Homework 2

EE232E - Graphs and Network Flows

Team:

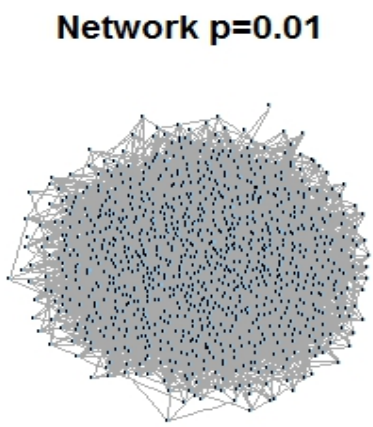
Rutuja Ubale (UID: 404558257)

Usha Amrutha N (UID: 204590772)

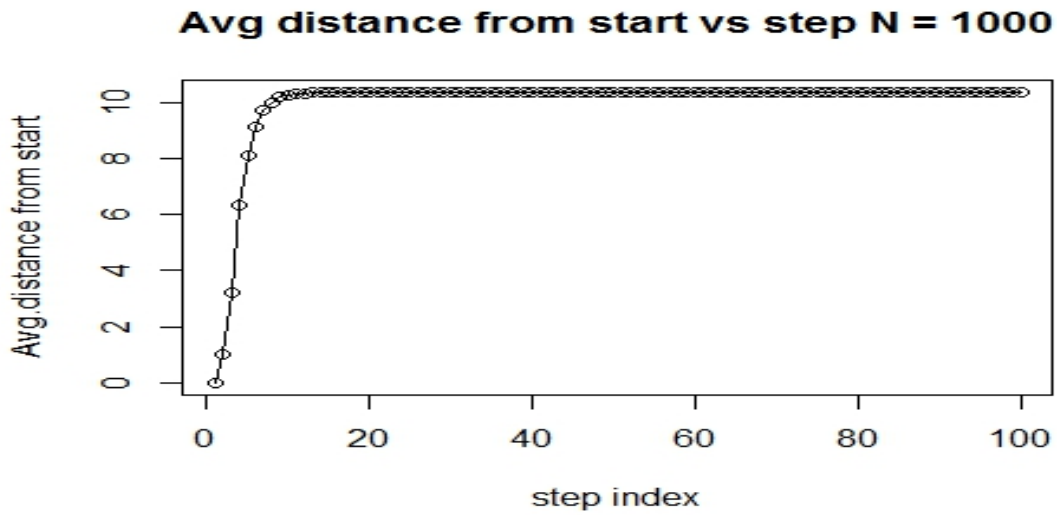
Pallavi Chakraborty (UID: 404519609)

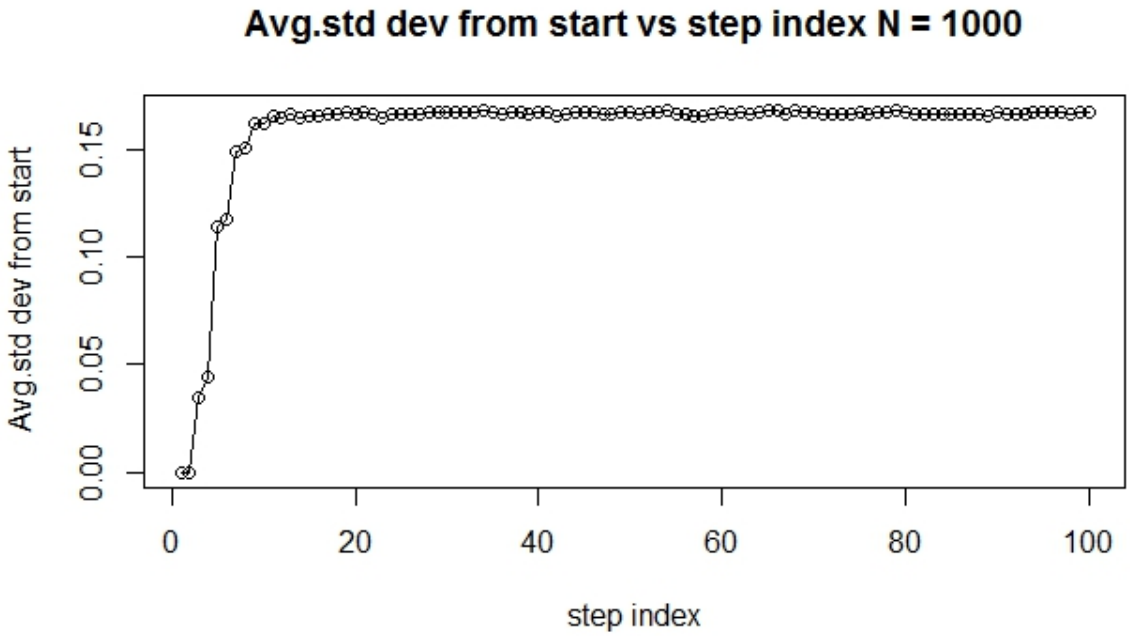
Question 1: Creating Random Networks

a.) An undirected graph was created with 1000 edges and by taking 0.01 as a probability of drawing an edge the graph was obtained as below:



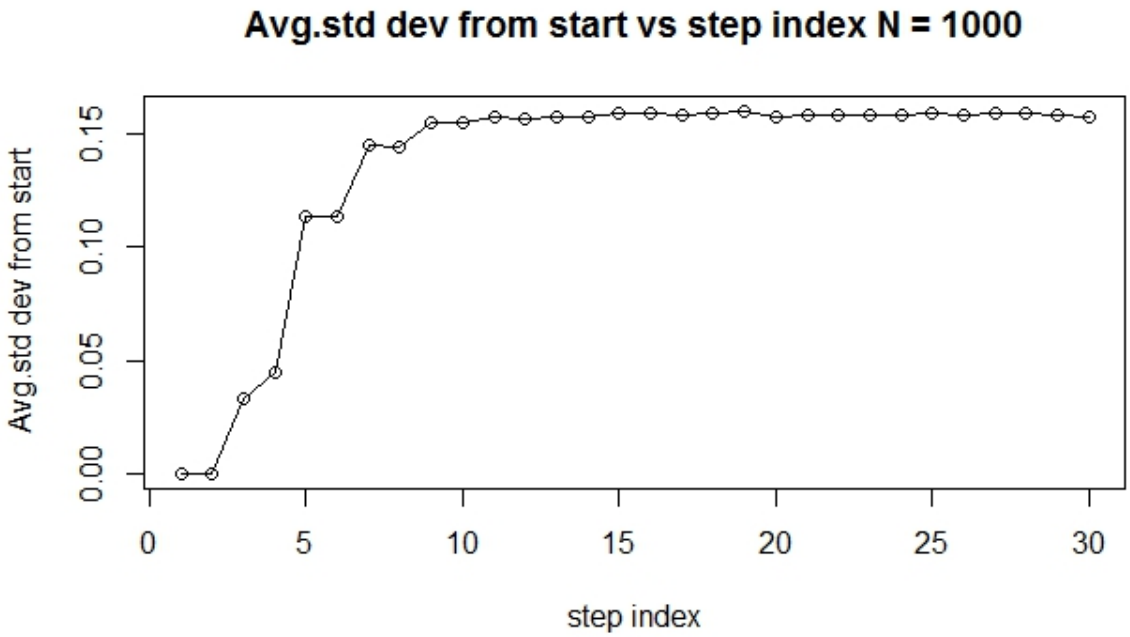
b.) Random walk was simulated on the random graph generated in question 1 part a. All possible starting nodes in the graph and different runs of the random walk were selected to perform random walk. Number of random walkers was set to 100. The standard deviation and the average distance <s(t)> of the walker from the starting point at step t is plotted below. Also, the shortest distance between was computed to measure the distance between two nodes. The diameter of the graph was found to be 6.





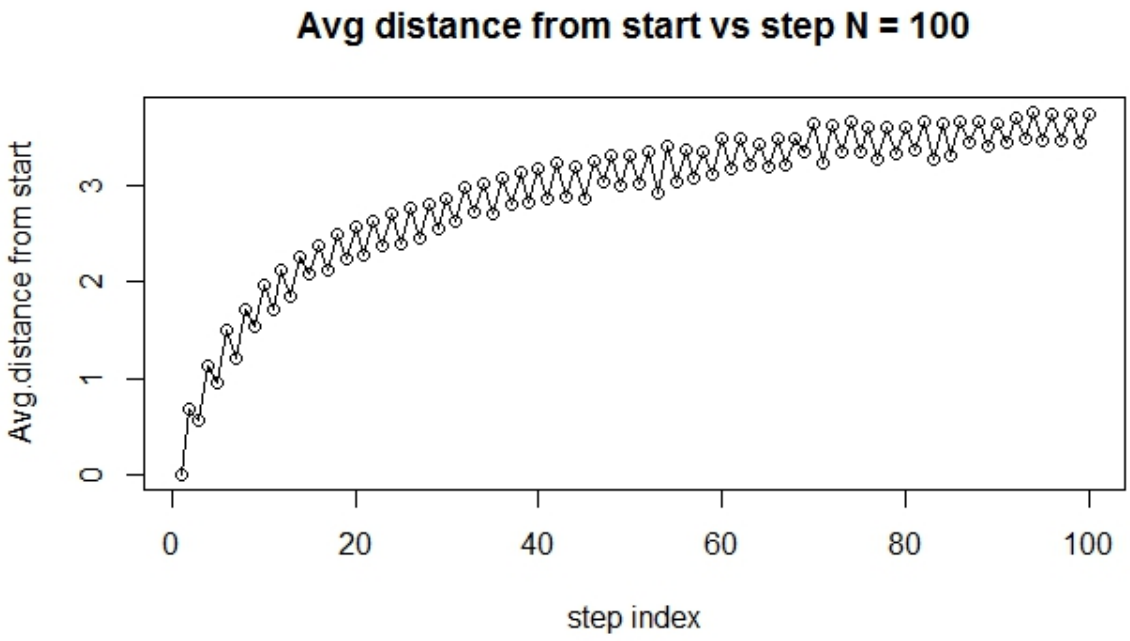
c.) It is given that in d dimensional lattice average distance is 0 and $\sqrt{\langle s(t)^2 \rangle}$ is proportional to $\sqrt{\ }$. However, the average distance measured in part b is not same as the dimensional lattice average because, it does not take into account the directionality (hence signed values). The average distance measured will always be a positive value.

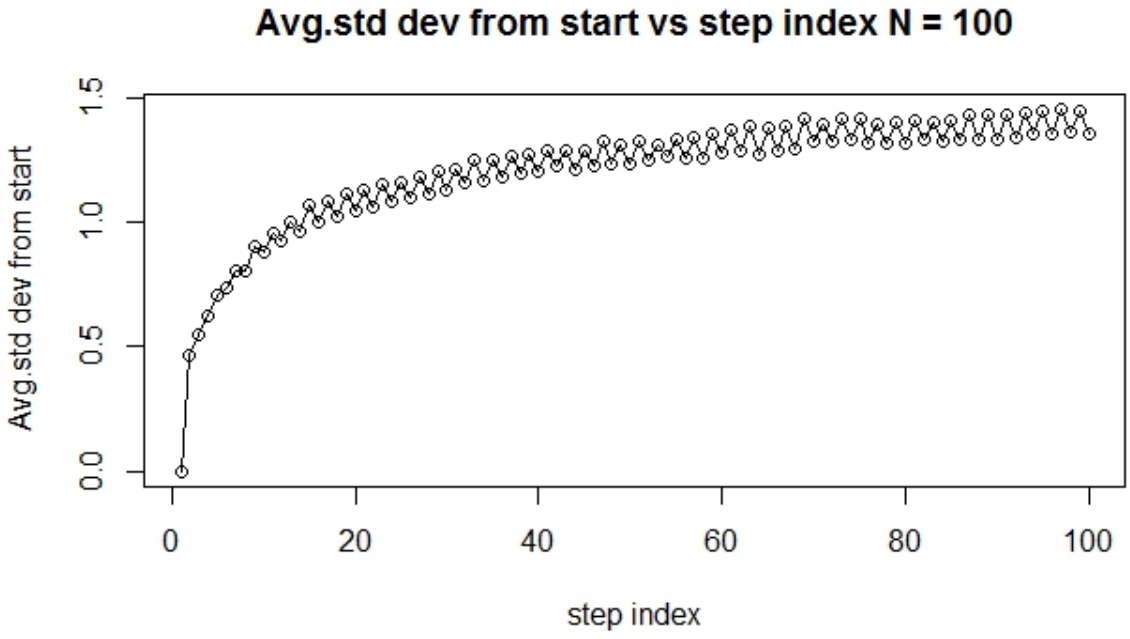
The standard deviation values are proportional to $\sqrt{\ }$ according to the d dimensional lattice structure. However, in part b the plot can be analysed as follows:
The average standard deviation values are linear to step index t until a threshold value after which the standard deviation values reach saturation.



d.)

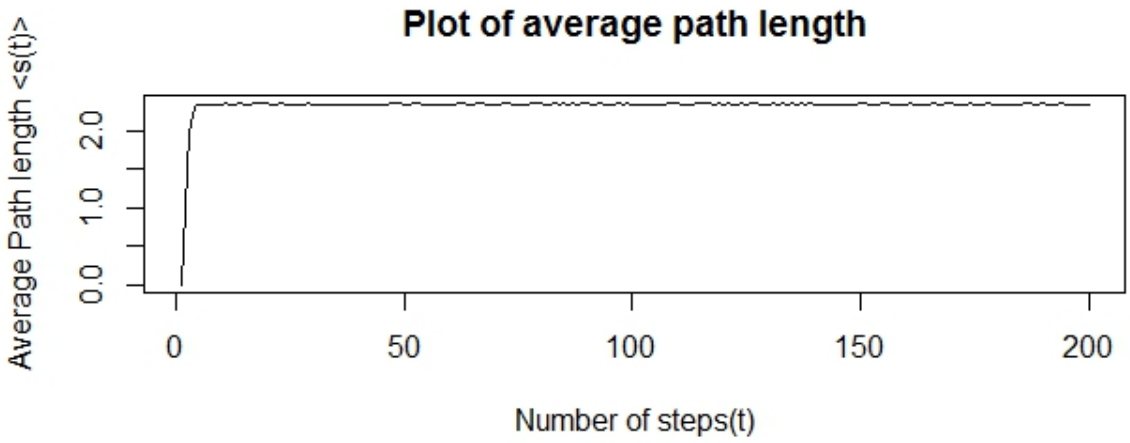
The random walk was repeated with 100 nodes and 10000 nodes in the network.

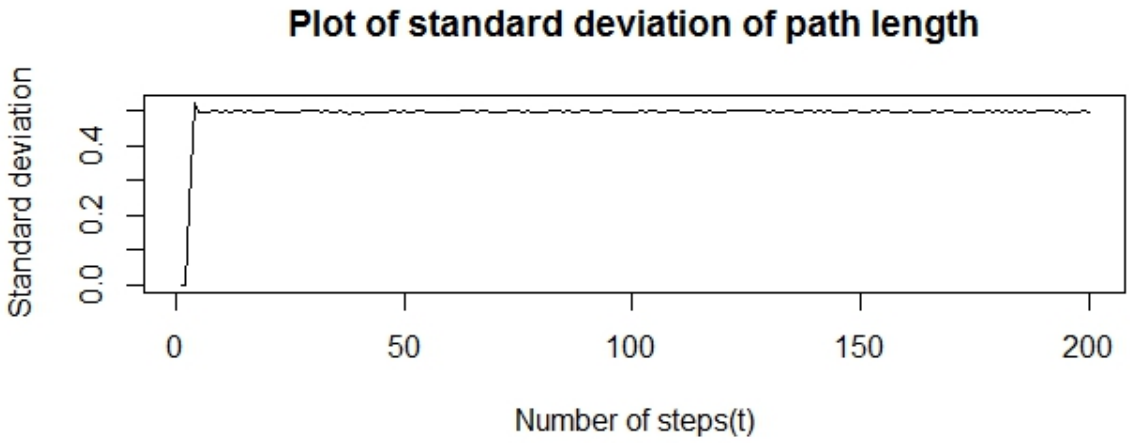




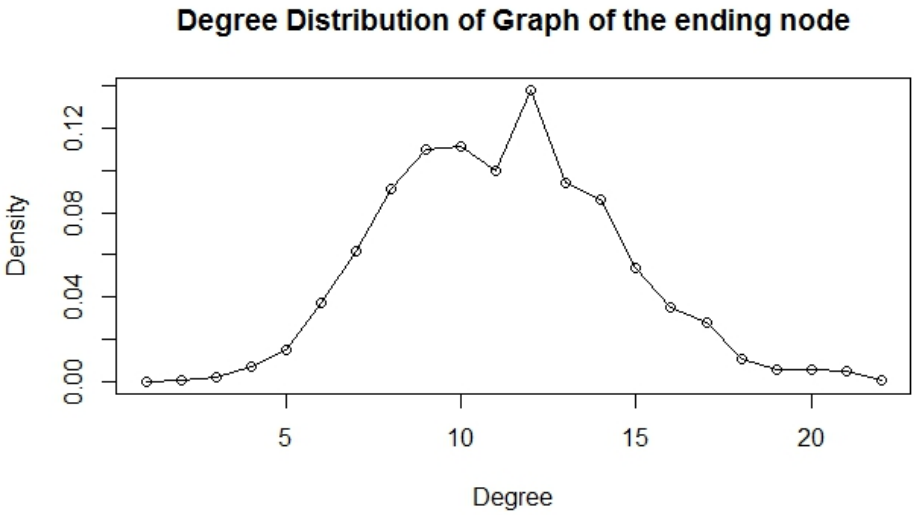
The diameter of the nodes N=100 is around 5. We see that the average distance moves forward and backward along the same path. This can also be attributed to the fact that the graph is highly disconnected. The standard deviation also shows similar variations for the same reasons.

The number of nodes in the experiment was set to 10000. The diameter was found to be around 3. From the average distance plot we can see that the average distance increases as the step index increases. However, the average distance saturates to a value proportional to the diameter of the network. In this case, it is approximately around 2.33. Also, we see that as the number of nodes increases the network becomes more densely connected and hence smaller values of the diameter can result. The average distance of start node from step t and average standard deviation are plotted as below:

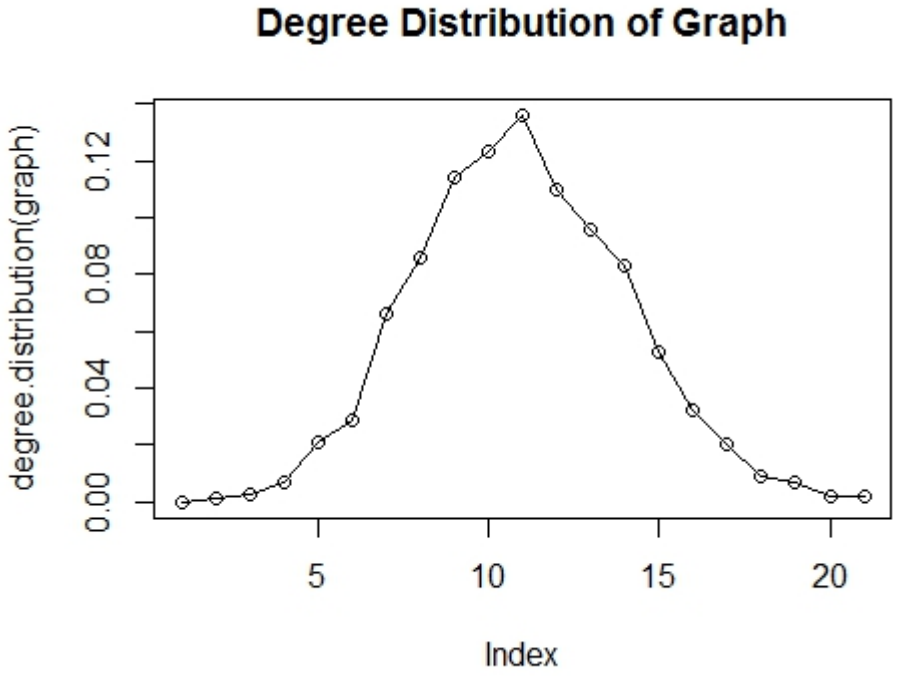




e.) The degree distribution and corresponding of the ending node of the random walk are as below



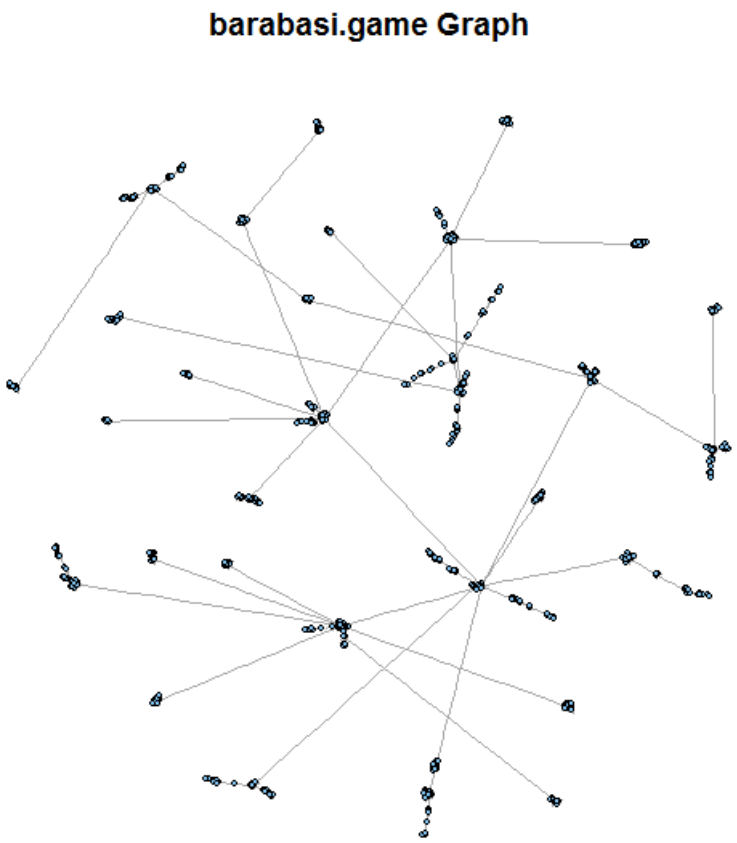
The actual degree of distribution of the graph is given below:



It can noted that the degree distribution of the graph is in line with the previous plot of the degree distribution of the ending node. The graph is well connected; the ending node in a random walk is equivalent to selecting a possibility of selecting a random node at the end of random walk. This corresponds to the actual degree distribution of the graph. Hence they are similar.

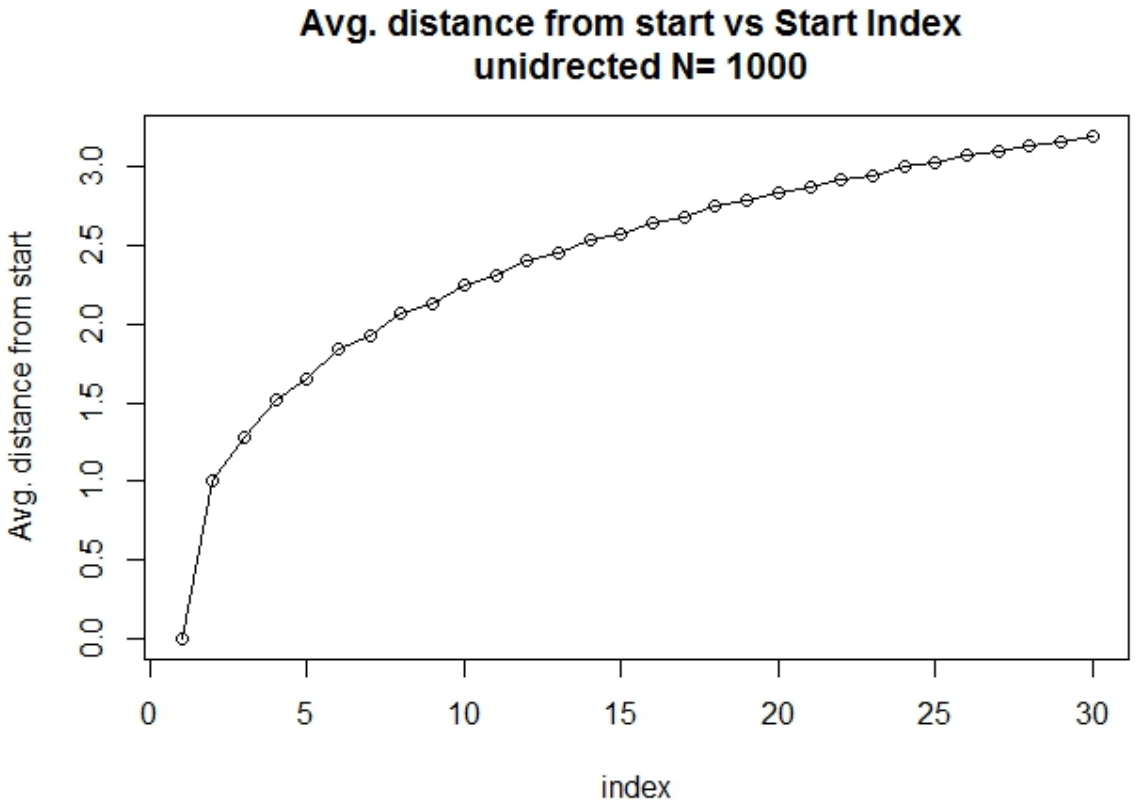
Question 2: Random walks on networks with fat-tailed degree distribution

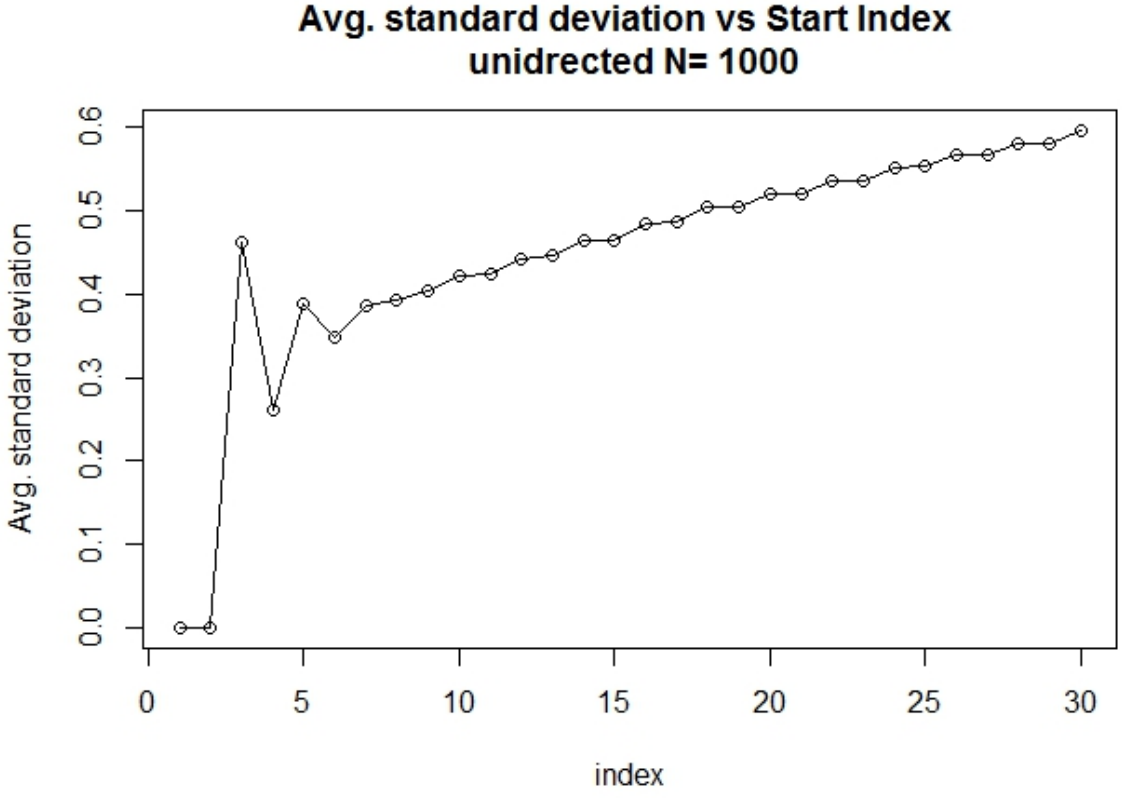
a) We used the barabasi.game function to generate an undirected graph with 1000 nodes.



b) We simulate a random walker on the graph and set damping factor to 1. Damping factor is the probability that an imaginary surfer will continue clicking on links. A damping factor of 1 would mean that the surfer doesn't stop clicking, however it is usually assumed to be 0.85.

We set number of random walkers to 100 (N) and time steps to 30 (t). Next, we plot the average distance from start vs the number of time steps as well as the average standard deviation vs the number of time steps in the graphs below.


Avg. distance from start vs Start Index
unidrected N= 1000
Avg. distance from start
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
25
30
index

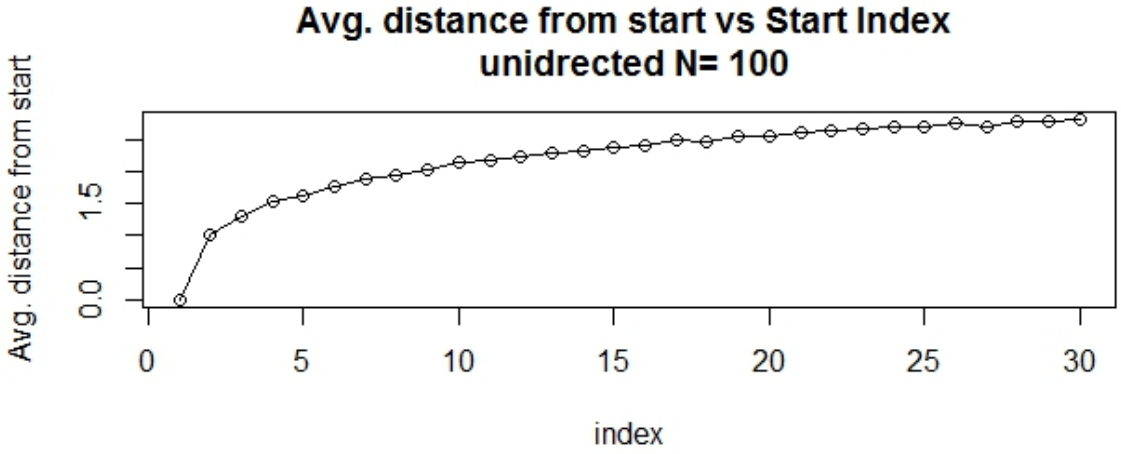

Avg. standard deviation vs Start Index
unidrected N= 1000
Avg. standard deviation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
index

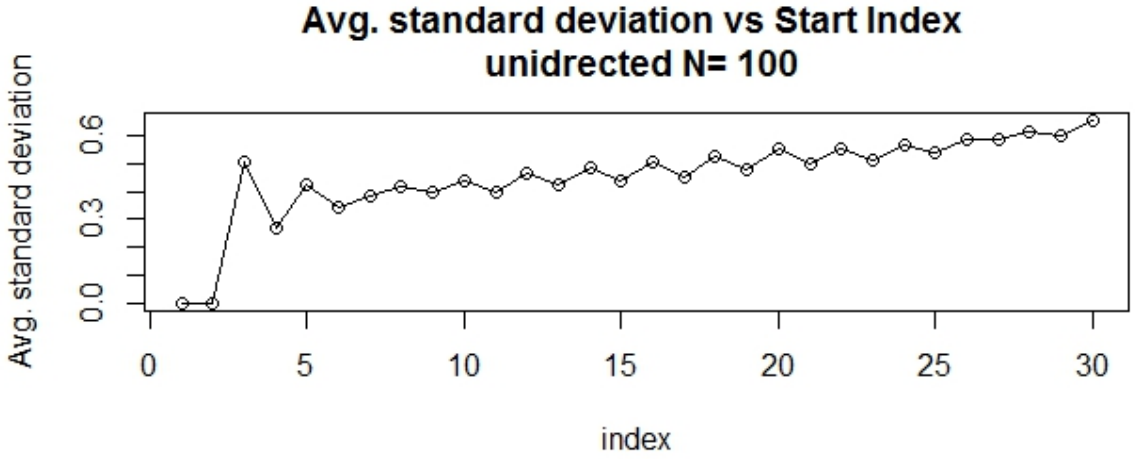
c) We compare the results of random walks in part b with results of random walks in d dimensional lattice. We know that the average (signed) distance is 0 for d dimensional lattice and √< s(t)^2 > is proportional to √t. The average distance obtained in part b is not 0 and hence is not like in a d dimensional lattice. It is possibly because we generate undirected graphs, and while finding the distance, direction is not considered, as a result of which it will always be a positive value. We can also see from the graph that the standard deviation is not always directly proportional to t and hence does not look like it is in the case of a d dimensional lattice.

d) We repeated our exercise in part (b) on the network we generated for 100 and 10000 nodes. The plots for average distance vs time steps and average standard deviation vs time steps for both 100 and 10000 nodes are shown below.

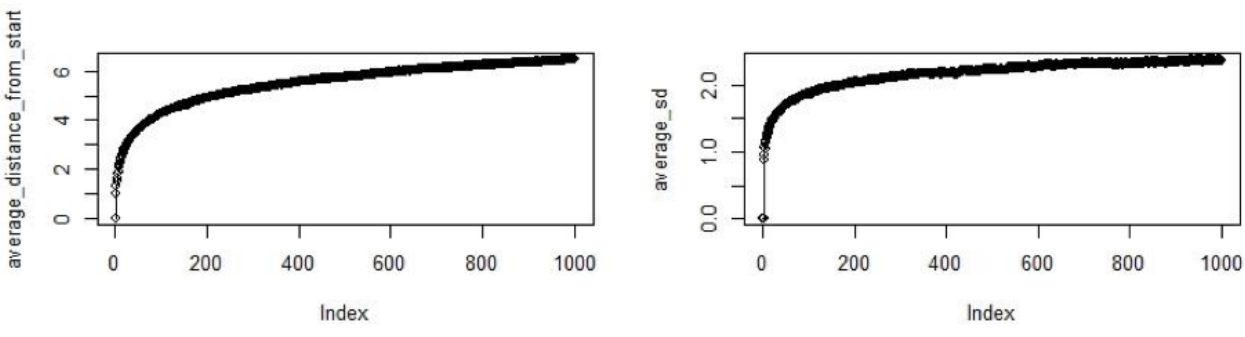
The diameter was found to be 10 for network with 100 nodes and the diameter for the network with 10000 nodes was found to be 23.

For 100 nodes:





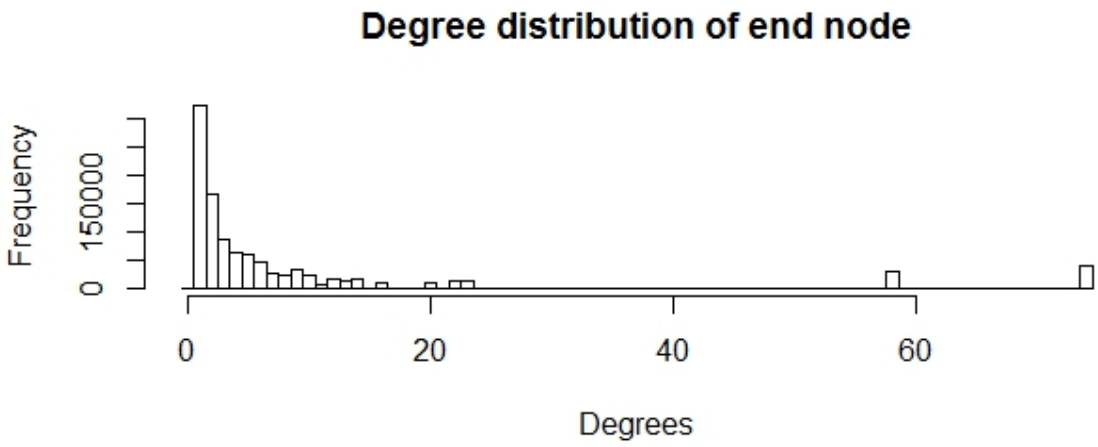
For 10000 nodes

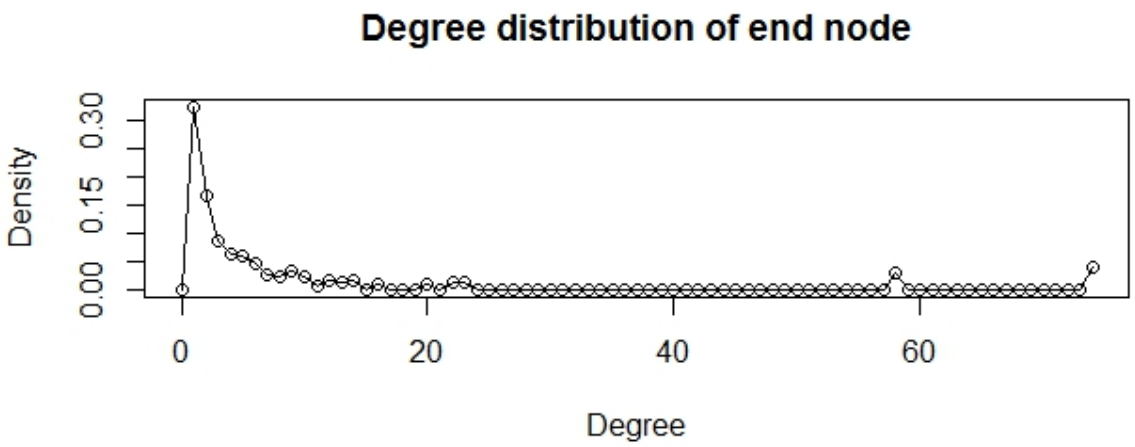


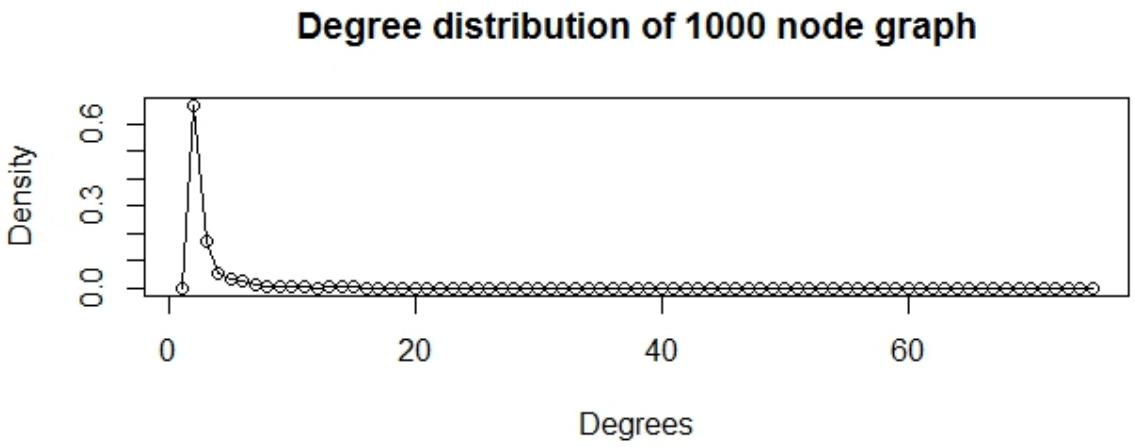
The average distance from the start node for network with 10000 nodes is linear with number of steps until a threshold is reached, beyond which the average distance saturates to a constant proportional to the diameter of the graph.

e.) We have plotted below the degree distribution of the end node reached during the random walk simulation in part 2b.

The degree distribution of the network follows a power law of x^{-3} as can be seen from the plot. The distribution of the end node is similar to that of the actual degree distribution of the network.

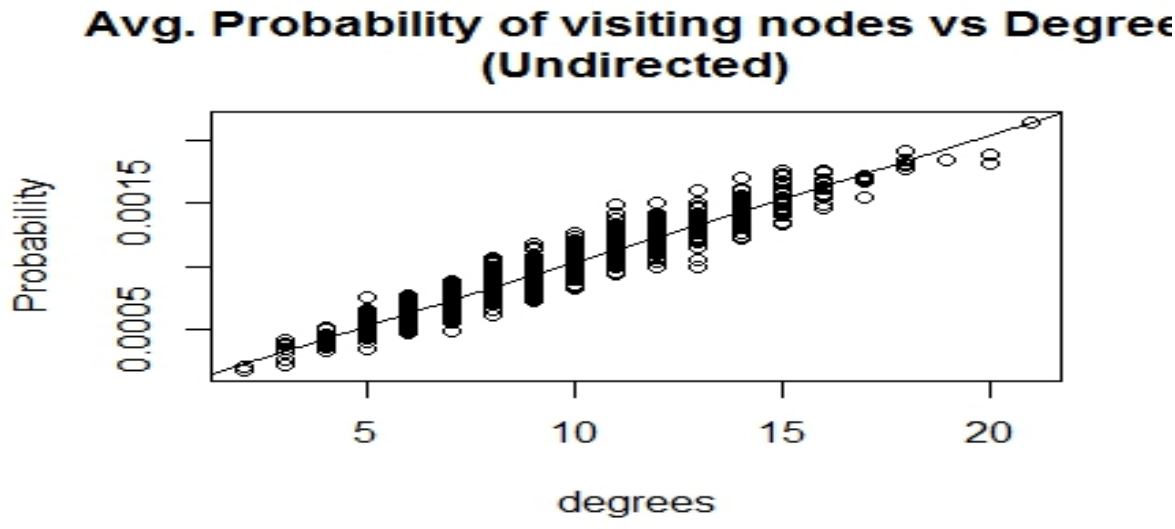






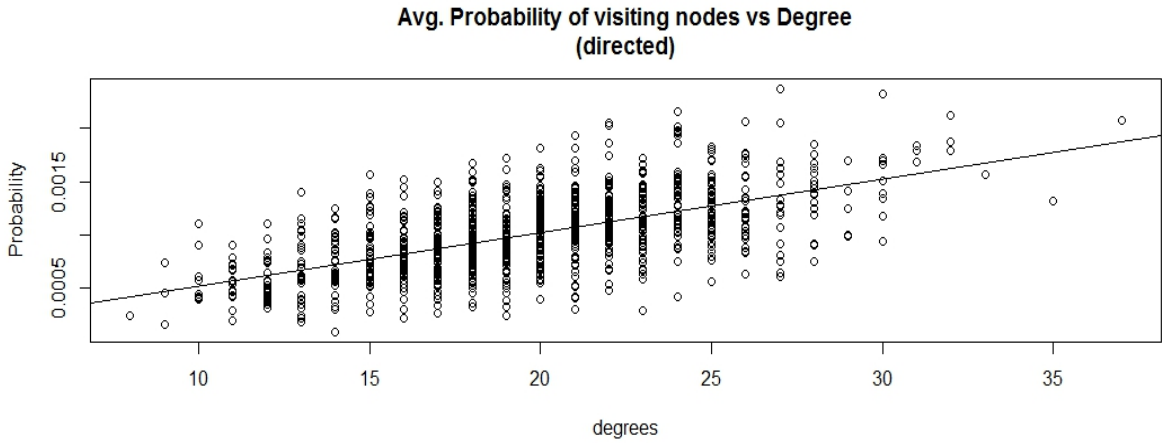
Question3: PageRank

a.) The plot for the probability of visiting node vs the degree of the node for an undirected graph shown below:



The average probability of visiting a node follows a linear relation and fits the linear regression model very well. Also we note that there are not many outliers. This can be attributed to the fact that the random graph generate is an undirected graph which highly connected. Therefore probability of reaching a node is also high.

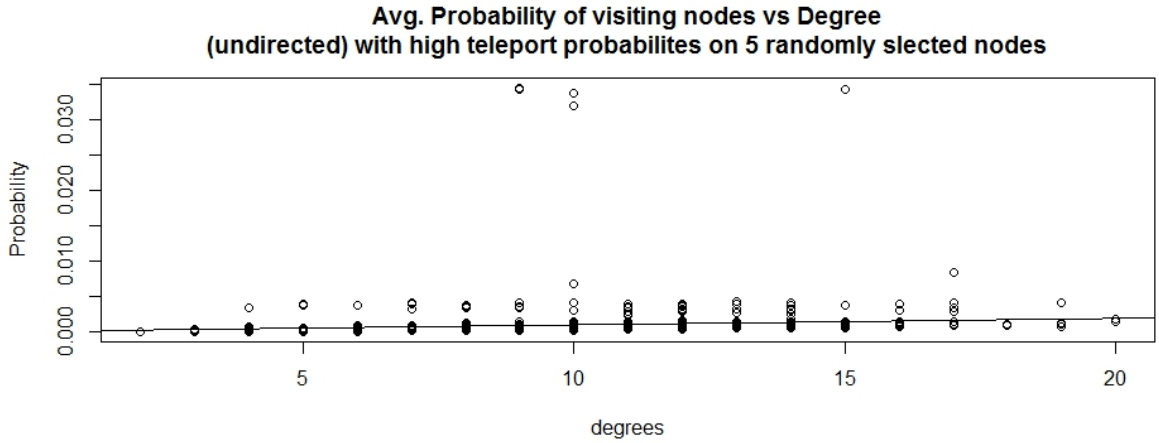
b.) The plot for the probability of visiting node vs the degree of the node for a directed graph shown below:



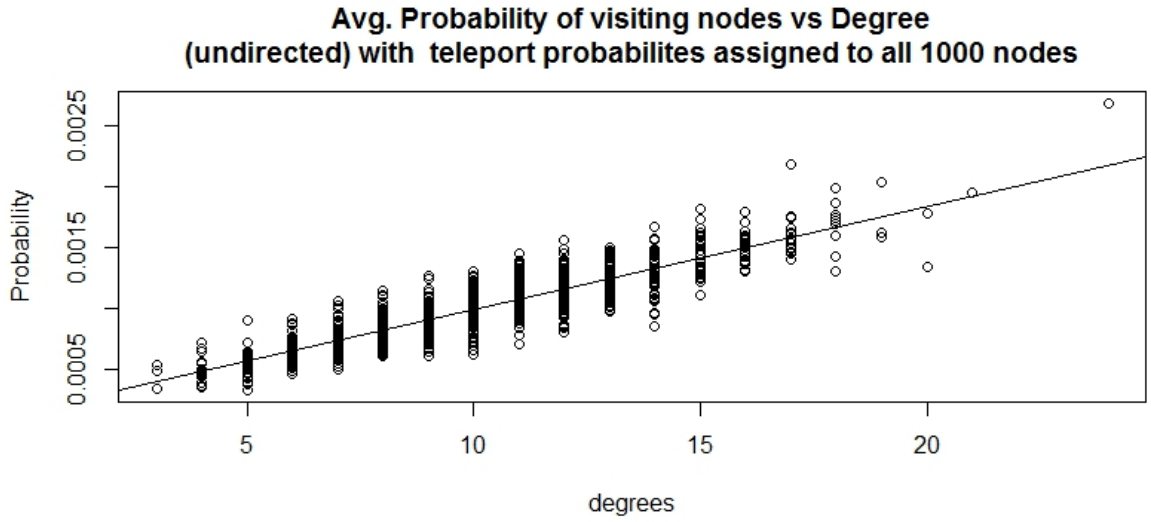
In this case since the graph is directed, the probability of reaching a node is dependent on in-degree distribution and not on the entire degree of the node. Therefore, the regression model does not fit well and has many residuals.

c.) The plot for the probability of visiting node vs the degree of the node for an undirected graph with damping factor 0.85 is as shown below.

In this experiment, we first set the teleport probability factor in netrw function to a vector containing 5 random 1s and 995 0's. We note that the average probability of reaching these nodes turns out to be very high in comparison to other nodes. This can be related to the fact that when the random walk reaches the end, it is teleported to one of the 5 nodes. Hence, the visiting probability of these nodes becomes very high.



The plot for the probability of visiting node vs the degree of the node for an undirected graph with damping factor 0.85 is as shown below:

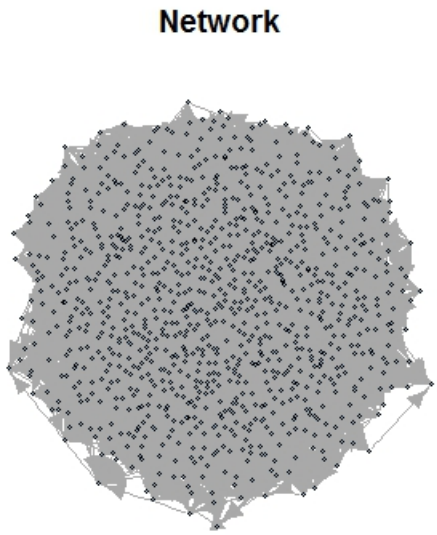


As the degree of a node increases, the probability of visiting it in a random walk also increases. However, the linear regression model has more residuals as compared to the undirected graph with no damping factor. This is because when there is damping factor, a node gets teleported to random node. Therefore, the degree of a node is only partially influential in deciding the average probability of visiting a node. If the damping factor is decreased, the teleportation probability increases, and hence the dependence of the probability of visiting the node with the degree of the node also decreases.

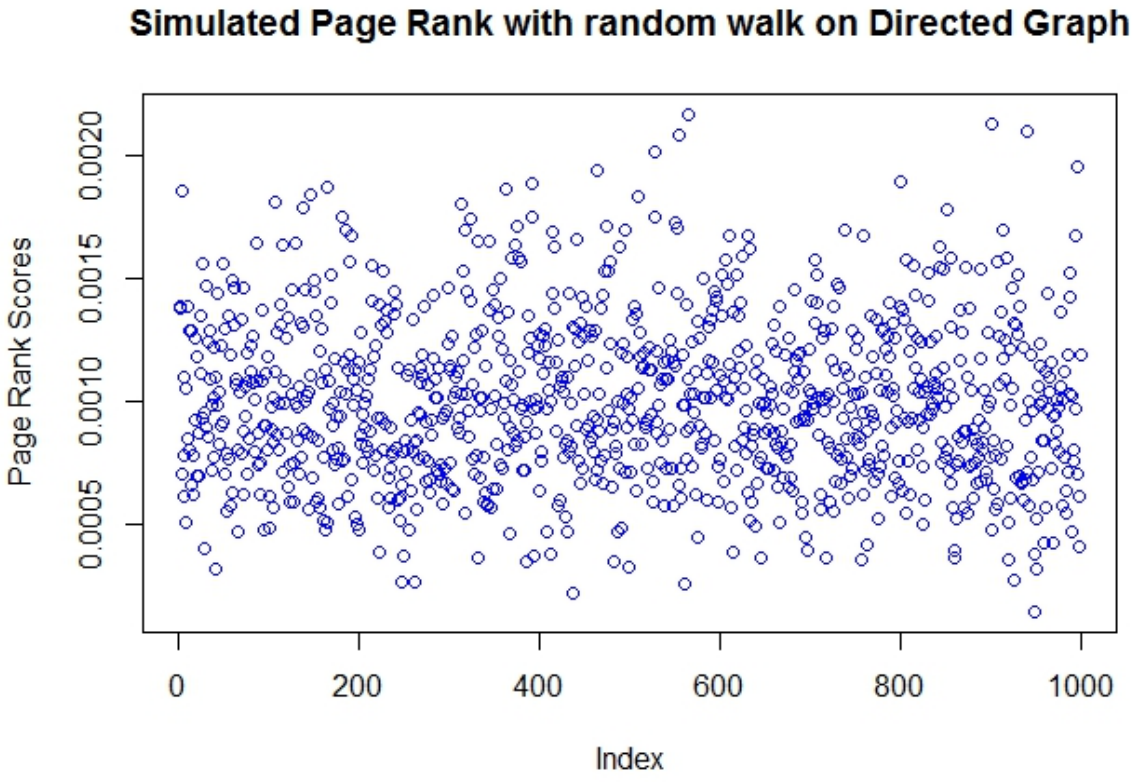
Question 4: Personalized Page Rank

a.) Page rank scores on a directed Random Network

A directed random network with 1000 nodes and probability p of drawing an edge between any pair of edges as 0.01 was created. A random walk was simulated on the network with damping parameter 0.85 and the page rank scores of the nodes was found. The network is shown below.

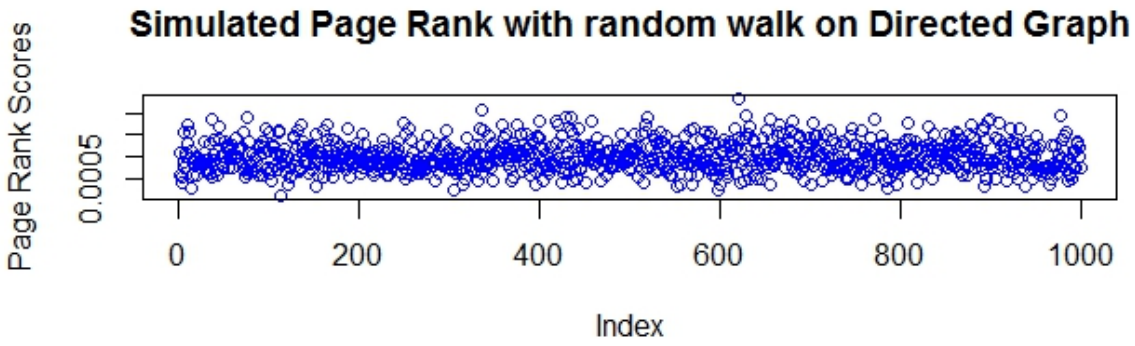


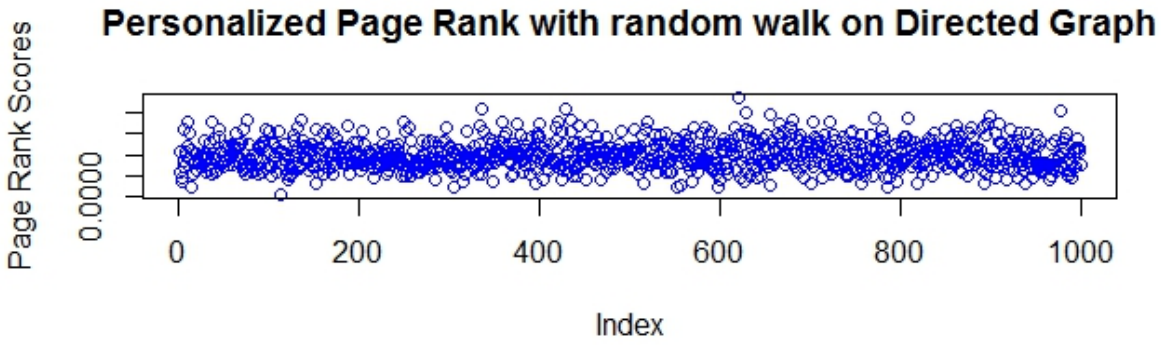
We determined the Page rank scores of the 1000 Nodes by setting the teleportation probability of all nodes equal and set to value 1/N . The plot for page rank scores of all the nodes is given below. It can be seen from the plots below that almost all the nodes have very similar page rank scores.



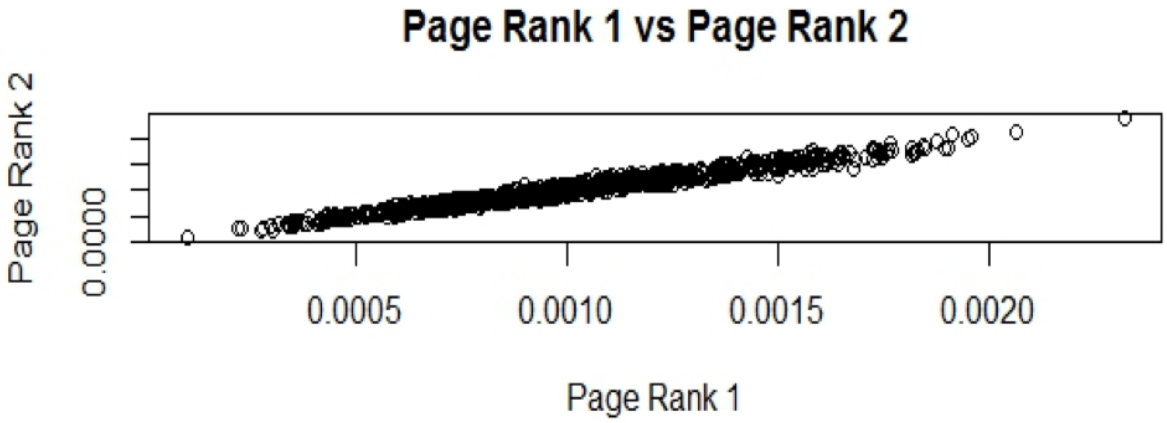
b.) Personalized Page Rank

In this section, we have simulated regular Page Rank and personalized Page Rank on the directed network created in Part a. In the regular page rank the teleportation probability to all nodes are same and equal to 1/N. In the personalized page rank the page rank scores found through a random walk were used as the teleportation probabilities for all the nodes in the network. The personalized page rank scores are plotted over all the nodes and it can seen that it not very different from the regular page rank. The personalized page rank scores have more large values and more small values.





The plot shown below Regular Page Rank vs Personalized Page Rank shows that both of them are very similar, which is evident from the **y = mx** kind of straight line nature of the plot below.



c.) Mathematical Formulation of Personalized Page Rank:

The PageRank can be formulated by the following linear system of equations which can be compactly written in matrix form as given below –

$$\mathbf{M*PR=(1 - d)/N}$$

where $0 < d < 1$ denotes the damping factor, PR is the N-dimensional page rank vector and M is a NxN matrix. N is the number of nodes in the network in our case. For example, the i^{th} component of the vector PR, i.e. PR_i is the page rank of the site i. The matrix M is given by

$$\mathbf{M=1-dT}$$

Where T represents the transition matrix. The components of T are given by the number of outgoing links:

1. $\mathbf{T_{ij} = 1 / C_j}$ **(if node j is linking to node i)**
2. $\mathbf{T_{ij} = 0}$ **(otherwise)**

C_j is the number of links on node j. The solution to the linear system of equations is

$$\mathbf{PR=M\text{-}1\ *(1–d)/N}$$

The solution for M^{-1} is done analytically through iteration scheme generally the Jacobi Iteration. Now the page rank equation can be modified to make it a personalized page rank equation. We can introduce a new personalized vector V so then the equation becomes -

$$\mathbf{PR = M^{-1} * V*(1 – d)}$$

Where the elements of V are the personalized scores of the nodes in the network.