Homework 2

Aozhu Chen 004773895
Ruoxi Zhang 404753334
Ning Xin 704775693

1. Random walk on random networks

(a) Create undirected random networks with 1000 nodes, and the probability p for drawing an edge between any pair of nodes equal to 0.01.

Firstly, we are going to define a function called randomWalker, which is used to perform random walker. There are two parameters in this function which are the number of nodes and the probability p for drawing an edge between any pair of nodes equal to 0.01.

Next, we used the random.graph.game() method which is already exist in the igraph library. There are three parameters which are the number of nodes, the probability p and the directed(False).

Finally, we successfully created the undirected random networks with 1000 nodes and by using cat method we can also find the diameters of random networks.

(b) Let a random walker start from a randomly selected node (no damping). We use t to denote the number of steps that the walker has taken. Measure the average distance ⟨ s(t)⟩ of the walker from his starting point at step t. Also measure the standard deviation σ2(t) = ⟨ (s(t) − ⟨ s(t)⟩)2⟩ of this distance. Plot ⟨ s(t) v.s. t and σ2(t) v.s. t. Here, the average ⟨ ·⟩ is over all possible starting nodes and different runs of the random walk (or different walkers). You can measure the distance of two nodes by finding the shortest path between them.

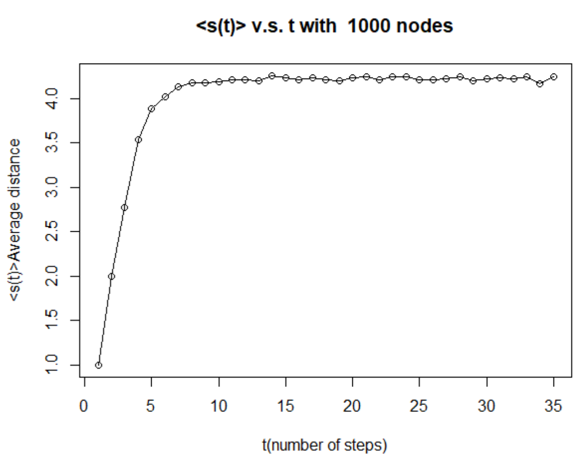


Figure 1.1 ⟨ s(t)⟩ v.s. t

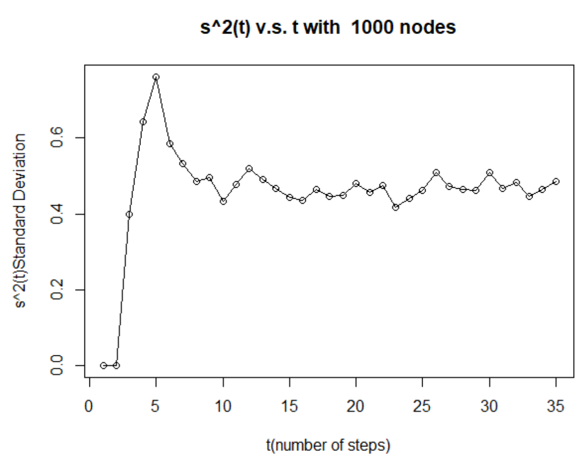


Figure 1.2 σ2(t) v.s. t

In this question, we used the netrw package provided by TA to simulate the random walker network. The number of steps t starts from 1 to 35. The number of nodes is 1000. We measured the average distance $\langle s(t) \rangle$ of the walker from his starting point at step t and measured the standard deviation $\sigma2(t) = \langle (s(t) - \langle s(t) \rangle)2 \rangle$ of this distance. The plots of above shows the relationship of $\langle s(t) \rangle$ v.s. t and $\sigma2(t)$ v.s. t.

(c) We know that a random walker in d dimensional has average (signed) distance $\langle s(t) \rangle = 0$ and $\langle s(t)2 \rangle = \sigma \propto \sqrt{t}$. Compare this with the result on a random network. Do they have similar relations? Qualitatively explain why.

The relationship is not similar, they are different when compare to d dimensional. As we all know, in d dimensional space, the average distance <s(t)> can be negative. But, the average distance of random walker in our graphs are all positive. The diameter of the network can influence the random walk. In part(b), after analyzing the graphs, we get the distance goes to around 4.0 and the standard deviation goes to 0.4.

(d) Repeat (b) for undirected random networks with 100 and 10000 nodes. Compare the results and explain qualitatively. Does the diameter of the network play a role?

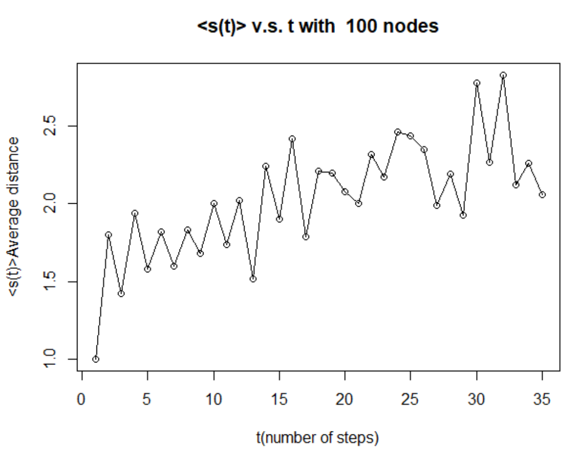


Figure 1.3 $\langle s(t) \rangle$ v.s. t

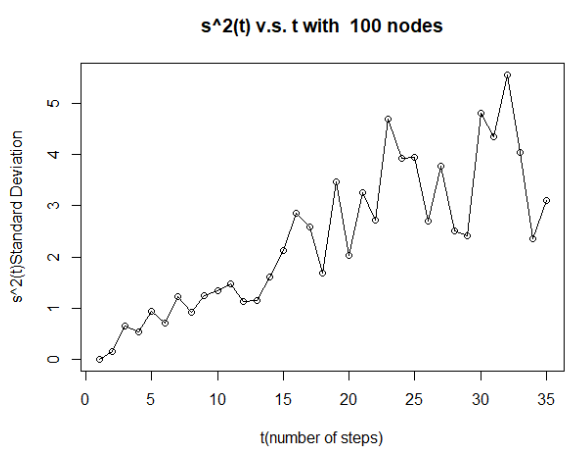


Figure 1.4 $\sigma2(t)$ v.s. t

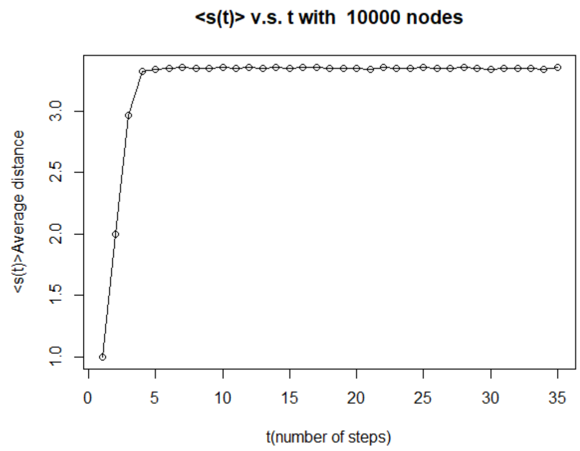


Figure 1.5 ⟨ s(t)⟩ v.s. t

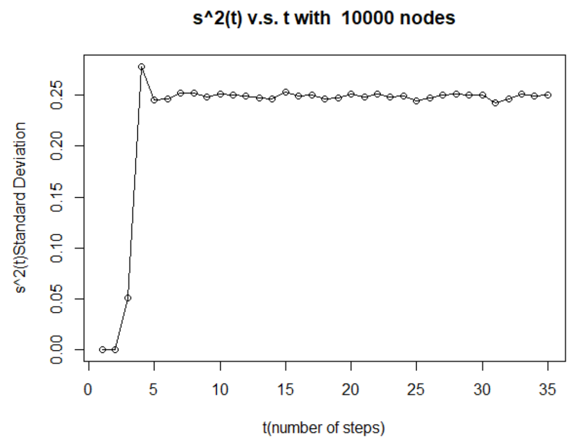


Figure 1.6 σ2(t) v.s. t

By calculating the diameters correspond to the number of nodes, we can get that: when the number of nodes is 100, the diameter is 8; when the number of nodes is 1000, the diameter is 6; when the number of nodes is 10000, the diameter is 3.

By comparing the result of part (b) and part (d), we can figure out that the average distance s was confined by the value of diameter.

(e) Measure the degree distribution of the nodes reached at the end of the random walk on the 1000-node random network. How does it compare with the degree distribution of graph?

After plotting the degree distribution for random undirected graph with 1000 nodes and degree distribution at end of random walk, we can figure out that the two graphs are almost similar.

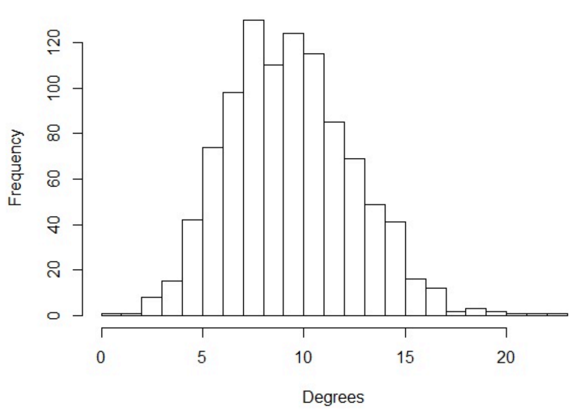


Figure 1.7 Random undirected graph

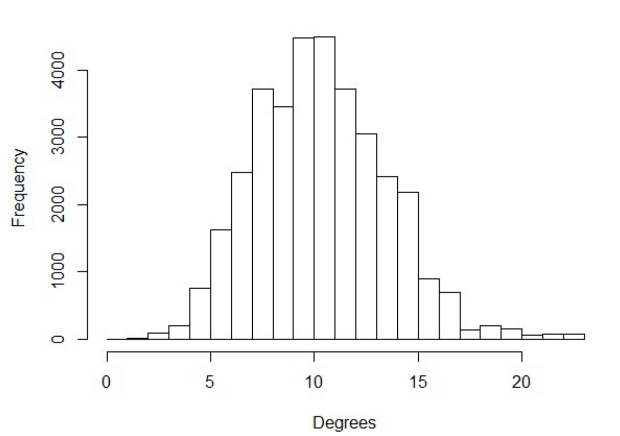


Figure 1.8 At end of random walk

2. Random walk on networks with fat-tailed degree distribution

(a) Using barabasi.game function, we can generate network with 1000 nodes, if we plot the network layout in a figure, it looks like the figure below.

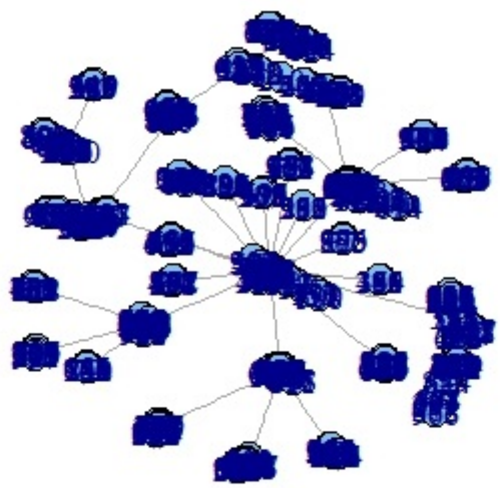


Figure 2.1 Layout of network with 1000 node

(b) $\langle s(t)\rangle$ and $\langle \sigma^2(t)\rangle$ are plot below, which indicates the average distance from the original place at time stamp t. The figures are plotted as below.

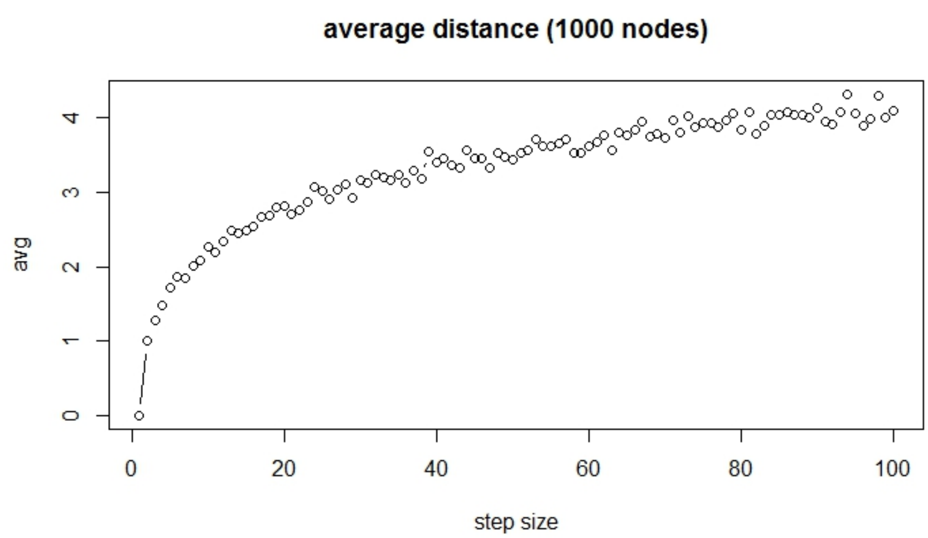


Figure 2.2 Average distance at time t

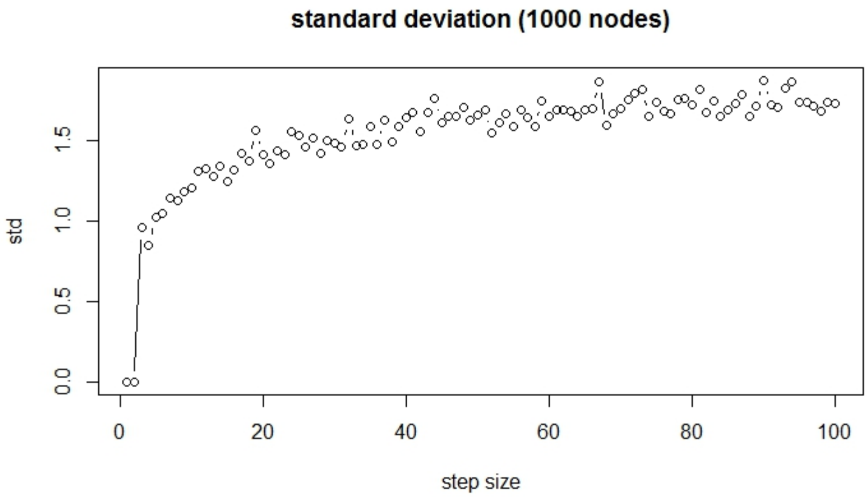


Figure 2.3 Standard Deviation vs. t

From the figure we can see that even if the steps we take could approach 100, the average distance never greater than 4. And the standard deviation is also within 1.5.

(c) According to the conclusion we got from question 1 and class, a d dimensional random walk will have an average distance of 0 and standard deviation proportional to square root of t. That's because the distance is d-dimensional space, the distance can be negative, and hence the combined value is 0. But in our example, the distance value are all positive, that's why the average std-dev will converge to about 3.5 rather than 0.

(d) For network with 100 node, the figures are showed below,

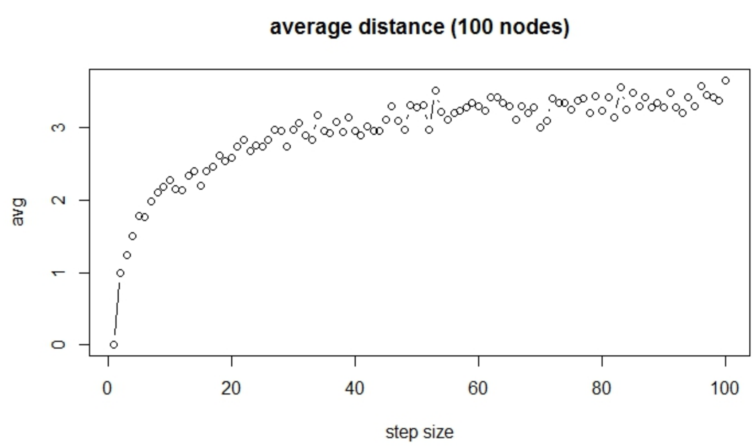


Figure 2.4 Average distance at time t for network with 100 node

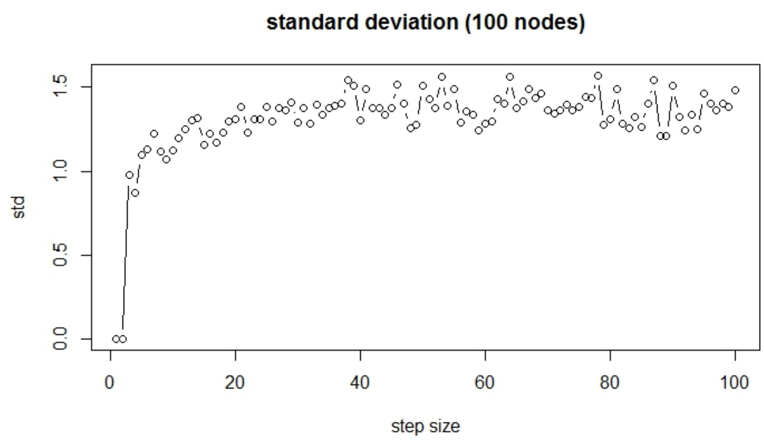


Figure 2.5 Standard deviation for network with 100 nodes

For network with 10000 nodes, the figures are plotted below:

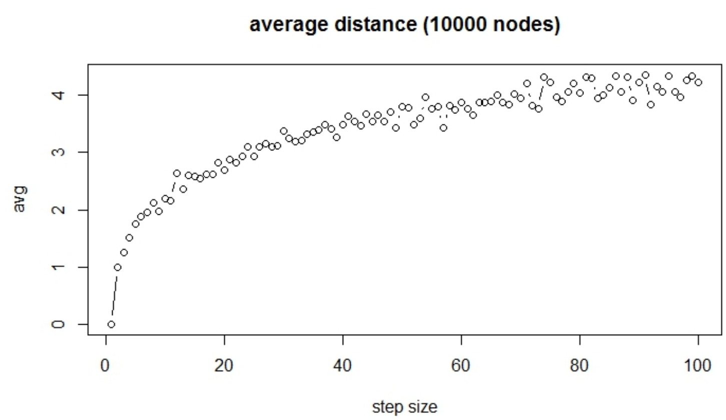


Figure 2.6 Average distance at time t for network with 10000 nodes

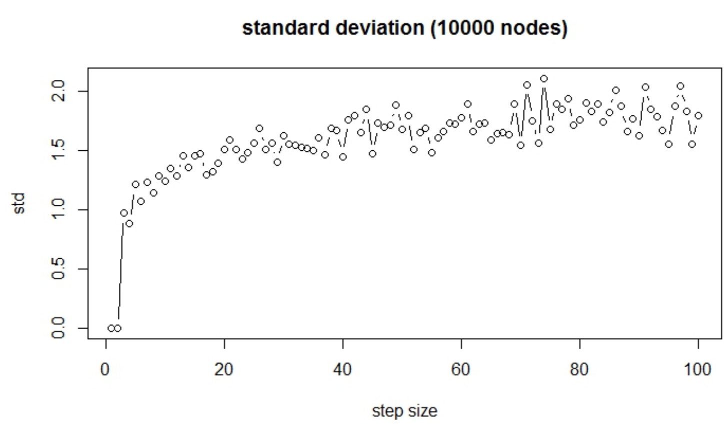


Figure 2.7 Standard deviation for network with 10000 nodes

From the figure we can see that, there's not much difference between the figure of 100 nodes and that of 10000 nodes, basically, they shows similar shape and characteristics. However, the average diameter of the 100-node network is approximately 8 while that of 10000-node network is 27.

(e)

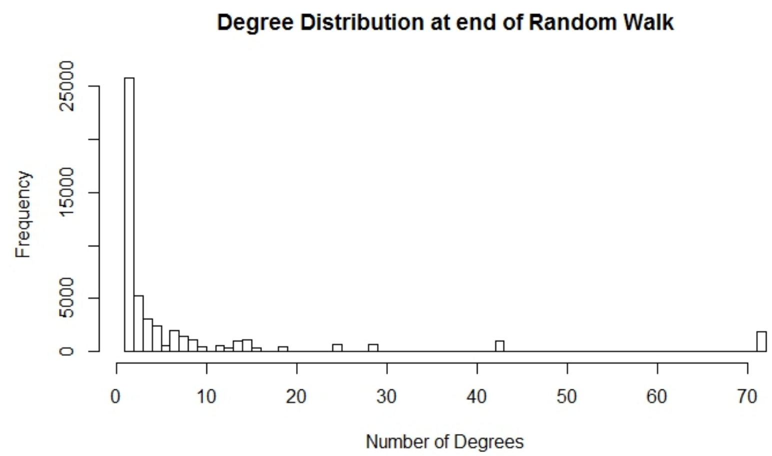


Figure 2.8 Degree distribution at end of random walk

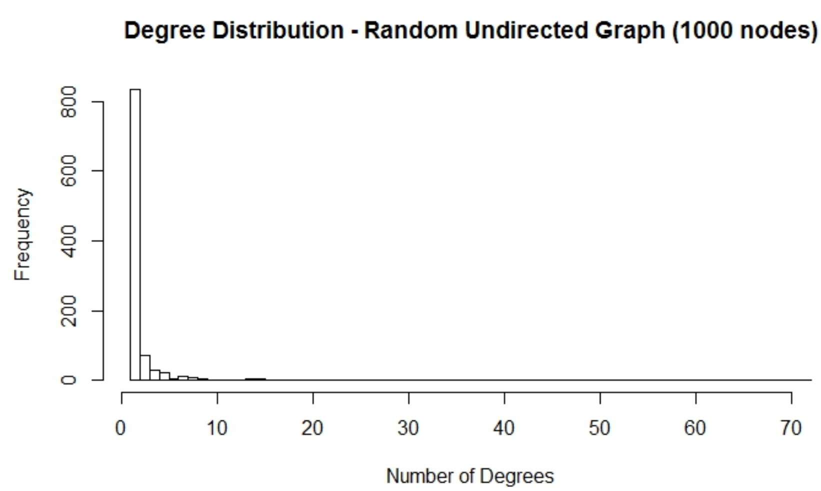


Figure 2.9 Degree distribution of random network

According to the figures we showed above, we see that the degree distribution of original network doesn't have much difference with the distribution after random walk.

3. PageRank

a) We use random walk to simulate PageRank. In this problem, we random walk on a undirected random network with 1000 nodes, and the probability p is equal to 0.01. The netrw() is used for random walk. The damping is set to 1. The number of the walker is 1000. And, there are 5000 time steps to run the simulation. We calculate the correlation between degree and the probability that a node be visited. The result we get as following.

```
Correlation between degree and visit probability:  0.9987972
```

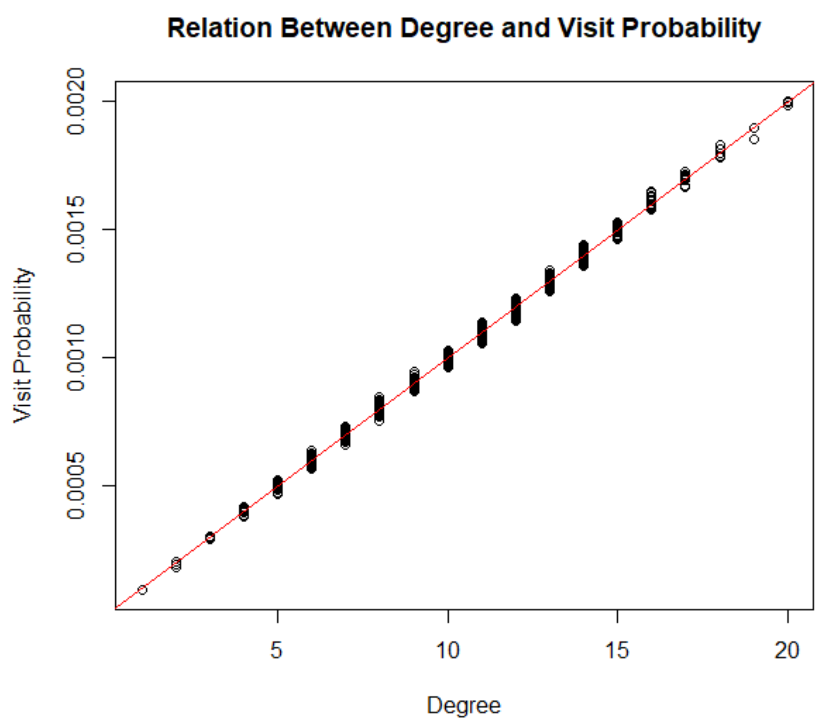
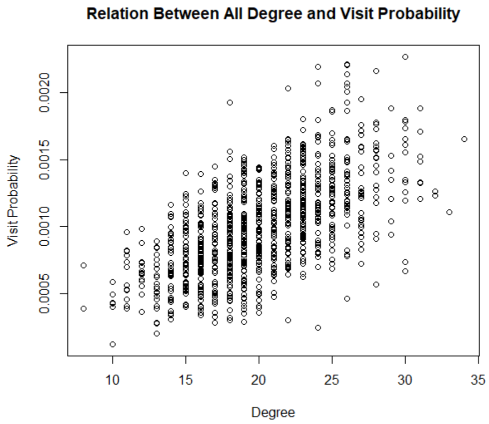


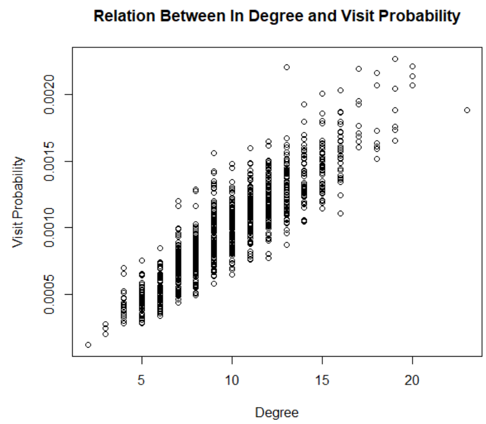
Figure 3.a.1 Relationship between degree and visit probability

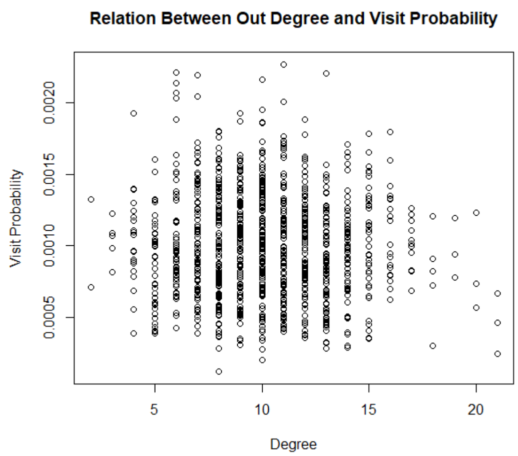
From the correlation value and figure we can see that the visit probability is highly correlated to the degree of the node. From the figure, we can see that a linear model can be fitted to the data.

b) In this problem, we random walk on a directed random network with 1000 nodes, and the probability p is equal to 0.01. The netrw() is used for random walk. The damping is set to 1. The number of the walker is 1000. And, there are 5000 time steps to run the simulation. We calculate the correlation between in degree, out degree, all degree and the probability that a node be visited. The result is summarized as following.

relation	correlation
All-degree	0.6082226
In-degree	0.8878079
Out-degree	-0.0499281



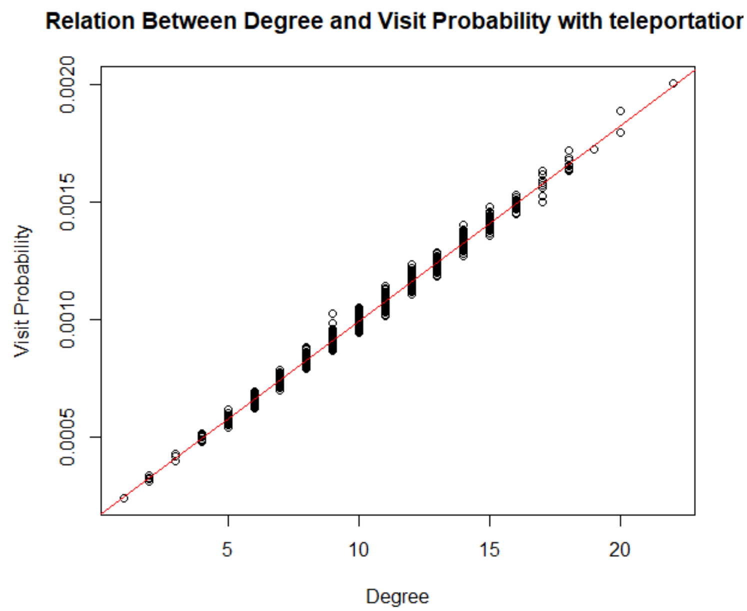




From the correlation value and figure we can see that the in degree of the nodes is correlated to the visit probability. However, the relationship is not as strong as the relationship in the undirected map. Except that, the all degree and the out degree do not have any strong, simple relation with visit probability.

c) In previous problems, the damping factor is 1. However, in this section we choose 0.85 for damping factor of the random work of an undirected network. The other parameter remains same as those in previous sections. We calculate the correlation between degree and the probability that a node be visited. The result we get as following.

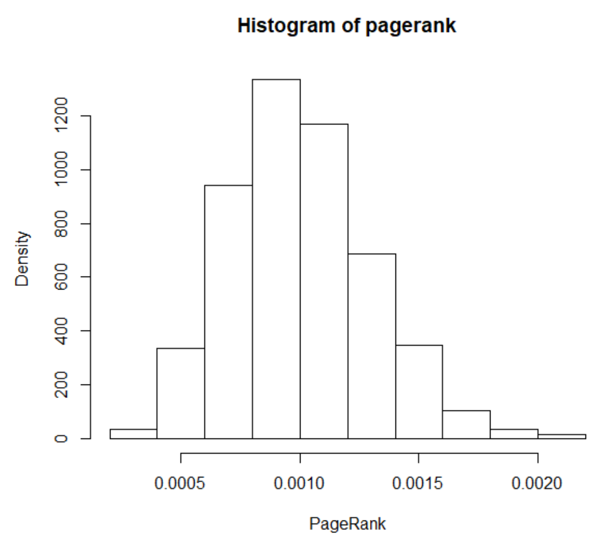
```
Correlation between degree and visit probability:  0.996222
```



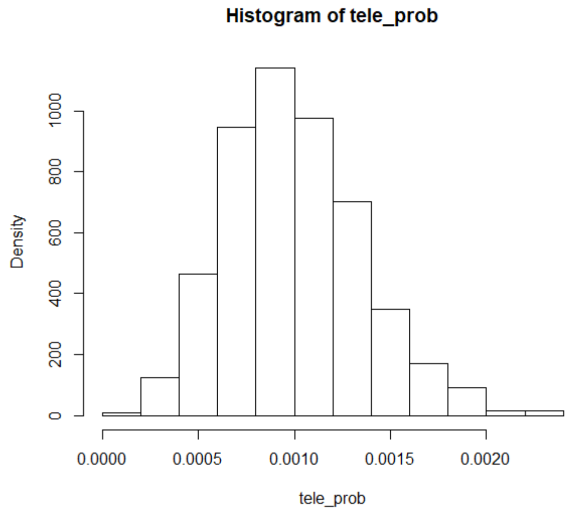
From the correlation value and figure we can see that the visit probability is highly correlated to the degree of the node. It is very close to linear model in the figure which is the red line. However, compared to the figure in the first section, the relation is weaker.

4. Personalized PageRank

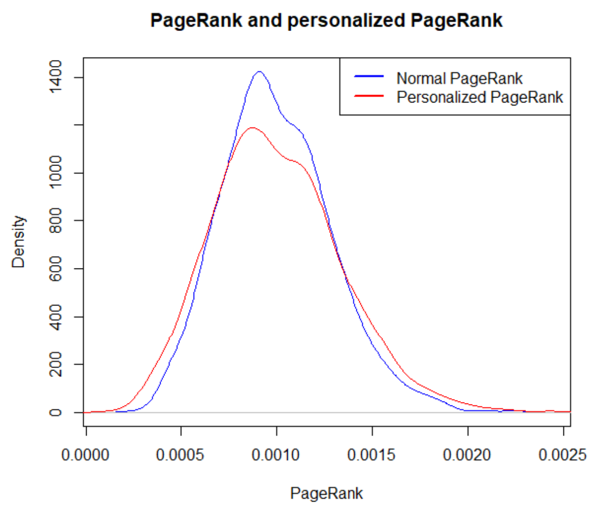
a) In this section, we use the random walk of a directed graph with damping parameter 0.85 to simulate the PageRank of the nodes. The other parameters remain same as previous sections. We calculate the distribution of the PageRank. The result we get as following.



b) To simulate a personalized PageRank, a random direct network with damping factor equals 0.85 is created. The teleportation probability of the random walk is determined by the PageRank of the network which is the visit probability of the network. Again, the damping factor is 0.85. The new PageRank is as follow.



By compare the figure of two sections above together, the result is as following.



We can see that the personalized PageRank has more large value and small values.

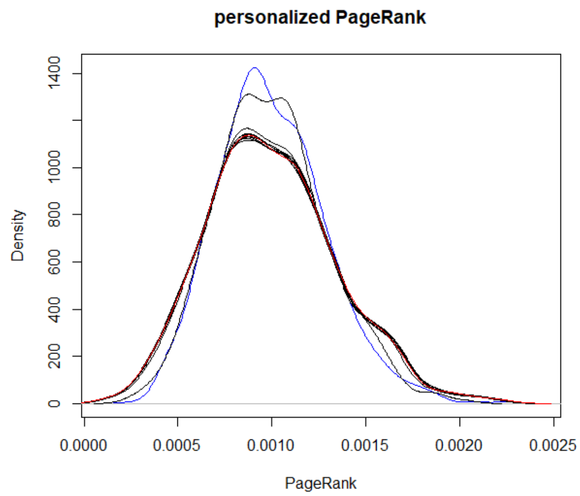
c) The teleportation probability to a node is controlling the interest to the node. By recursively substitute the teleportation probability with the PageRank of the nodes. We can consider the effect of the self-enforcement and adjust the equation to satisfied the real world constrain. The adjusted equation is as the following.

$$\Pr(pi, 0) = \frac{1}{N}$$

$$\Pr(pi, t+1) = \frac{1-d}{N} + d \sum_{Pj \in M(pi)} \frac{\Pr(Pj, t)}{L(Pj)}$$

where p1, p2, …pN are the pages under consideration, M(pi)is the set of pages that link to pi, L(pi) is the number of outbound links on page pj, and N is the total number of pages.

The result we get is as the following.



We can see that after several iterations, the curve converges to the red line.