UCLA

232E - Large-Scale Social and Complex Networks: Design and Algorithms

Report on

# Project 1: Random Graphs and Random Walks

Authors
**Jayanth** Shreekumar (805486993)
**Leo** Ly (805726182)
**Yuheng** He (505686149)

Spring 2022

# Part 1 - GENERATING RANDOM NETWORKS

**Question 1 a**

In this part, we created undirected random networks with n = 1000 nodes, and we used different probability $p$ for drawing an edge between two arbitrary vertices. The probability $p$ are 0.003, 0.004, 0.01, 0.05, and 0.1.

Erdös-Rényi Models $G(n, p)$ is a random graph with $n$ vertices where each possible edge has probability $p$ of existing. An Undirected Erdös-Rényi Models is defined as a random network where there is no direction in the edges of the graph. a graph in $G(n, p)$ has on average $\binom{n}{2}p$ edges. The distribution of the degree of any particular vertex is binomial and is shown as follows.

$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

The degree distributions are plotted in Fig. 1. As we have explained earlier, the degree distribution follows binomial distribution and we can see from Fig. 1 that binomial distribution is more visible with larger $p$.

To derive the theoretical mean and variance for the model, we have the following.

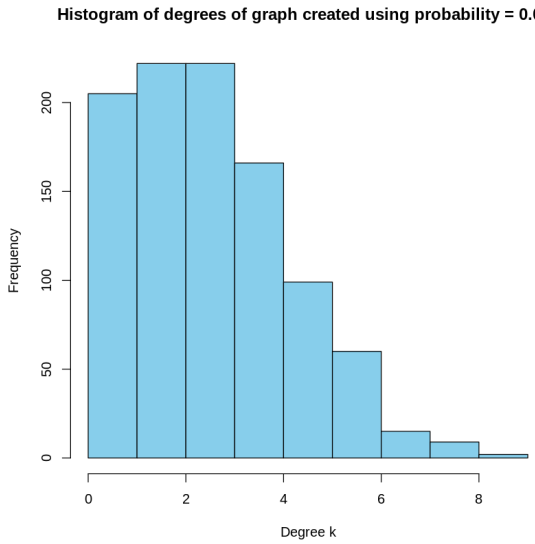
$$E[\deg(v)] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-1-k} = np$$

$$Var[\deg(v)] = np(1-p)$$

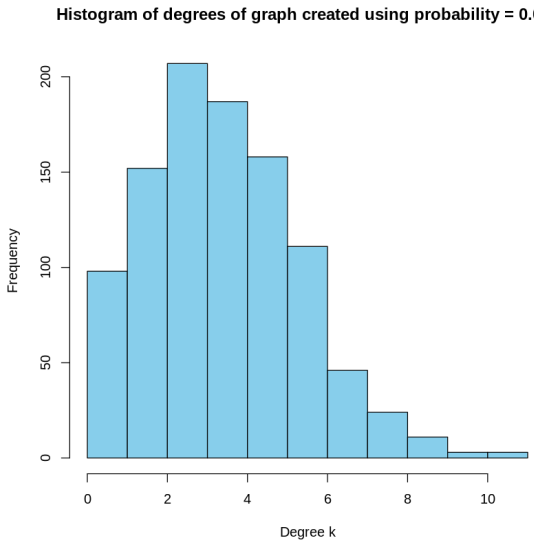
Therefore, we know that for a model with given $n$ and $p$, the mean for the model is $np$, and the variance is $np(1-p)$. The expected mean (Exp. Mean), observed mean (Obs. Mean), expected variance (Exp. Var), and observed variance (Obs. Var) are presented in Table 1. As we can see from the table, the expected values and the observed ones are close, meaning that the models are theoretically correct.

| $p$ | Exp. Mean | Obs. Mean | Exp. Var | Obs. Var |
|---|---|---|---|---|
| 0.003 | 3 | 2.980 | 2.991 | 2.929 |
| 0.004 | 4 | 3.884 | 3.984 | 3.750 |
| 0.01 | 10 | 10.122 | 9.9 | 9.628 |
| 0.05 | 50 | 49.906 | 47.5 | 46.079 |
| 0.1 | 100 | 100.076 | 90 | 90.080 |

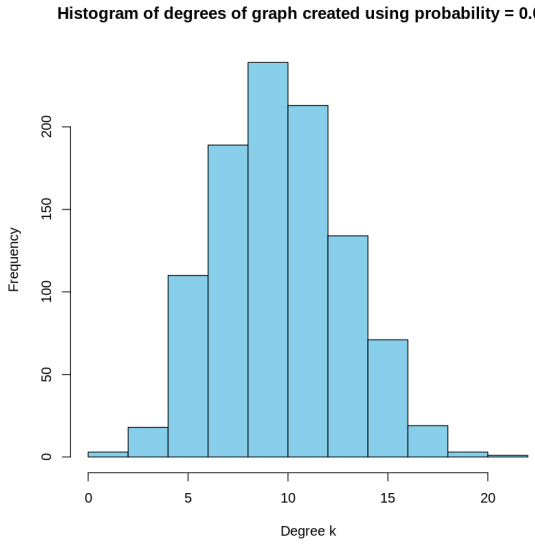
Table 1: Expected and Observed Mean and Variance for Different $p$



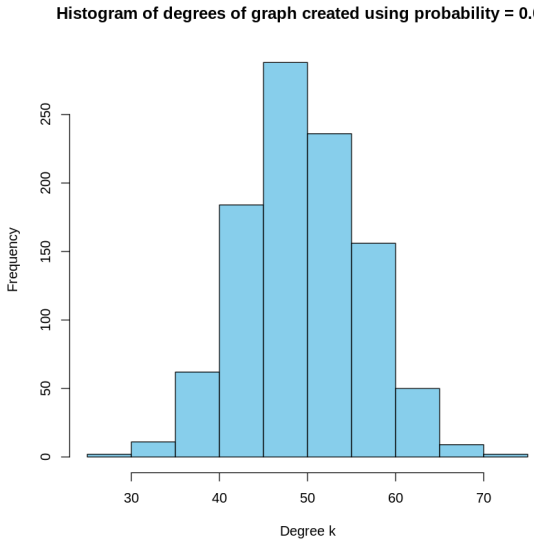

(a) Degree Distribution with $p = 0.003$



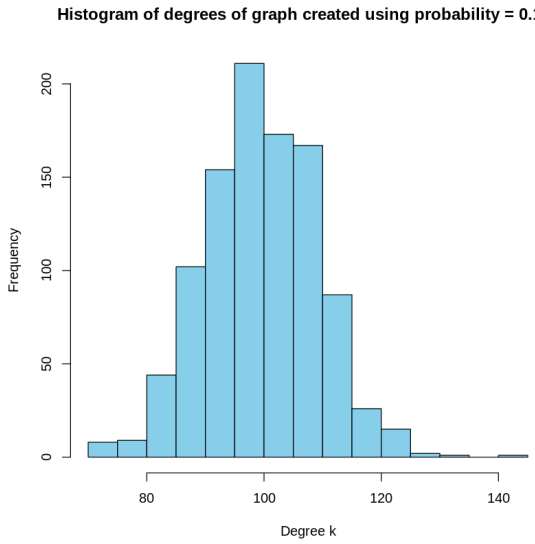

(b) Degree Distribution with $p = 0.004$




(c) Degree Distribution with $p = 0.01$




(d) Degree Distribution with $p = 0.05$




(e) Degree Distribution with $p = 0.1$

Figure 1: Degree Distribution with Varying $p$

### Question 1 b

In this part, we stepped further from the last question to see if all realizations of the random ER network are connected. We found the giant connected component (GCC) for the instance $p$ where the network is not connected. We also obtained the probability of the connectedness of the network, the diameter of the GCC, the number of nodes and edges for the non-connected networks. The results are given in Table 2. **For $p = 0.05$ and $p = 0.1$, we observe fully connected networks and thus the GCC does not exist and hence these $p$ are not included in the table.** It can be seen from Table 2 that for the non-connected instances, as the probability increases, the diameter decreases and the number of nodes and edges increase.This is expected because as the probability increases, the network evolves toward a connected realization and the number of nodes and edges will increase to a fully-connected instance. The approximation of the diameter of the GCC can be given as follows.

$$d = \frac{ln(p)}{ln(np)}$$

It explains why when $p$ increases, the diameter of GCC decreases.

| $p$ | Probability of Connectedness | Diameter of GCC | Number of Nodes | Number of Edges |
|---|---|---|---|---|
| 0.003 | 0 | 14 | 946 | 1502 |
| 0.004 | 0 | 11 | 970 | 1937 |
| 0.01 | 0.959 | 5 | 999 | 4852 |

Table 2: Connectedness and GCC of the non-connected Networks with Different $p$

### Question 1 c

Giant components are a prominent feature of the Erdős–Rényi model. Each possible edge connecting pairs of a given set of n vertices is present independently of the other edges, with probability $p$. This means that a giant connected component appears when the probability exceeds a certain threshold value. Similarly, there is another threshold value for probability at which almost all nodes belong to the GCC, meaning that the graph is almost fully connected. As required, **we swept over values of $p$ from 0 to 0.009 in steps of 0.0001.** The results are displayed in the scatterplot shown below.

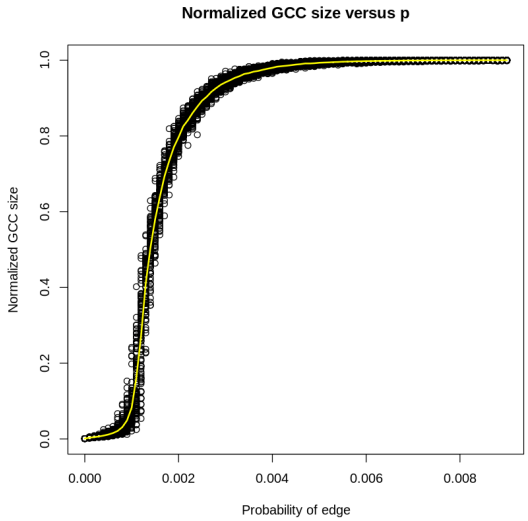



Figure 2: GCC size vs probability of an edge for undirected Erdős–Rényi model with n = 1000

**(i)** The criterion of emergence is the probability value at which we observe a GCC of size = np. For our experiment, we picked $p = 0.01$ for the criterion of emergence, meaning that we considered that we obtained a GCC when it was about 0.1 percent the size of a fully connected graph with $n = 1000$ nodes. Also, the value of $p$ at which a GCC appears is theoretically equal to

$$p_c = \frac{1}{n} = 0.001$$

From the graph, we see that our results match this theoretical value, and we obtain a GCC at around 0.0013.

**(ii)**

Theoretically, the value at which we have an almost fully connected graph is given by:

$$p_c = \frac{ln(n)}{n} = 0.006907$$

This is again observed in the graph, where we see that the graph is almost fully connected around $p = 0.0068$

$$p_c = \frac{1}{n} = 0.001$$

**Question 1 d**

**(i)** In this part, we studied the GCC of ER network with the average degree of nodes $c = n \times p = 0.5$. We swept over the number of nodes $n$ which ranges from 100 to 10000 (with step size 50). The result is plotted in Fig. 3.

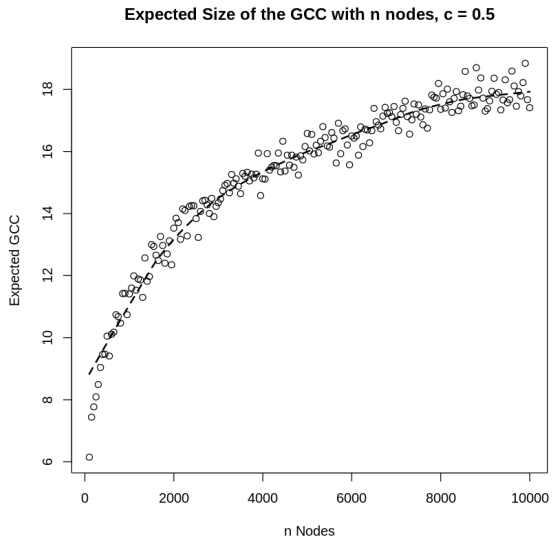



Figure 3: Expected Size of the GCC with n nodes, $c = 0.5$

As is observed from Fig. 3, the expected size of GCC increases as the number of nodes $n$ increases in a logarithmic trend. As presented by Erdös and Rényi [ER$^+$60], If $c = np < 1$, then a graph in $G(n, p)$ will almost surely have no connected components of size larger than $O(ln(n))$. This is explained by the fact that both $n$ and $p$ are constrained by $c$. Since $c$ is a constant number, when $n$ grows, $p$ decreases accordingly. For $c < 1$, it is guaranteed that the size of GCC is restricted by $\frac{\ln(n)}{\ln(c)}$ as is explained in Question 1(b). This property give us an insight to why the trend is logarithmic.

**(ii)** In this section, we will repeat the same process for $c = 1$. Similarly we swept over the number of nodes $n$ which ranges from 100 to 10000 (with step size 50). The result is plotted in Fig. 4.

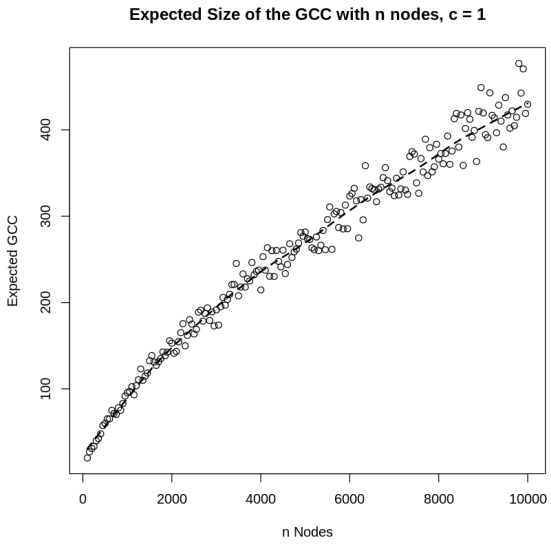



Figure 4: Expected Size of the GCC with n nodes, $c = 1$

As is observed from Fig. 4, the expected size of GCC increases as the number of nodes $n$ increases. The trend is higher than that of $c = 0.5$ as previously presented. As presented by Erdös and Rényi [ER$^+$60], If $c = np = 1$, then a graph in $G(n, p)$ will almost surely surely have a largest component whose size is of order $n^{2/3}$. The trend we observe conforms to this theory.

**(iii)** In this section, we will repeat the same process for $c = 1.1, 1.2, 1.3$. Similarly we swept over the number of nodes $n$ which ranges from 100 to 10000 (with step size 50). The result is plotted in Fig. 5.

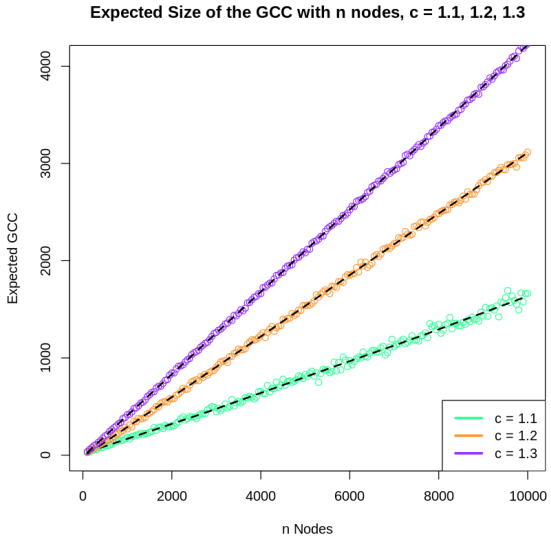



Figure 5: Expected Size of the GCC with n nodes, $c = 1.1, 1.2, 1.3$

We can see from Fig. 5 that expected size of GCC increases as the number of nodes $n$ increases in a linear trend. This is explained by Erdös and Rényi [ER$^+$60], that If $np = c > 1$, where $c$ is a constant, then a graph in $G(n, p)$ will almost surely have a unique giant component containing a positive fraction of the vertices. No other component will contain more than $O(log(n))$ vertices.

**(iv)** The relations are given as follows.

- $np = c < 1$: $E(GCC) \propto ln(n)$
- $np = c = 1$: $E(GCC) \propto n^{2/3}$
- $np = c > 1$: $E(GCC) \propto n$

**Question 2 a**

The Barabási model is an algorithm for generating random graph using a preferential attachment mechanism where nodes with higher degrees are more likely to be attached to by new nodes.As a new node will connect to m existing nodes, by construction, Barabási graphs are connected graphs.

As required by this question, we created a random graph using the function **sample_pa()** where each new node attaches to 1 previously existing node (m = 1). As m = 1, we expect a tree to be formed as there will be no cycles in the random graph. The generated graph is shown below:

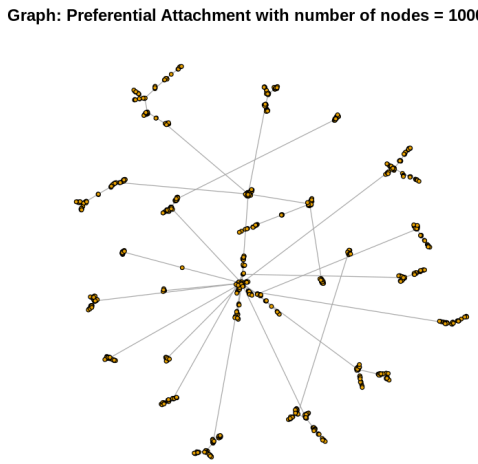



Figure 6: Preferential Attachment graph with m = 1, n = 1000

To prove that a network created using preferential attachment is always connected, we repeated the graph creation process 10000 times and the obtained graph was connected every single time.

**Question 2 b**

Graph nodes are often from groups that are almost independent from the rest of the graph, with which they share a few edges, but the edges between nodes in that small group is denser. Such groups of nodes are called communities and the community structure describes the communities existing in a given graph. Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into communities.

The community structre of a preferential attachment graph with m = 1 and n = 1000 is shown below. **The modularity of this network was found to be 0.93266.**

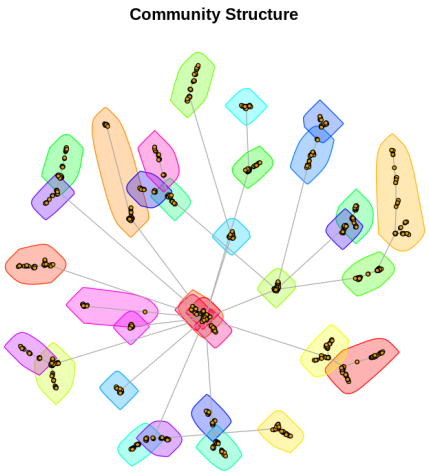
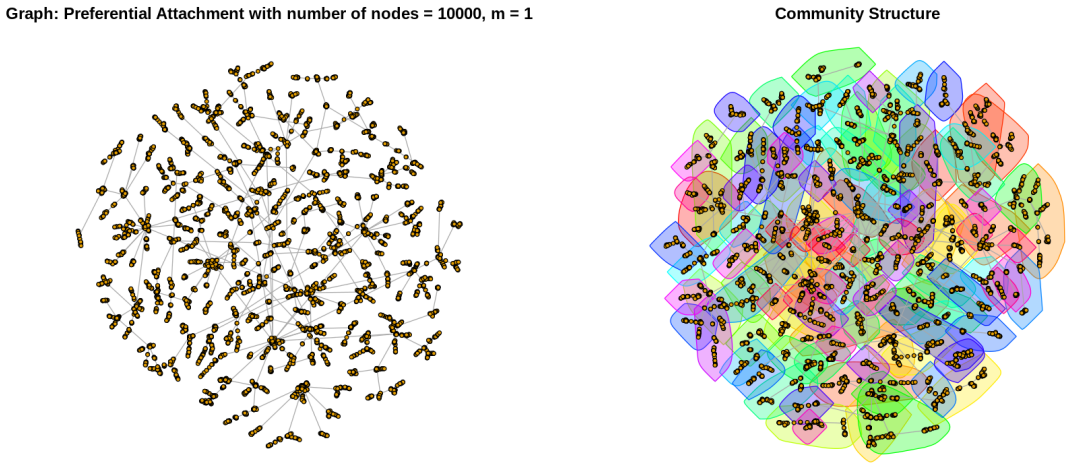



Figure 7: Community Structure for a Preferential Attachment graph with m = 1, n = 1000

### Question 2 c

In this question, we repeated the above steps for a graph with n = 10000, m = 1. The graph and its community structure are shown below. **The modularity of this network was found to be 0.97814.**




(a) Preferential Attachment graph with m = 1, n = 10000

(b) Community Structure for a Preferential Attachment graph with m = 1, n = 10000

Figure 8: A random preferential attachment graph with n = 10000, m = 1 and its community structure

The modularity of a graph with n = 10000 is higher than the one with n = 1000. This means that the larger graph can be separated into bigger clusters of communities that are also more independent from the connected graph. This is as expected, as the more the number of nodes, the more

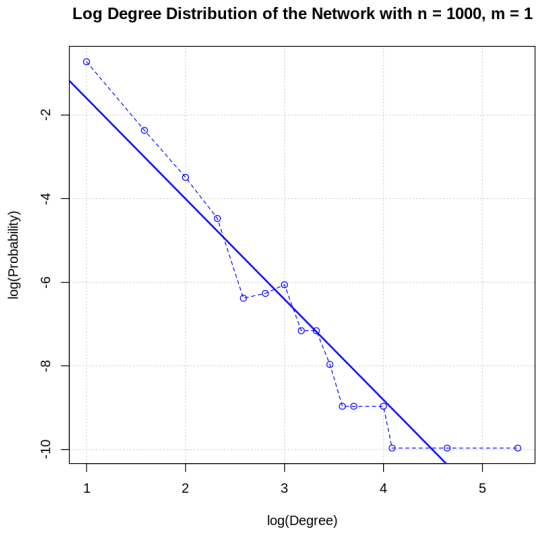
the preference to existing nodes with higher degree while building the model.

### Question 2 d

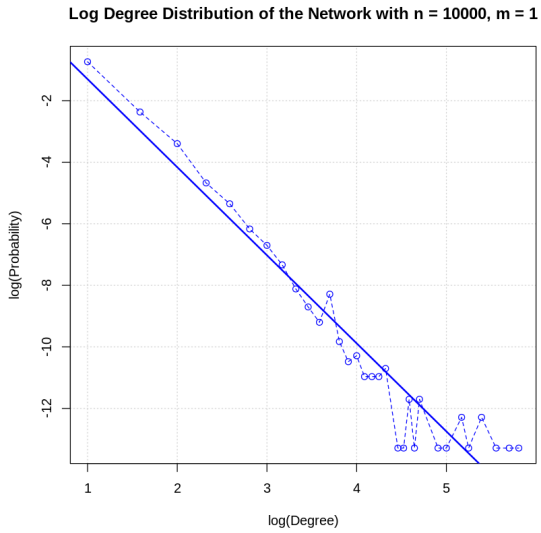
We studied in class that the preferential attachment model is a power-law network, and that its degree distribution follows the curve given by:

$$p_k \propto \frac{1}{k^3}$$

The plots required in this question are shown below:




(a) Log-Log degree distribution for Preferential Attachment graph with m = 1, n = 1000



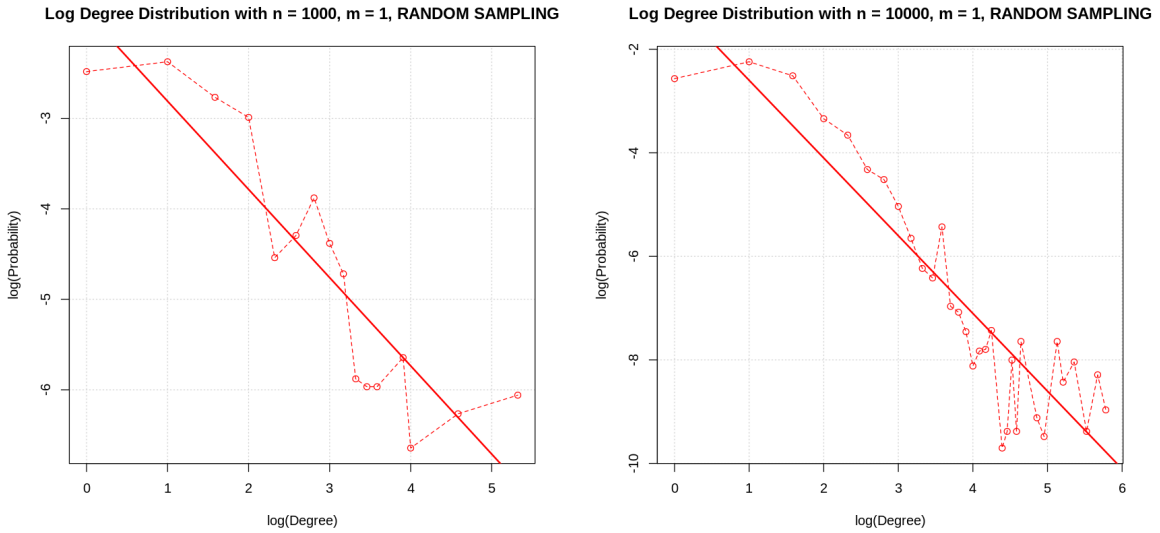

(b) Log-Log degree distribution for Preferential Attachment graph with m = 1, n = 10000

Figure 9: Degree distributions on a log-log scale

The slopes were calculated by linear regression using the function lm(). The slope for the PA graph with n = 1000 was found to be **-2.4048** and the slope for the PA graph with n = 10000 was found to be **-2.862.** This indicates that graphs with more nodes tend to follow the distribution more closely.

### Question 2 e

In this question, we are asked to repeat (d), but this time, we randomly sample nodes. The plots obtained are shown below. The slopes were calculated by linear regression using the function lm(). The slope for the PA graph with n = 1000 was found to be **-0.9752** and the slope for the PA graph with n = 10000 was found to be **-1.501**. Although the distributions can be crudely classified as linear(both decrease), they are highly unstable. This indicates that randomly sampling nodes is not a reliable method to find the degree distribution. This procedure can be interpreted as a random walk. As stated in class, at least ln(n) steps are required to converge on to a node with a steady-state degree, where n is the number of nodes. So a single step is insufficient for this interpretation and we do not go to a steady-state node.




(a) Log-Log degree distribution for Preferential Attachment graph (Random Sampling) with m = 1, n = 1000

(b) Log-Log degree distribution for Preferential Attachment graph (Random Sampling) with m = 1, n = 10000

Figure 10: Degree distributions on a log-log scale(Random Sampling)

### Question 2 f

In this question, we plotted a graph that shows the relationship between a node's age and its expected degree. The plot is shown below:

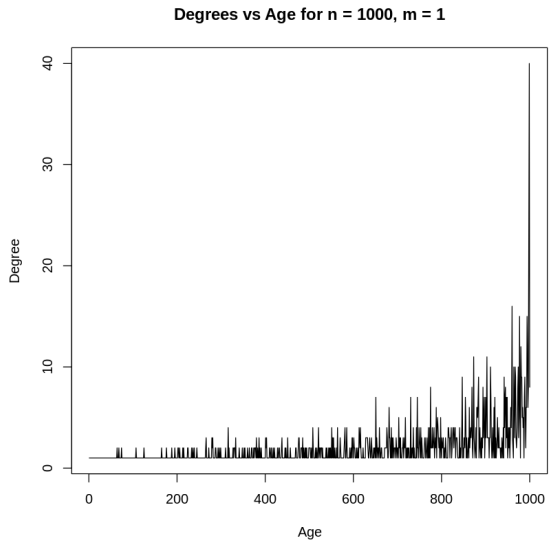



Figure 11: Expected degree vs age of node for an undirected Preferential Attachment graph with n = 1000, m = 1

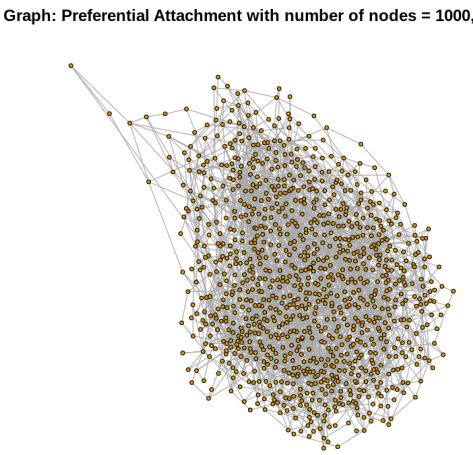
As expected, we see that a node's degree increases as it gets older as by construction, older nodes tend to have a be preferred for connections, leading to a higher degree.

The expected degree of a node added at time i is estimated as:

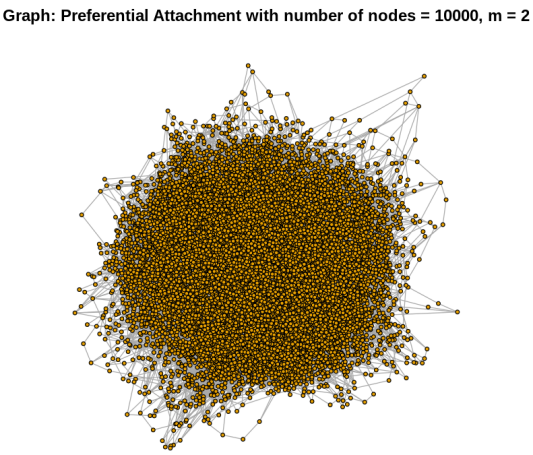
$$k = m\sqrt{\frac{t}{i}}$$

**Question 2 g**

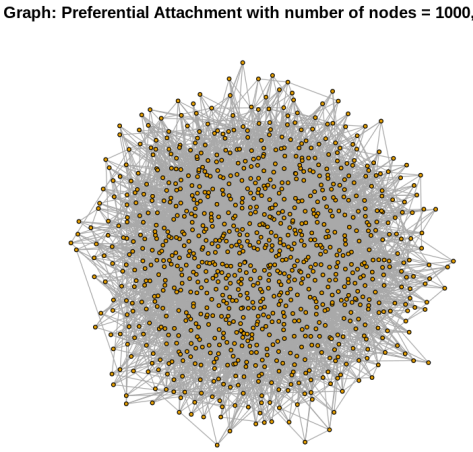
The four random graphs generated using preferential attachment are shown below:



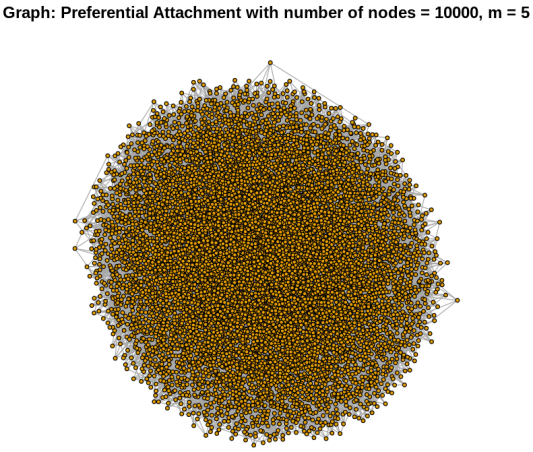

(a) Preferential Attachment graph with m = 2, n = 1000




(b) Preferential Attachment graph with m = 2, n = 10000



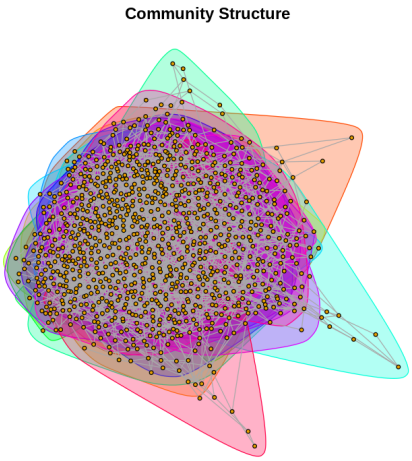

(c) Preferential Attachment graph with m = 5, n = 1000



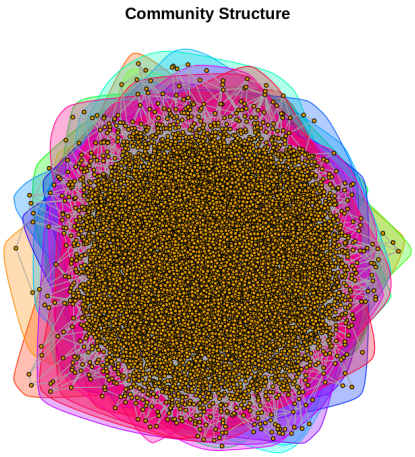

(d) Preferential Attachment graph with m = 5, n = 10000

Figure 12: Preferential Attachment Graphs

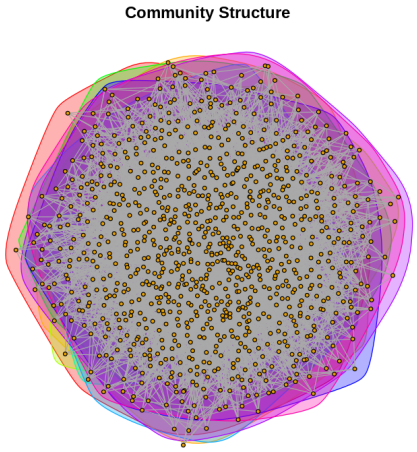
The four corresponding community structures are shown below:



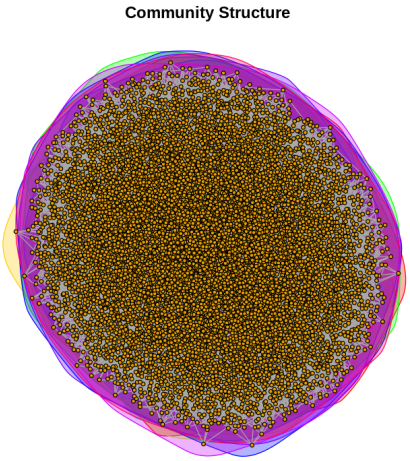

(a) Community Structure of Preferential Attachment graph with m = 2, n = 1000




(b) Community Structure of Preferential Attachment graph with m = 2, n = 10000




(c) Community Structure of Preferential Attachment graph with m = 5, n = 1000




(d) Community Structure of Preferential Attachment graph with m = 5, n = 10000

Figure 13: Community Structures of Preferential Attachment Graphs

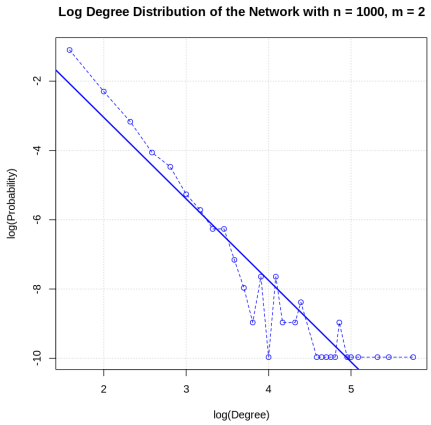
**The modularities were as follows:**

1. n = 1000, m = 2: Modularity = 0.52423
2. n = 10000, m = 2: Modularity = 0.52961
3. n = 1000, m = 5: Modularity = 0.27642
4. n = 10000, m = 5: Modularity = 0.27752

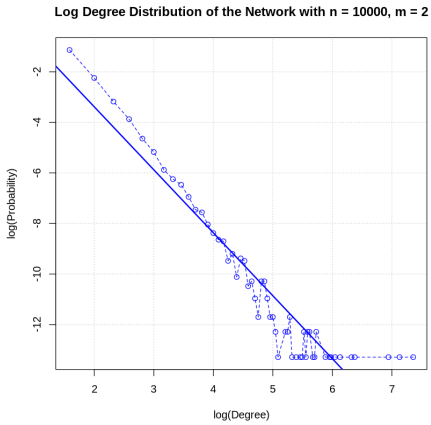
**We observe two important trends:**

1. We see that as n increases, the modularity score increases, meaning that the number of communities increases, and the larger graph also has bigger communities.
2. We see that as m increases, the modularity score decreases, meaning that the number of communities decreases, and each community has a larger number of nodes in it.

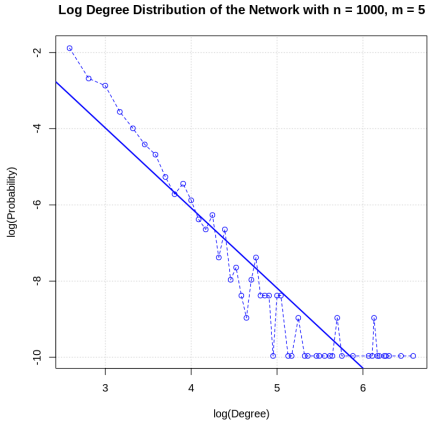
The following figure shows the corresponding degree distributions of the above graphs:



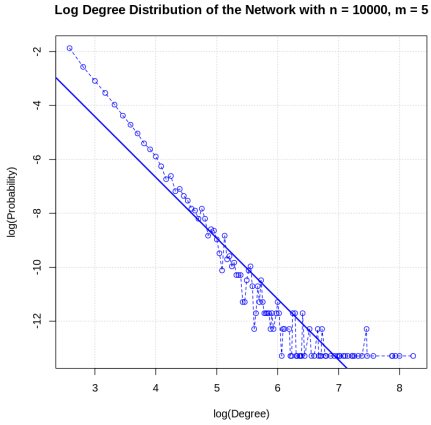

(a) Degree Distribution for Preferential Attachment graph with m = 2, n = 1000




(b) Degree Distribution for Preferential Attachment graph with m = 2, n = 10000




(c) Degree Distribution for Preferential Attachment graph with m = 5, n = 1000




(d) Degree Distribution for Preferential Attachment graph with m = 5, n = 10000

Figure 14: Degree Distributions for Preferential Attachment Graphs

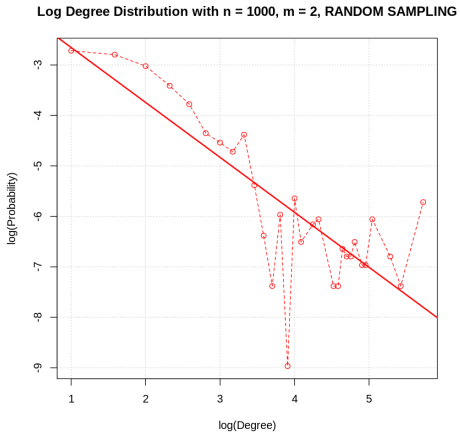
**The slopes were as follows:**

1. n = 1000, m = 2: Slope = -2.351
2. n = 10000, m = 2: Slope = -2.491
3. n = 1000, m = 5: Slope = -2.103
4. n = 10000, m = 5: Slope = -2.256

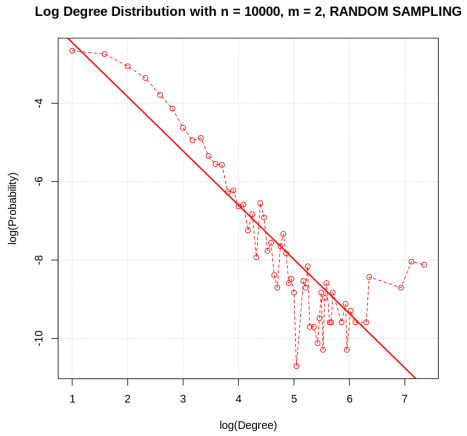
**All the plots are roughly linear on the log-log scale. We observe two important trends:**

1. For the same m, graphs with more nodes tend to follow the distribution more closely.
2. As m increases, the slope increases as well, that is, on average, the expected degree increases for a node. This is obvious, as we create more edges.

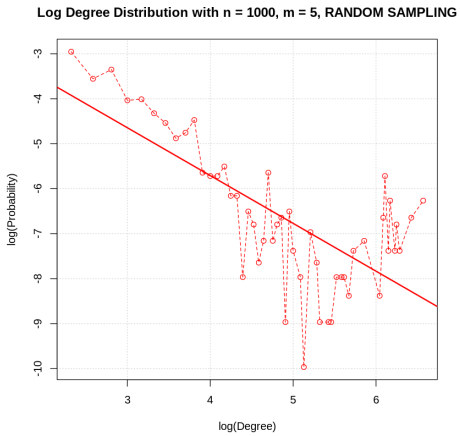
The following figure shows the corresponding degree distributions obtained through random sampling of the above graphs:



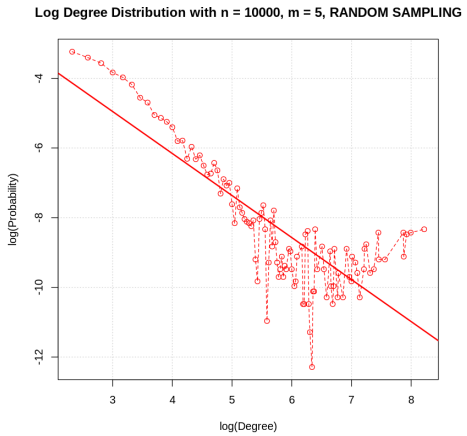

(a) Degree Distribution (Random Sampling) for Preferential Attachment graph with m = 2,
n = 1000




(b) Degree Distribution (Random Sampling) for Preferential Attachment graph with m = 2,
n = 10000




(c) Degree Distribution (Random Sampling) for Preferential Attachment graph with m = 5,
n = 1000




(d) Degree Distribution (Random Sampling) for Preferential Attachment graph with m = 5,
n = 10000

Figure 15: Degree Distributions (Random Sampling) for Preferential Attachment Graphs

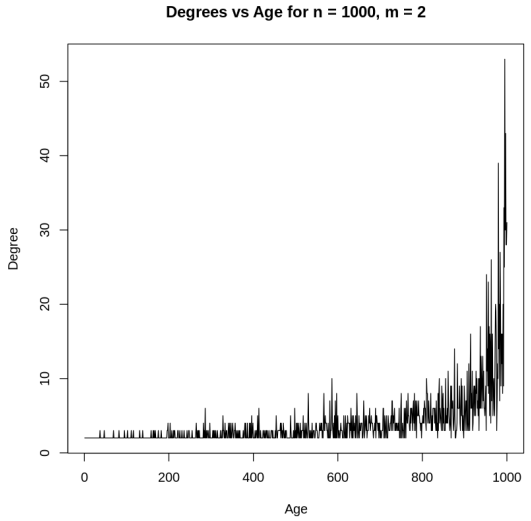
**The slopes were as follows:**

1. n = 1000, m = 2: Slope = -2.351
2. n = 10000, m = 2: Slope = -1.383

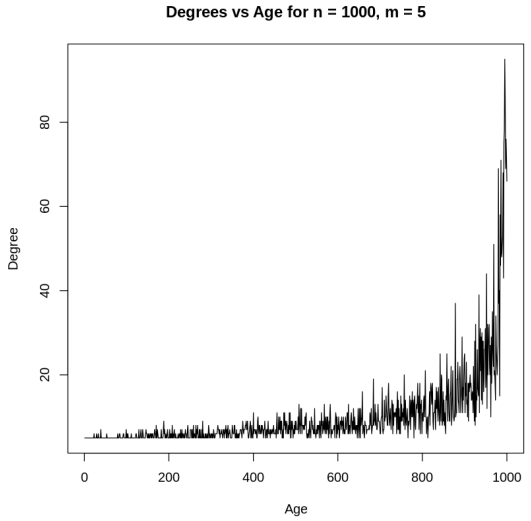
3. n = 1000, m = 5: Slope = -1.063
4. n = 10000, m = 5: Slope = -1.205

**We observe that as m increases, the slope increases as well, meaning that the degree distribution increases(shifts to the right). This is as expected, as the expected degree increases due to the fact that we bring in more edges to the graph.**

The graphs containing a node's age and its expected degree is plotted below for n = 1000, m = 2, 5:




(a) Expected degree vs age of node for an undirected Preferential Attachment graph with n = 1000, m = 2



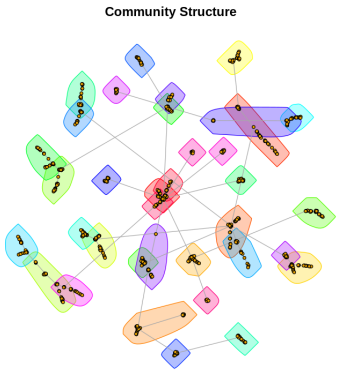

(b) Degree Distribution (Expected degree vs age of node for an undirected Preferential Attachment graph with n = 1000, m = 5

**As expected we observe that as m increases, the expected degree for a node of the same age increases as well. This happens because we bring in more edges to the graph as m increases.**

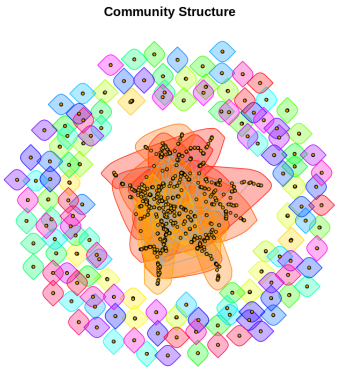
**Question 2 h**

In this question, we are asked to use a stub matching procedure to generate a graph, record its degrees, and generate a completely new graph with the same degrees. We used the function **sample_degseq** for this purpose.

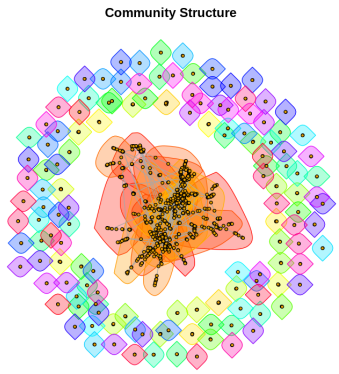
The generated graphs and their community structures are shown below:



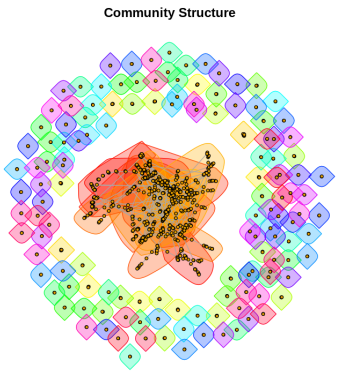

(a) Original Preferential Attachment graph with m = 1, n = 1000




(b) Preferential Attachment graph (Viger-Latapy) with m = 1, n = 1000




(c) Preferential Attachment graph (Simple, No Multiple) with m = 1, n = 1000




(d) Preferential Attachment graph (Simple) with m = 1, n = 1000

Figure 17: An undirected preferential attachment graph and its stub-matched graphs and their community structures

Their modularities were recorded as follows:

1. Original: 0.93253
2. Viger-Latapy: 0.83146
3. Simple, No Multiple: 0.83146
4. Simple: 0.83146

We make the following observations:

1. A large number of nodes are unconnected in graphs generated using stub-matching procedures.
2. The number of communities are more in stub-matched graphs.
3. Modularity of graphs generated using the preferential attachment model have a higher modularity.

**Question 3 a**

The graph generated using the preferential attachment model with age penalization is shown below:

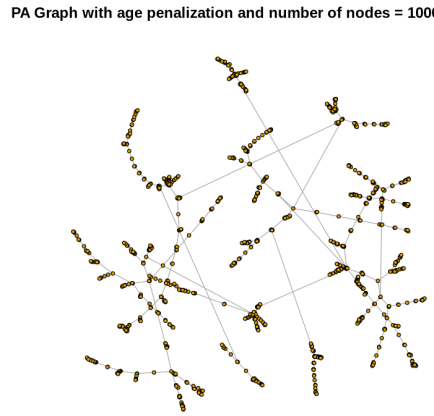



Figure 18: Preferential Attachment graph with m = 1, n = 1000 with age penalization

The degree distribution of the above graph is shown below:

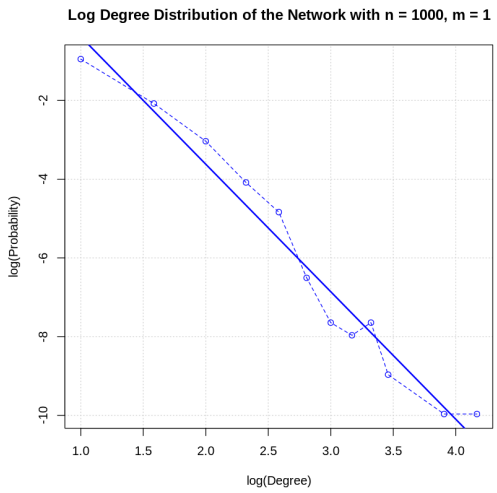



Figure 19: Degree Distribution of Preferential Attachment graph with m = 1, n = 1000 with age penalization

The slope of the linear regression line that estimates the log-log degree distribution was found to be equal to **-3.239**. This is approximately equal to the power law exponent value of 3 that is expected for graphs generated using a preferential attachment model.

### Question 3 b

The community structure of the graph generated using the preferential attachment model with age penalization is shown below:

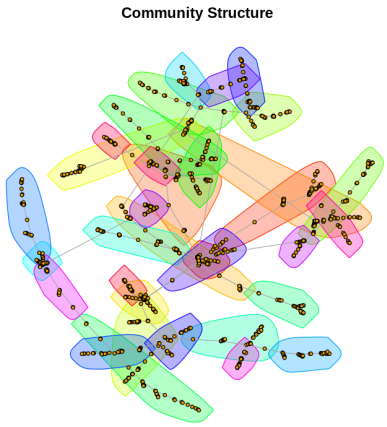



Figure 20: Community Structure of Preferential Attachment graph with m = 1, n = 1000 with age penalization

The modularity of the graph was found to be **0.93518** which is very similar to a preferential attachment graph without age penalization. However, from inspection of the figure, we see that the communities are very well separated. This is expected, as the model discourages new nodes to form edges with nodes that are too old.

# Part 2 - RANDOM WALK ON NETWORKS

### Question 1 a

In this part, we created an undirected random network with 1000 nodes, and the probability $p$ for drawing an edge between any pair of nodes equal to 0.01. The network is presented in Fig. 21.

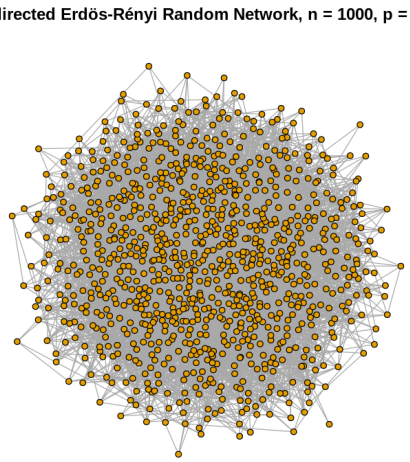
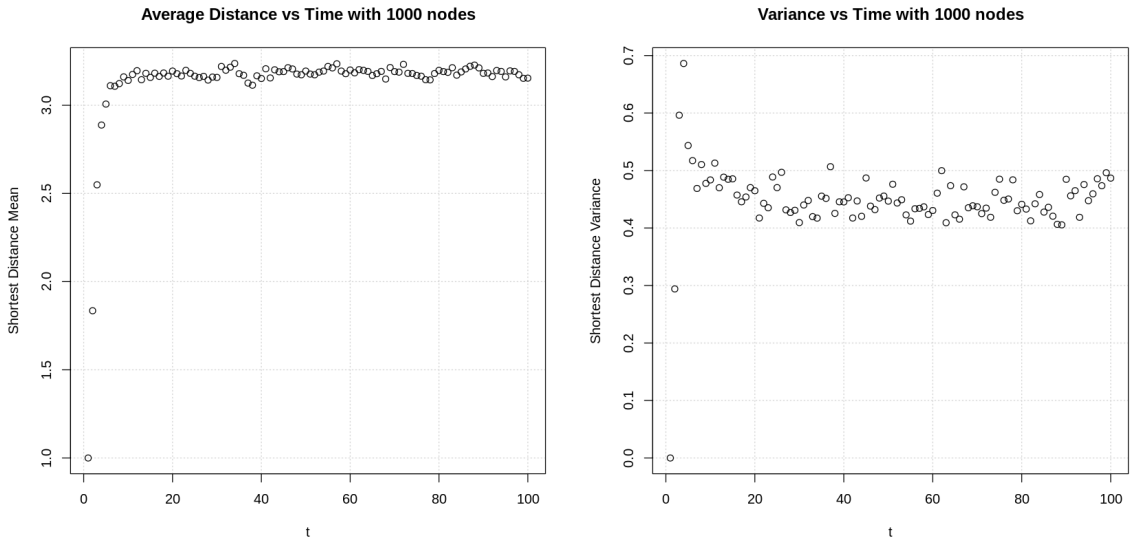



Figure 21: Undirected Erdös-Rényi Random Network, n = 1000, p = 0.01

### Question 1 b

in this part, we let a random walker start from a randomly selected node. The number of steps that the walker has taken is denoted by $t$. The shortest path length $\langle s(t) \rangle$, denoted as the average distance and the variance $\sigma^2(t) = \langle (s(t) - \langle s(t) \rangle)^2 \rangle$ of the distance are measured.

The average distance and the variance of the distance are presented in Fig. 22, with Fig. 22a representing the average distance against $t$ and Fig. 22b representing the variance against $t$.



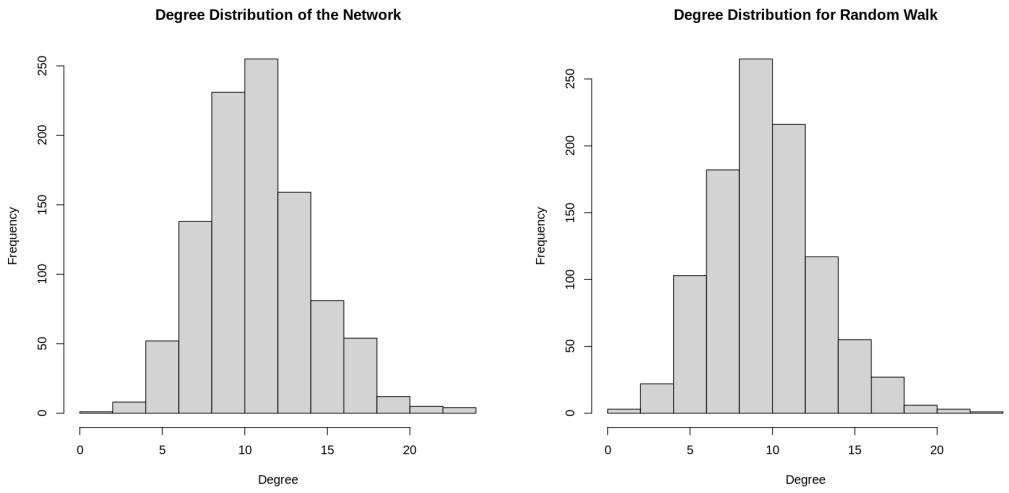

(a) Average Distance vs t with 1000 nodes (b) Variance of Distance vs t with 1000 nodes

Figure 22: Average Distance and Variance vs t with 1000 nodes

As is presented in Fig. 22, the average distance increases steeply at the initial stage where $t$ is closer to 0, and then it converges to a steady value with $\langle s(t) \rangle \approx 3.15$. The variance $\sigma^2(t)$ changes sharply on the inception and then oscillates between 0.4 and 0.5. We obtained the **diameter of the network d = 6**.

### Question 1 c

In this section, we compared the degree distribution between that of the nodes reached at the end of the random walk and that of the network. The result is given in Fig. 23 with Fig. 23a being the degree distribution of the network and Fig. 23b being that of random walk ending node.



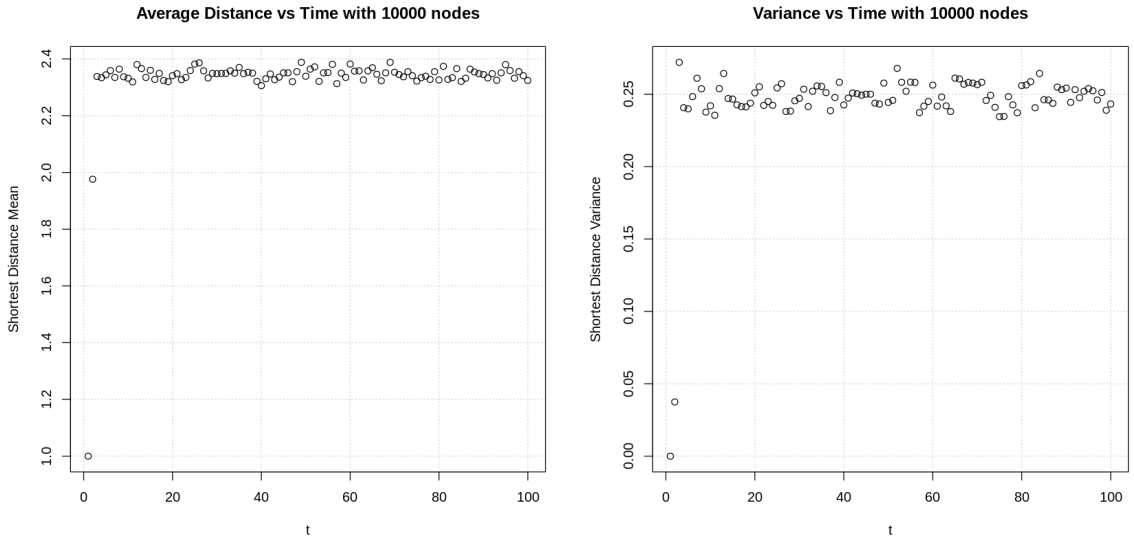

(a) Degree Distribution of the Network (b) Degree Distribution for Random Walk

Figure 23: Degree Distribution with 1000 nodes

It is observed that the degree distribution of both cases are similar with similar mean and variance, indicating that the distribution of the random walk ending node follows that of the entire network.

**Question 1 d**

In this part, we increased the number of nodes to 10000, and repeated the same process as previous steps. Similarly, we let a random walker start from a randomly selected node. The number of steps that the walker has taken is denoted by $t$. The shortest path length $\langle s(t)\rangle$, denoted as the average distance and the variance $\sigma^2(t) = \langle (s(t) - \langle s(t)\rangle)^2\rangle$ of the distance are measured. The results are plotted in Fig




(a) Average Distance vs t with 10000 nodes

(b) Variance of Distance vs t with 10000 nodes

Figure 24: Average Distance and Variance vs t with 10000 nodes

We can see in Fig. 24, the average distance increases steeply at the initial stage where $t$ is closer to 0, and then it converges to a steady value with $\langle s(t)\rangle \approx 2.35$. The variance $\sigma^2(t)$ changes sharply on the inception and then oscillates between 0.2 and 0.3. We obtained the **diameter of the network d = 3**.

The diameter of the network ($d = 3$) with more nodes (10000) is shorter than that (5) with less nodes (1000), and we can observe from the plots that the average distance is also smaller (2.35 with 10000 nodes compared with 3.15 with 1000 nodes), and the variance is also smaller (0.2-0.3 with 10000 nodes compared with 0.4-0.5 with 1000 nodes). This is explained by the compactness of the connected nodes in a larger network. Therefore, the distance for the walker is shorter than that of a smaller network. Since the connected nodes in the larger network are closer and more compact, the variance is smaller than that of a smaller network. In here, larger network means networks with more nodes and smaller network means that with less nodes.

### Question 2 a

We generated an undirected graph using the preferential attachment model with n = 1000, m = 1. The graph is displayed below:

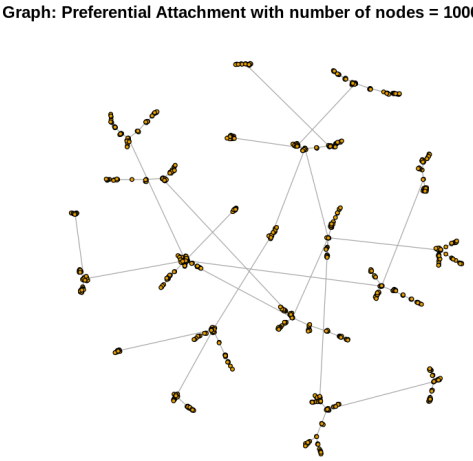
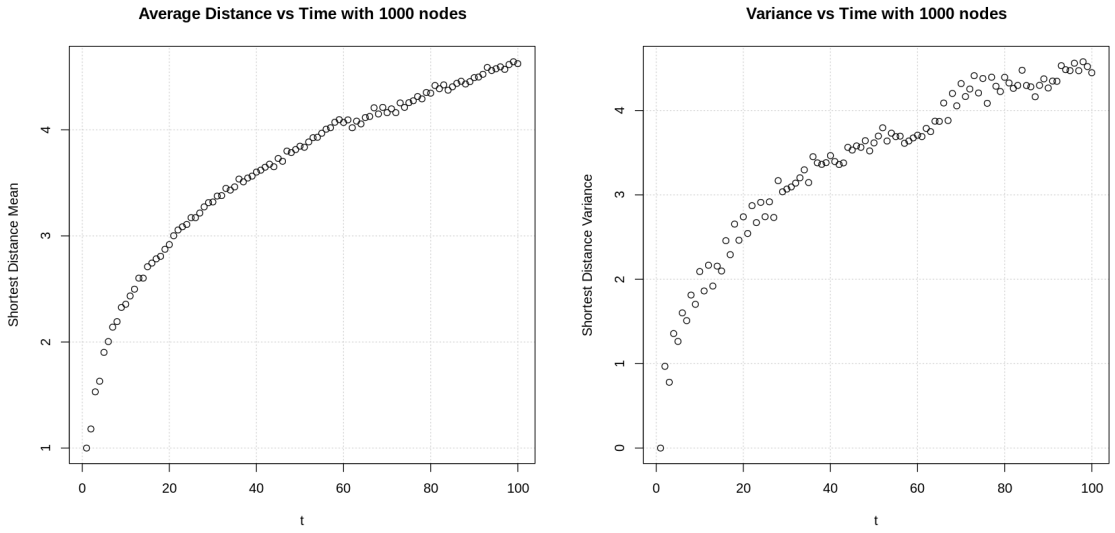



Figure 25: Undirected Preferential Attachment Network, n = 1000, m = 1

### Question 2 b

We performed random walks 1000 times with different initializations. The plots of mean distance and variance of a random walker vs time t is shown below:




(a) Average Distance vs t

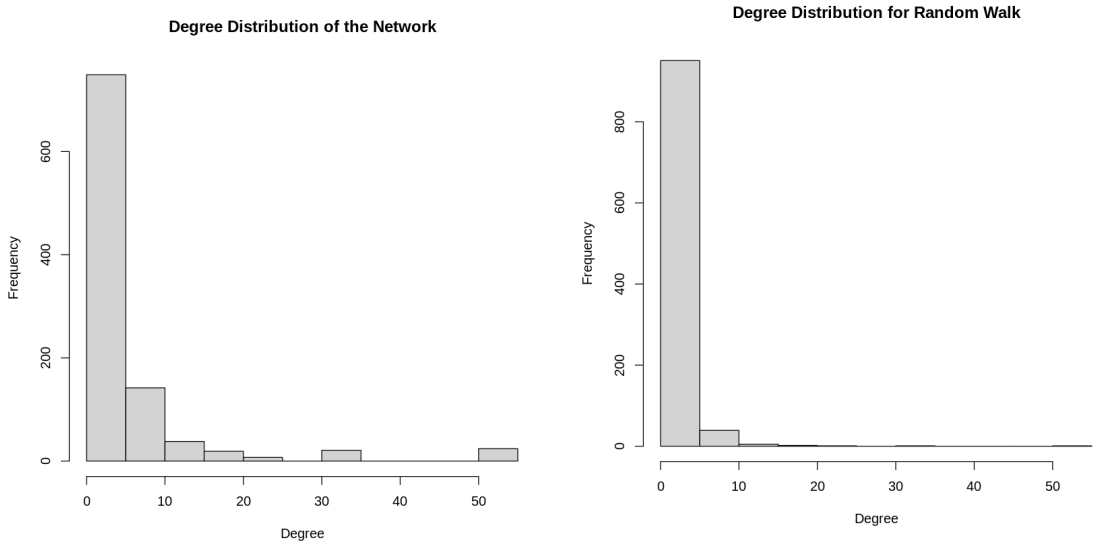
(b) Variance of Distance vs t

Figure 26: Average Distance and Variance vs t For Preferential attachment network with n = 1000, m = 1

We observe that both the mean and variance increase as time increases and it seems to be reaching a constant value(the rate of increase decreases as we take more steps). This is expected because as we take more steps, we reach the steady state of Preferential Attachment graphs, and we have a higher chance of moving between nodes with higher degrees. **We observe that the diameter of the network is 18.**

### Question 2 c

The degree distributions for the entire graph and the path that we went on during a random walk are displayed below:




(a) Degree Distribution of entire graph

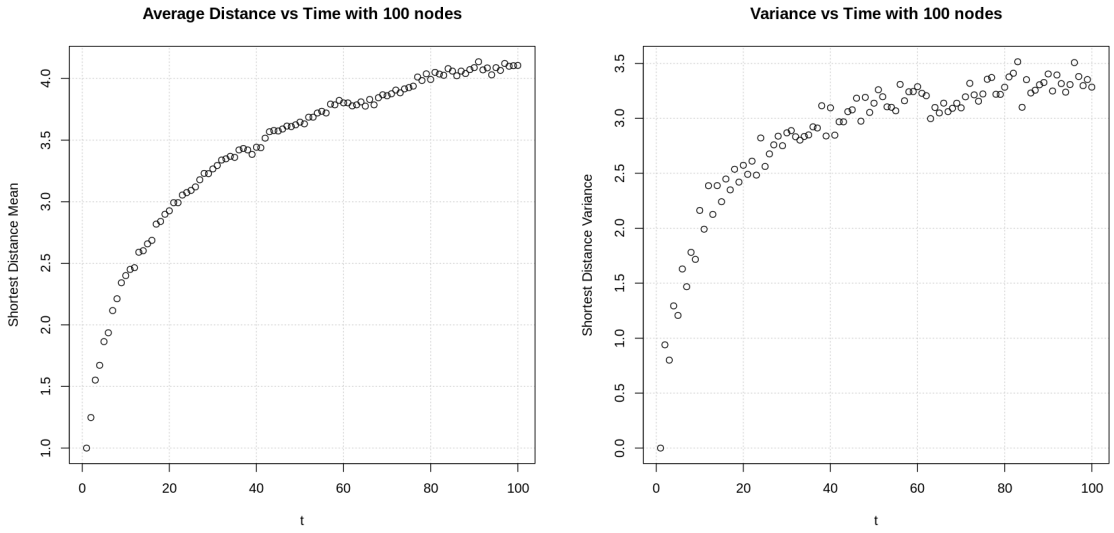
(b) Degree Distribution of nodes visited during random walk

Figure 27: Degree Distributions

We see that both the degree distributions are very similar and this is due to the nature of Barabasi networks and the way they are preferentially constructed (we reach a steady state after a certain number of steps).

**Question 2 d**

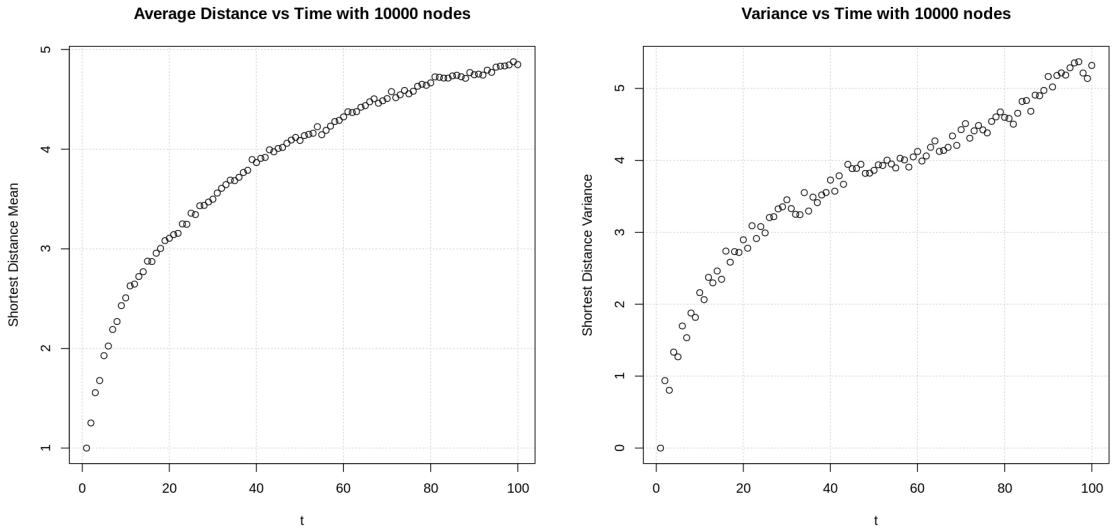
The plots of mean distance and variance of a random walker vs time t for a preferential attachment graph with n = 100, m = 1 is shown below:




(a) Average Distance vs t (b) Variance of Distance vs t

Figure 28: Average Distance and Variance vs t For Preferential attachment network with n = 100, m = 1

The plots of mean distance and variance of a random walker vs time t for a preferential attachment graph with n = 100, m = 1 is shown below:




(a) Average Distance vs t (b) Variance of Distance vs t

Figure 29: Average Distance and Variance vs t For Preferential attachment network with n = 10000, m = 1

The diameters of the networks were as follows:

1. n = 100, m = 1: diameter = 12
2. n = 1000, m = 1: diameter = 18
3. n = 10000, m = 1: diameter = 30

We make two observations:

1. As number of nodes increases, diameter increases. This is because bigger networks have more communities and better modularity, with very few connections between two communities, so the diameter is bigger.
2. As number of nodes increases, mean increases. This is because as the number of nodes increases, the probability of visiting a far away node with high degree also increases, thereby increasing the mean distance.
3. As number of nodes increases, variance increases. This is because as number of nodes increases, the uncertainty of traversing a path is greater, as we can have many more paths. This means that we will need more steps to reach steady state as well.

**Question 3 a**

In this part, we used random walk to simulate the PageRank algorithm [PBMW99]. The PageRank algorithm, used by the Google search engine, exploits the linkage structure of the web to compute global "importance" scores that can be used to influence the ranking of search results.

We first created a directed random network with 1000 nodes with $m = 4$, using the preferential attachment model. However, creating such a random network may have a risk of a walker runs into a "black hole". That is, the first node does not have outbounding edges, and the walker has no way to escape. To address this issue, we generated another random network with 1000 nodes and $m = 4$. We then merged these networks by adding the edges of the second graph to the first graph with a shuffling of the indices of the nodes. This way, all nodes are guaranteed to have at least one outbounding edge. Fig. 30 represents the degree distribution of the simulated webpage or random network.

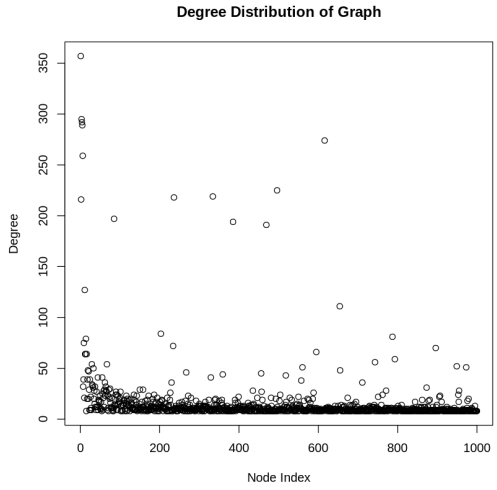
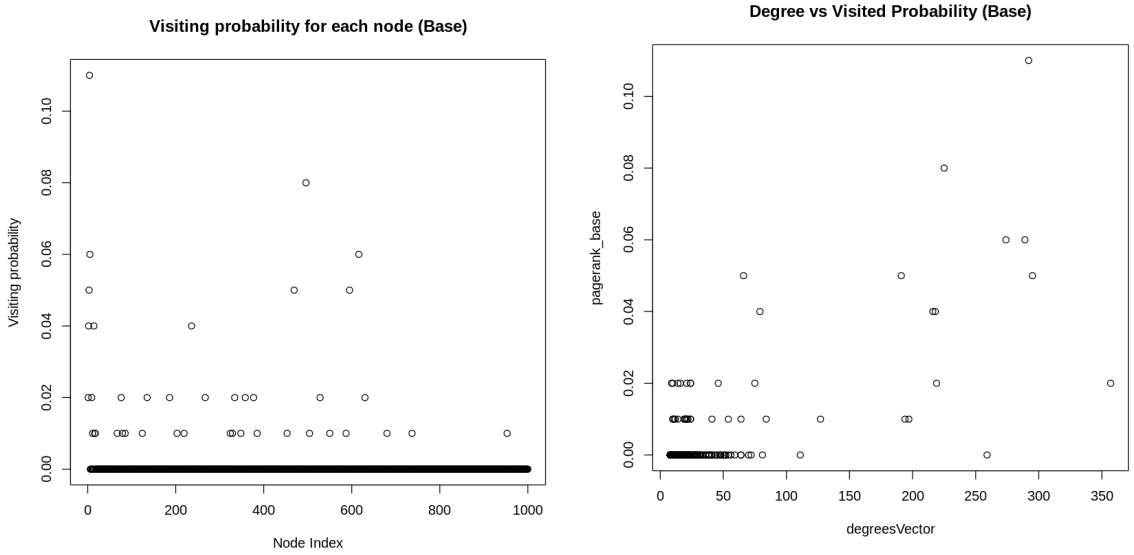



Figure 30: Degree distribution of the simulated network with m = 4

We run the walker for 500 steps and 100 iterations. The last node that the walker visits in each iteration is counted. The probability of being visited for each node is calculated by the number

of visits divide by total visits (i.e. 100). Fig. 31a and Fig. 31b show the results. The Pearson correlation coefficient for this relationship is 0.7407. A high correlation is expected as the more connected the node is, the higher the chance it gets visited.



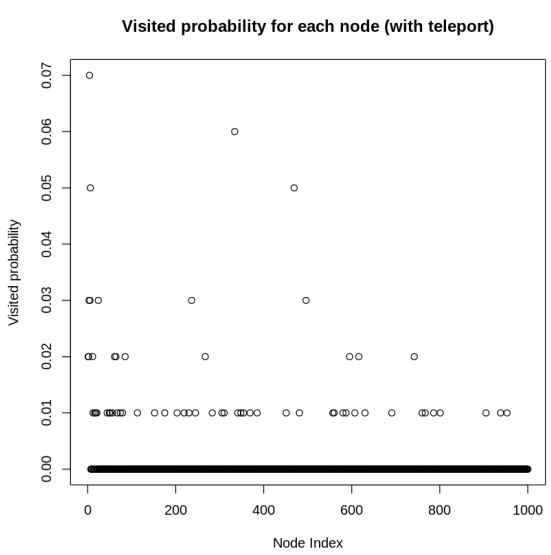

(a) Probability of walker randomly visiting each node

(b) Relationship between node degree of visited probability

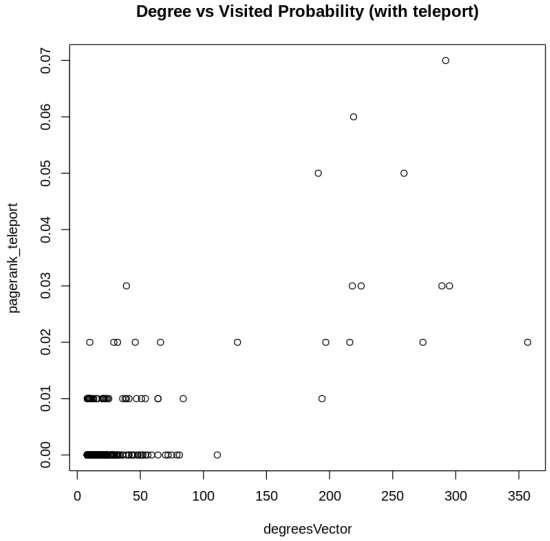
Figure 31: Probability of random walker visiting each node (Modified preferential attachment network, $n = 1000$, $m = 4$. No teleportation.)

**Question 3 b**

Using the same network in part 3a, we run the walker differently in this part with teleportation probability of $\alpha = 0.15$. That is, each time a walker has the probability of 85% runs to an adjacent connecting node, and 15% chance of 'teleporting' to another node randomly, with each node has equal chance to get visited. Fig. 32a shows the probability of node being visited with teleportation; and Fig. 32b shows the relationship between node degree and the visited probability. The Pearson correlation coefficient is 0.7554, which is higher than the previous experiment. Teleportation results in a higher correlation as the walker has is able to teleport to high-degree nodes.




(a) Probability of walker randomly visiting each node

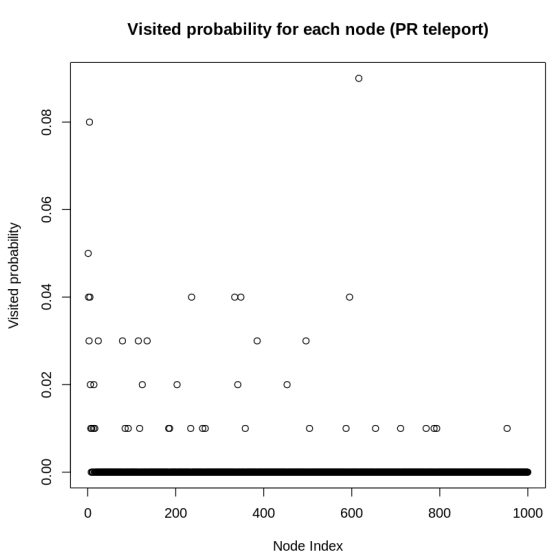



(b) Relationship between node degree of visited probability

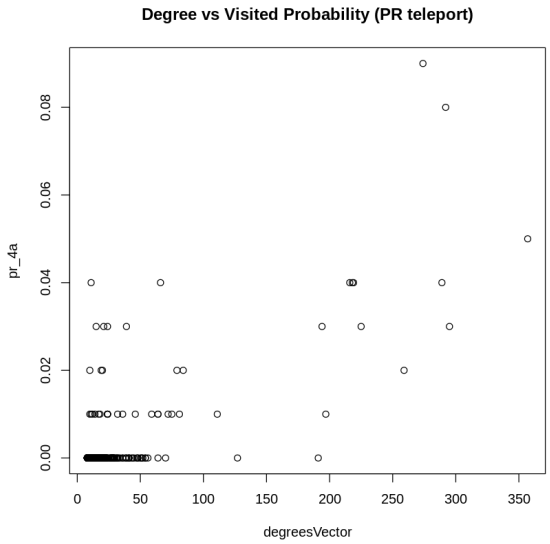
Figure 32: Probability of random walker visiting each node (Modified preferential attachment network, $n = 1000$, $m = 4$. Teleportation probability $\alpha = 0.15$.)

**Question 4 a**

In this question, we tried to simulate the Personalized PageRank. In other words, users are able to define their own notion of importance for each individual query. We used the same random network from parts 3a and 3b. However, the teleportation probability to each node is proportional to its PageRank from part 3a (as opposed to part 3b, where at teleportation, the chance of visiting all nodes are the same at $1/N$). Fig. 33a and Fig. 33b show the relationship between walker probablity and node degree in this scenario. The Pearson correlation coefficient is 0.7655, which is even higher than those in question 3. It is expected because PageRank algorithm marks higher probability to nodes with higher degree, accompanying with teleportation, the walker has even higher chance to teleport to a high-degree node than part 3a.




(a) Probability of walker randomly visiting each node

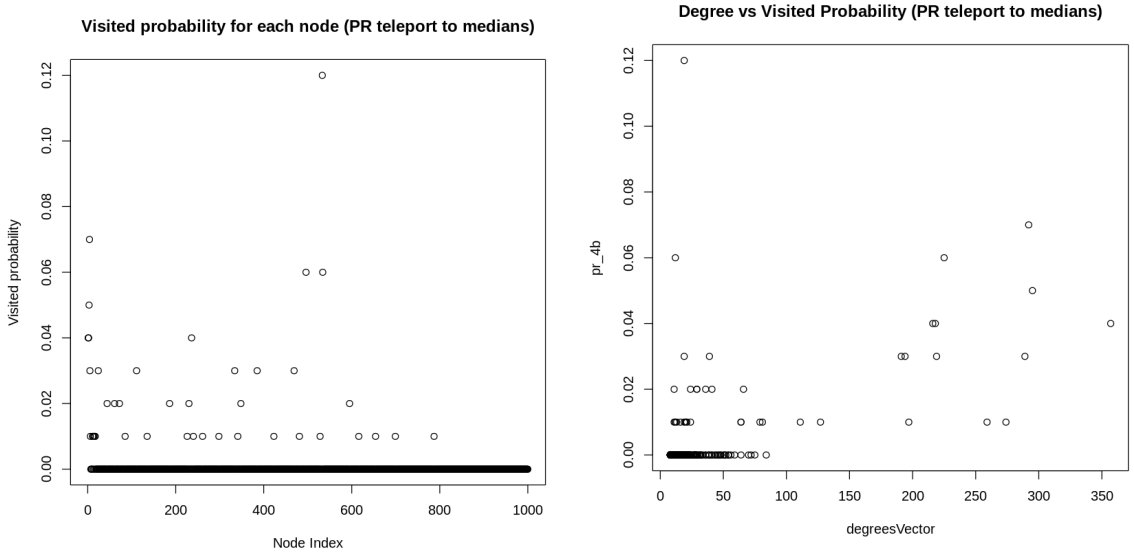



(b) Relationship between node degree of visited probability

Figure 33: Probability of random walker visiting each node (Modified preferential attachment network, $n = 1000$, $m = 4$. Teleportation probability $\alpha = 0.15$. Node teleportation probability is proportional to base PageRank.)

**Question 4 b**

Similar to part 4a, we experimented on node teleportation probability but keeping other configuration. In this case, instead of teleportation to any node, we only teleport to median nodes in PageRank algorithm. Fig. 34a and Fig. 34b show the relationship. The Pearson correlation coefficient is 0.6020, which is the lowest among PageRank experiments. It can be explained that as teleportation takes place, the walker has a much lower probability to reach high-degree nodes as the two choices it can teleport is the two median nodes. It is interesting that even with the teleportation, the walker still has a lower chance to visit high-degree nodes comparing to non-teleportation cases.




(a) Probability of walker randomly visiting each node (b) Relationship between node degree of visited probability

Figure 34: Probability of random walker visiting each node (Modified preferential attachment network, $n = 1000$, $m = 4$. Teleportation probability $\alpha = 0.15$. Node teleportation narrows to median nodes in PageRank algorithm.)

**Question 4 c**

First we need to identify the original PageRank equation. Let:

- A is the node-node incidence matrix of a directed graph.
- P is the node transition matrix for a random walker on the graph (with teleportation probability $\alpha$).
- $k_{out}(i)$ is the out-degree of node i.
- $P_{ij}$ is the element of $P$ representing the probability of node $j$ after the random walker has landed on node $i$.
- $n$ is the number of nodes in graph.

Hence the original PageRank equation:

$$P_{ij} = (1-\alpha)\frac{1}{k_{out}(i)}A_{ij} + \alpha\frac{1}{n} \tag{1}$$

Eqn. 1 assumes that people's interest in all nodes are the same. However, it is not the case in the real world, where a user browsing the web only teleports to a subset of trusted web pages. Let $T$ is a set of trusted nodes. The term $P_{ij}$ then changes into:

$$P_{ij} = \begin{cases} (1-\alpha)\frac{1}{k_{out}(i)}A_{ij} + \alpha\frac{1}{|T|}, i \in T \\ (1-\alpha)\frac{1}{k_{out}(i)}A_{ij}, \qquad\qquad i \notin T \end{cases} \tag{2}$$

At equilibrium, for every node $i$,

$$\pi(i) = \Sigma_{j=1}^{n} P_{ij}\pi(j) \tag{3}$$

The modified PageRank equation becomes:

$$\pi(i) = \begin{cases} (1-\alpha)\Sigma_{j=1}^{n}\frac{1}{k_{out}(j)}A_{ji}\pi(j) + \alpha\frac{1}{|T|}, i \in T \\ (1-\alpha)\Sigma_{j=1}^{n}\frac{1}{k_{out}(j)}A_{ji}\pi(j) \qquad\qquad , i \notin T \end{cases} \tag{4}$$

# Bibliography

[ER+60] Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.