_x = \Pr(X=x)$, $p_y = \Pr(Y=y)$ are given and δ is some suitably small positive number.
As the marginal distribution of X and Y are given by $\text{Poi}(\mu)$ and $\text{Poi}(\lambda)$ respectively, the row and column sums match with the marginal distributions: $p_i = \sum_j p_i q_j$ and $q_j = \sum_i p_i q_j$. Also the off diagonal sums also match with $X+Y \sim \text{Poi}(\mu+\lambda)$ as: $\Pr(X+Y=k) = \sum_i p_i q_{k-i}$.

□

According to the last theorem, if X, Y, $X+Y$ follows $\text{Poi}(\mu)$, $\text{Poi}(\lambda)$ and $\text{Poi}(\mu+\lambda)$ respectively then we cannot necessarily say that X and Y are independent but it eliminates one possible way to disprove independence.

Earlier in this section we took 100 pairs of observations to check independence. Similarly, now we took 1000 such pairs, added each pairs and tested if those 1000 sums follow $2 + \text{Poi}(2c)$ or not.

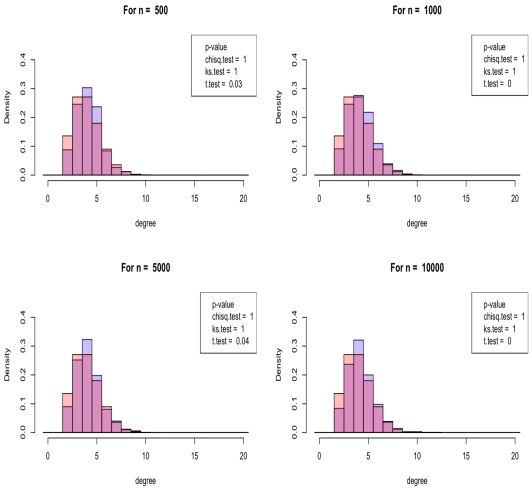

Figure 24: For $c = 1$

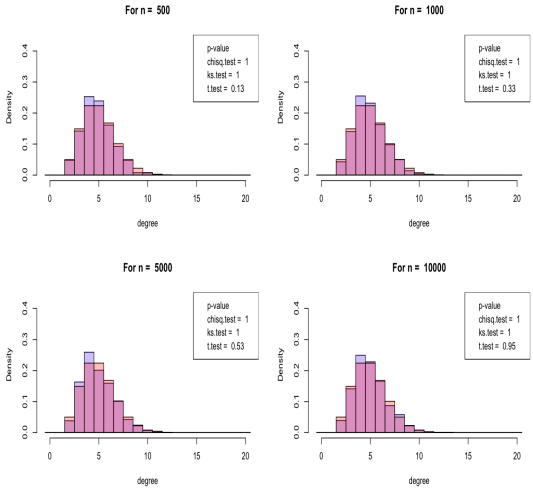

Figure 25: For $c = 1.5$

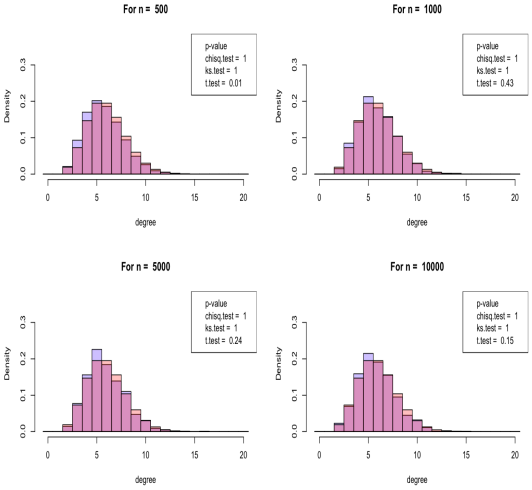

Figure 26: For $c = 2$

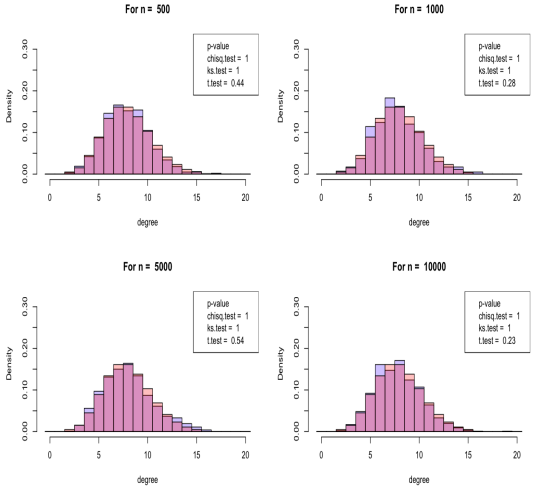

Figure 27: For $c = 3$

Observation:

- From the histograms it is clear that the sum of two randomly chosen vertices from d-neighbourhood follows $2 + \text{Poi}(2c)$.
- It has also been checked that sum of degrees of three such vertices follows $3 + \text{Poi}(3c)$.

3 Results

Lets summarize the observations we got from the previous section:

- For fixed d, c and suitably large n, the d depth neighbourhood of a randomly chosen vertex in a Erdős-Rényi binomial random graph from $G(n,p)$ with $p = c/n$ is a tree.
- The degree of that randomly chosen vertex follows Poisson distribution with parameter c.
- The degree of any other vertex from that tree follows distribution $1 + \text{Poi}(c)$.
- Degrees of all the vertices in this tree are independent

From these observations we can conclude that the d-depth neighbourhood of a randomly slected vertex is a tree which has distribution same as a d-depth neighbourhood of a Galton-Watson process with progeny distribution $\text{Poi}(c)$.

4 R codes

Use the following link to get the R codes for all the simulations:
https://github.com/Sohamdas-stat/Random-graphs.