EE 232E
Graphs and Network Flows
Homework 2
Winter 2016

Liqiang Yu, Rongjing Bai, Yunwen Zhu
904592975, 204587519, 104593417

04-20-2016

Contents

1 Problem1

In this part, we do random walk on random network and try to analyze the relationship between random walk and network structure.

1.1 Part a

We create undirected random networks with 1000 nodes, the probability of the existence of a edge between any two nodes p is 0.01 by using the random.graph.game() function easily.

1.2 Part b

We simulate the random walk process by assigning random number to neighborhood of certain node. And we choose 1000 random walkers, and 100 as the maximum steps in order to get a better simulation result. For each one random walkers, the start node is randomly choosen. In order to get the distance from each starting point at step t, we find the shortest paths between them and measure it. The average distance $\langle s(t) \rangle$, and standard deviation $\sigma^2(t) = \langle (s(t) - \langle s(t) \rangle)^2 \rangle$ v.s. t is plotted in Figure 1 and 2 respectively.

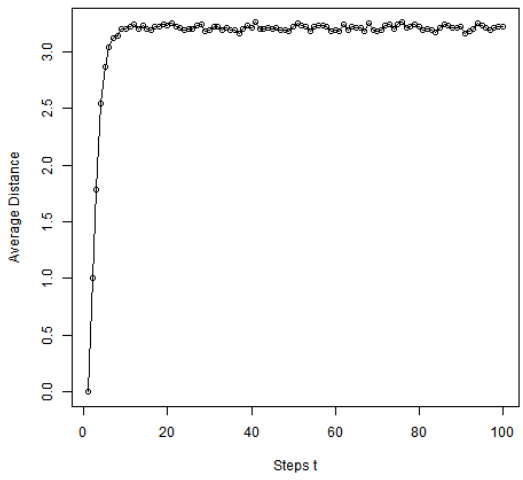

Figure 1: The average distance for random network with 1000 nodes

From the figures above, we can see that the average distance is around 3 and the

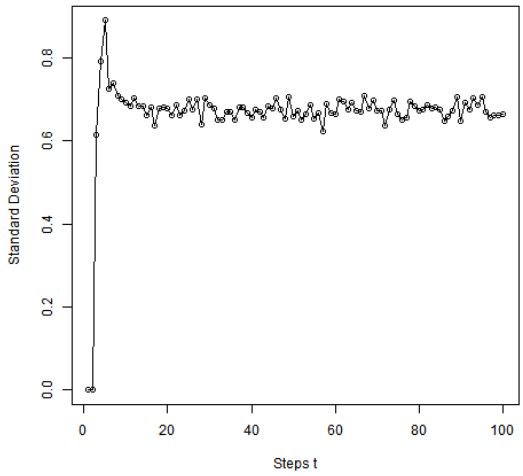

Figure 2: The standard deviation for random network with 1000 nodes

standard deviation is about 0.7. As step t increases, these two values converge to the fixed value.

1.3 Part c

In theory, values of a random walker in d dimensional are $\langle s(t) \rangle = 0$ and $\sqrt{\langle s^2(t) \rangle}$ $= \alpha \propto$t. However, the results in our random network are totally different from the theorical values. The standard deviation does not show any linear relationship with the number of steps t and average distance is not equal to zero. In fact, what we get in the random network, is that both the standard deviation and average distance converge to a constant.

As for the different result of average distance, this is due to the mechanism of our random walk algorithm: the distance of first step is adding one, after that, the minimum value of distance that the walker can ever achieve is zero (due to the fact that we cannot have signed distance). As the distance is always nonnegative, the average value cannot reach zero.

As for the different result of standard deviation, this is due to the structure of random network. We can see that the diameter is rather small in our random network which possible could make the nodes' distribution relatively seperate. The limited number of edges can not guarantee the walker can randomly walk in a widely distributed pattern. The result we get seems that at first $\sqrt{\langle s^2(t) \rangle}$ $= \alpha \propto$t, but later the random walker fell into a cluster and began walking in the local region.

1.4 Part d

When the number of nodes n is set to be 100, $\sigma^2(t) = \langle (s(t) - \langle s(t) \rangle)^2 \rangle$ v.s. t is plotted in Figure 3 and 4 respectively. And we choose 100 random walkers, and 100 as the maximum steps in order to get a better simulation result.

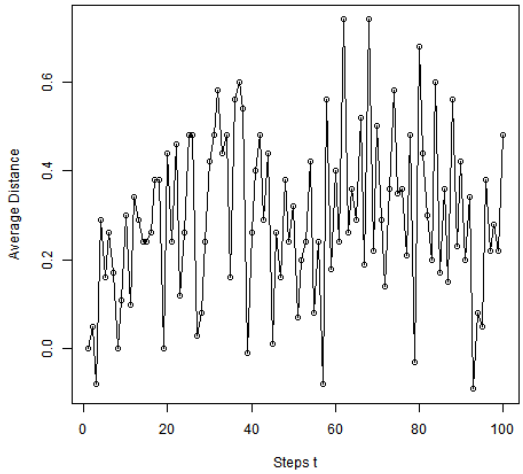

Figure 3: The average distance for random network with 100 nodes

When the number of nodes n is set to be 10000, $\sigma^2(t) = \langle (s(t) - \langle s(t) \rangle)^2 \rangle$ v.s. t is plotted in Figure 5 and 6 respectively. And we choose 1000 random walkers, and 100 as the maximum steps in order to get a better simulation result.

The diameter of these three network is shown in Table 1. Comparing with

Table 1: diameter of random network

	n=100	n = 1000	n = 10000
diameter	9	6	3

Figure 1 and 2, we can see the trend of average distance and standard deviation in random network with 1000 and 10000 nodes are similar that both converge to constants. However, as the number of nodes increases(size of network increases), these two convergence values become smaller. This is because when the network becomes larger,the walker are freer to randomly walk in a relatively widely distributed pattern. Thus the average distance and the standard deviation will be more closer to the theorical values.

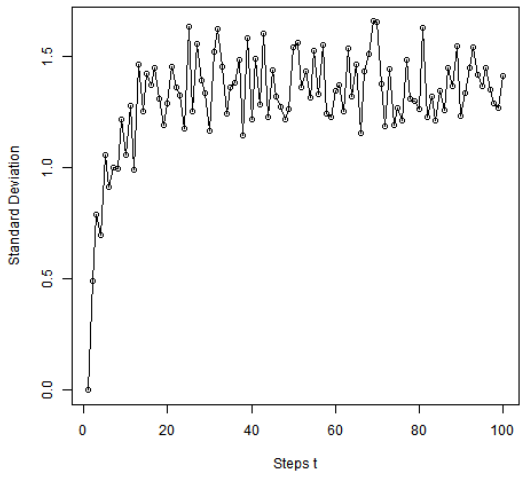

Figure 4: The standard deviation for random network with 100 nodes

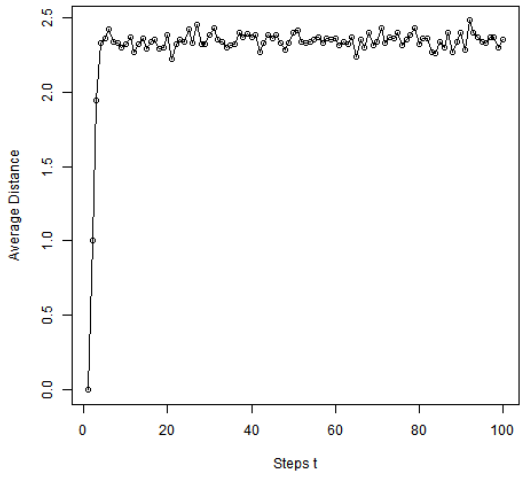

Figure 5: The average distance for random network with 10000 nodes

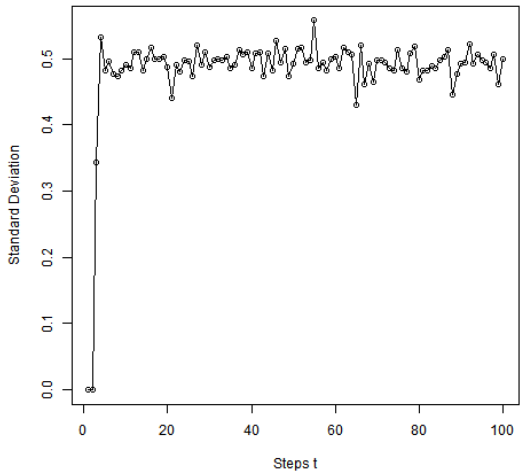

Figure 6: The standard deviation for random network with 10000 nodes

Note that the diameter value for these networks decreases with smaller average distance and standard deviation as the number of nodes increase. Furthermore, the average distance can never exceed the value of its diameter. Thus, to some extent,the diameter plays a role in the properties of network.

However, the random network with 100 nodes does not follow the same trend like the two other networks. This is possibly due to the fact that network with 100 nodes is disconnected and diameter is not applicable for disconnected graph.

1.5 Part e

The degree distribution at the end of random walk is shown in Figure 7 and the degree distribution of the graph is shown in Figure 8.

We can see from the figures above that the degree distribution at the end of the random walk and the degree distribution of graph are generally resembles each other. This similarity between two distributions means that nodes the walker went through are randomized, thus indicates the reliability of our random walk algorithm.

2 Problem2

In this part, we do random walk on networks with fat-tailed degree distribution and try to analyze the relationship between random walk and network structure.

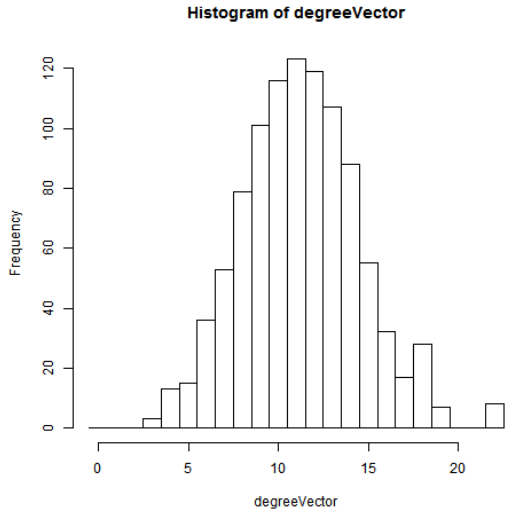

Figure 7: The degree distribution at the end of random walk

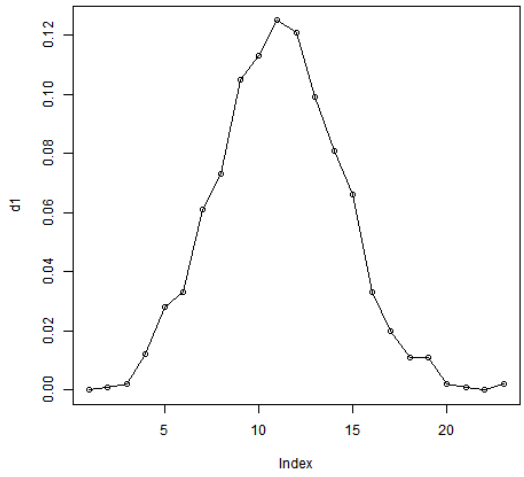

Figure 8: the degree distribution of the graph

2.1 Part a

We create a fat-tailed network with 1000 nodes and degree distribution proportional to x^3 by using the barabasi.game()function easily.

2.2 Part b

Similarly, we simulate the random walk process by assigning random number to neighborhood of certain node. And we choose 1000 random walkers, and 100 as the maximum steps in order to get a better simulation result. For each one random walkers, the start node is randomly choosen. In order to get the distance from each starting point at step t, we find the shortest paths between them and measure it. The average distance $\langle s(t) \rangle$, and standard deviation $\sigma^2(t) = \langle (s(t) - \langle s(t) \rangle)^2 \rangle$ v.s. t is plotted in Figure 9 and 10 respectively.

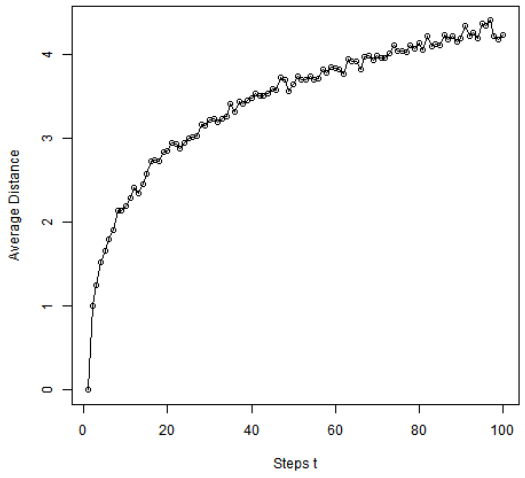

Figure 9: The average distance for fat-tailed network with 1000 nodes

From the figures above, we can see that the average distance and the standard deviation both grow proportional to the step t.

2.3 Part c

The results of average distance and standard deviation is different from the random network in problem 1. In the fat-tailed network, the degree distribution of the nodes is more tightly bounded and densely located in a small range, so the

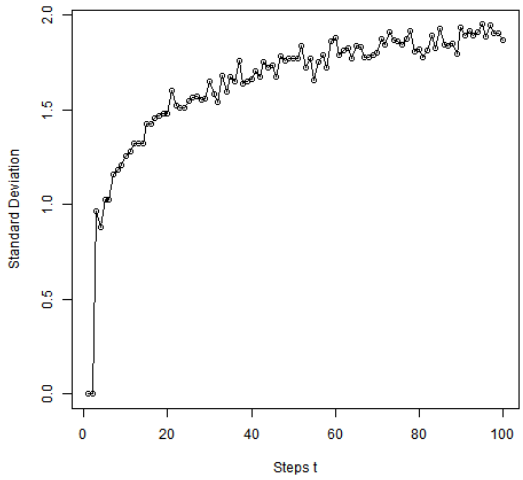

Figure 10: The standard deviation for fat-tailed network with 1000 nodes

degree of each nodes is more likely resemble. Thus, with the incrementation of the step t, the walker will be more likely to walk in a remote area. In this case, the average distance and the the standard deviation will be proportional to the t.

However, in the previous random network, the node is more sparsely distributed and the walk may be more likely to be trapped in a certain area, which could be the reason the the convergence of these two indicators.

The reason why the average distance still differs the theorical value zero is due to the mechanism of our random walk algorithm. The minimum value of distance that the walker can ever achieve is zero (due to the fact that we cannot have signed distance). As the distance is always nonnegative, the average value cannot reach zero.

2.4 Part d

When the number of nodes n is set to be 100, $\sigma^2(t) = \langle(s(t) - \langle s(t)\rangle)^2\rangle$ v.s. t is plotted in Figure 11 and 12 respectively. And we choose 100 random walkers, and 100 as the maximum steps in order to get a better simulation result.

When the number of nodes n is set to be 10000, $\sigma^2(t) = \langle(s(t) - \langle s(t)\rangle)^2\rangle$ v.s. t is plotted in Figure 13 and 14 respectively. And we choose 10000 random

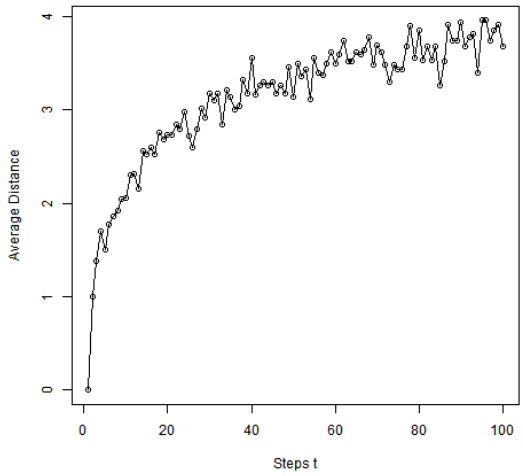

Figure 11: The average distance for fat-tailed network with 100 nodes

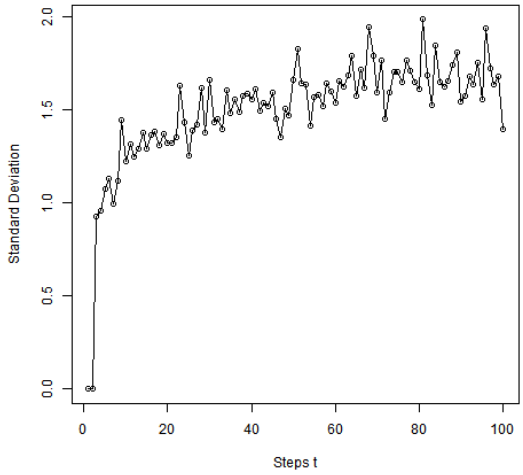

Figure 12: The standard deviation of average distance for fat-tailed with 100 nodes

walkers, and 100 as the maximum steps in order to get a better simulation result.

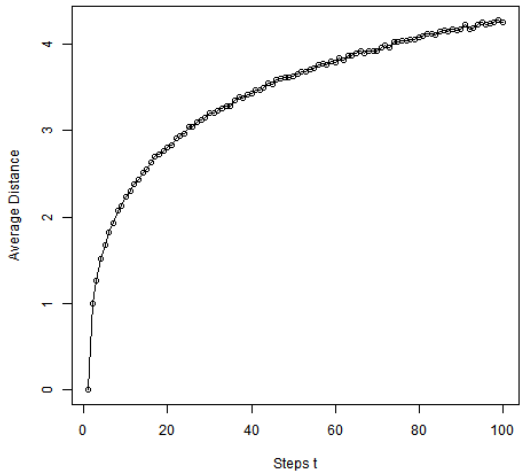

Figure 13: The average distance for fat-tailed network with 10000 nodes

The diameter of these three networks is shown in Table 2. Comparing with

Table 2: diameter of fat-tailed network

	n=100	n = 1000	n = 10000
diameter	11	19	27

Figure 9 and 10, we can see the trend of average distance and standard deviation in all three networks with 100, 1000 and 10000 nodes are similar that both proportional to t. Unlike 1(d), here 100 nodes fat-tailed network is connected, thus its result is similar to the other two and we can analyze these three results all together.

Since every network is connected, the random walker can travel to any node in the network, then diameter will not play a role. Thus, the average distances among the three cases are nearly the same. However, the standard deviation in the case of 100 nodes is a bit larger than other two cases. This is possibly due to the limited size in that case. If the network size is too small, walker are relatively easier to get trapped in certain subregion, thus the walker behavior becomes more uncertain.

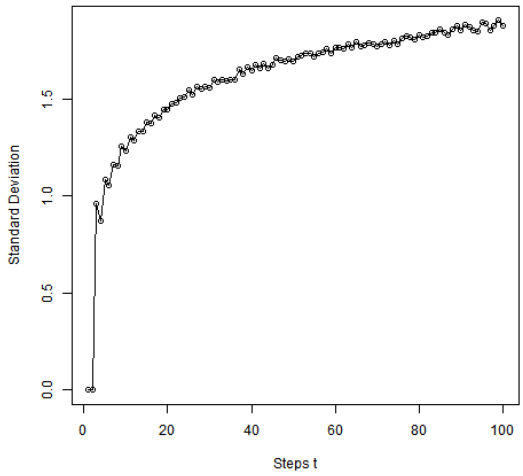

Figure 14: The standard deviation of average distance for fat-tailed network with 10000 nodes

2.5 Part e

The degree distribution at the end of random walk is shown in Figure 15 and the degree distribution of the graph is shown in Figure 16.

We can see from the figures above that the degree distribution at the end of the random walk and the degree distribution of graph are generally resembles each other.This similarity between two distributions means that nodes the walker went through are randomized, thus indicates the reliability of our random walk algorithm.

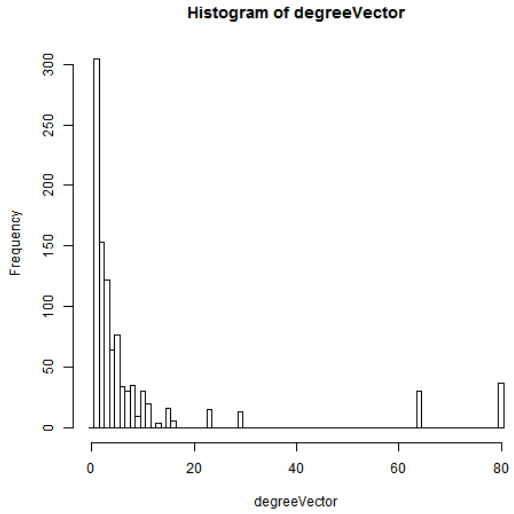

Figure 15: The degree distribution at the end of fat-tailed walk

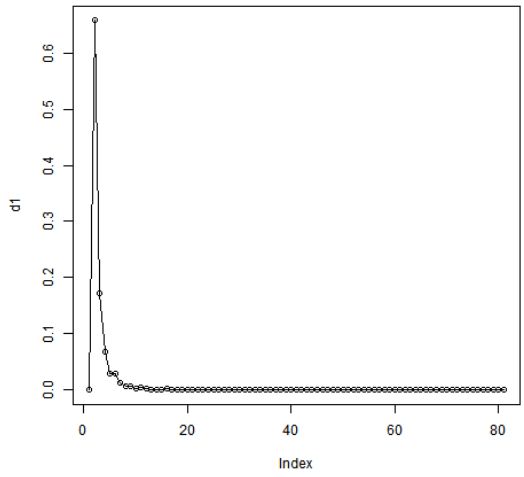

Figure 16: the degree distribution of the graph

3 Problem 3

We create a graph of 1000 nodes and

3.1 Part a

We plot the average visit probability, we can see that average visiting probability is closely RELATED to the degree. Low degree causes low average probability and high degree causes high average probabillity. From the second graph, we can see that the average visiting probability has linear relation to the degree of the node. The covariance AVP and degreee is about 0.9471915(about 0.95 after several tests)

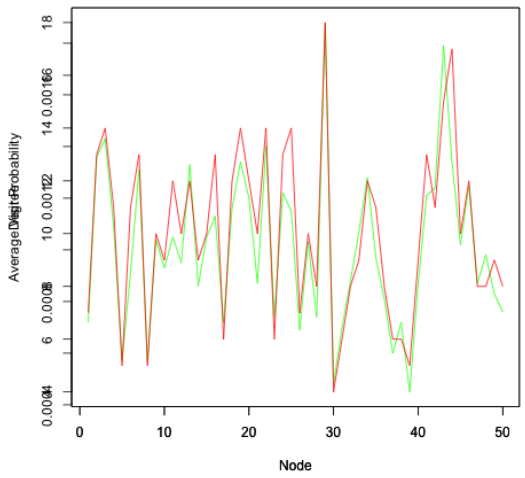

Figure 17: Statistic of degree and average visit probability for undirected graph with damp factor=1

3.2 Part b

For directed graph, we plot in degree and out degree respectively.
For in degree, we plot the relation in Figure 19 and Figure 20. It is similar with problem3a.
There is a close linear relation, and the coriance is 0.8631107(around 0.86 after several tests), which is little than the covariance of undirected graph.
However, for degree it is a different situation. We plot the relation in Figure 21The covariance is very low(0.01592023, around 0 after several tests), and there is no obvious relation between probability and out degree.

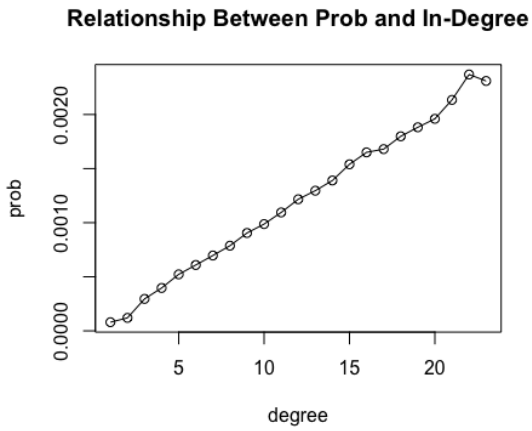

Figure 18: Relation between degree and average visit probability for undirected graph with damp factor=1

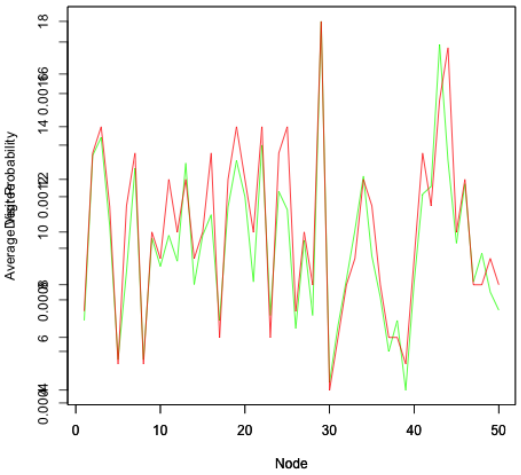

Figure 19: Statistic of indegree and average visit probability for directed graph with damp factor=1

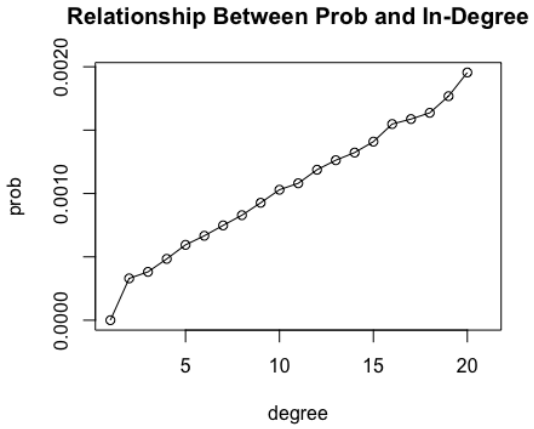

Figure 20: Relation of indegree and average visit probability for directed graph with damp factor=1

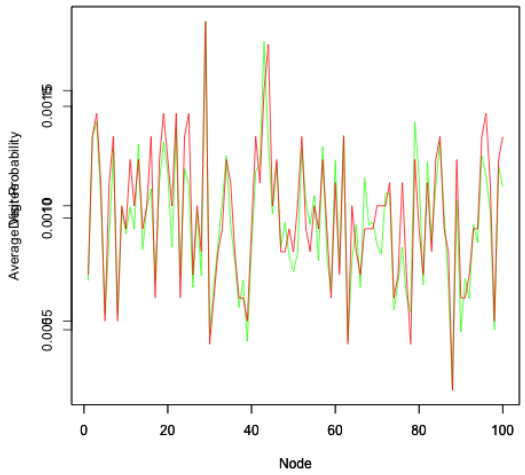

Figure 21: Statistics of outdegree and average visit probability for undirected graph with damp factor=1

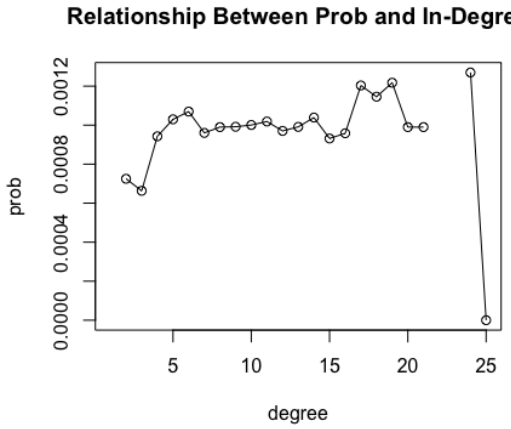

Figure 22: Relation between outdegree and average visit probability for undirected graph with damp factor=1

3.3 Part c

We plot degree and average visit probability in Figure 23 and Figure24. The relation is similar with Problem3a. The average visit probability is still closely linearly related, but the covariance is 0.919052(about 0.92 after several tests), which is slightly little than the covariance for damping factor =1, and greater than the value undirected graph.

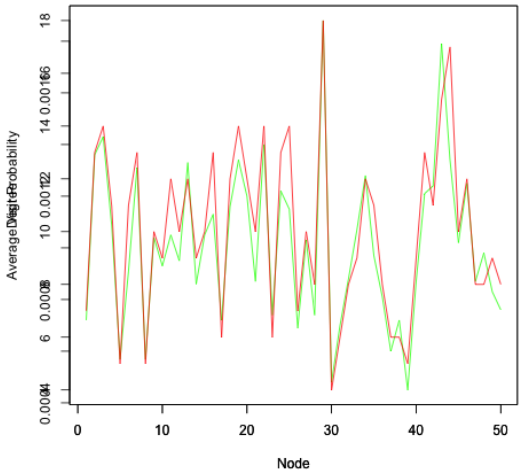

Figure 23: statistics of degree and average visit probability for undirected graph with damp factor=0.85

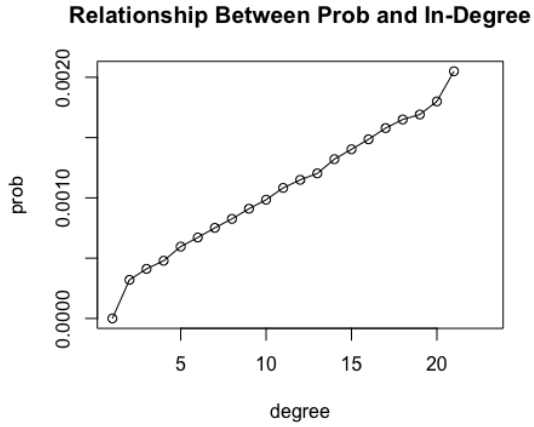

Figure 24: Relation between degree and average visit probability for undirected graph with damp factor=0.85

4 Problem 4

4.1 Part a

The ramdom directed network was created with the damping factor 0.85. It had 1000 nodes and we chose to run the random walk experiment for 1000 times. Figure 25 shows the result of the average page rank.

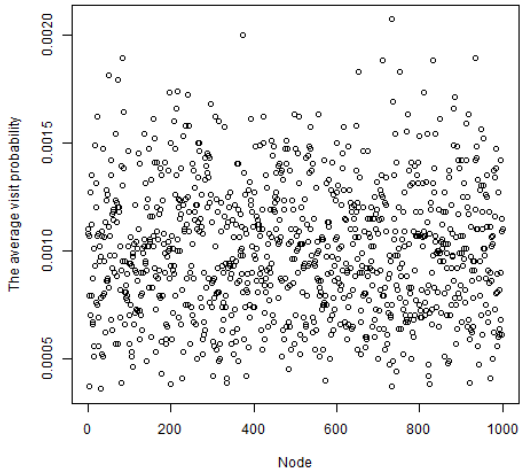

Figure 25: The page rank VS node index

4.2 Part b

The random directed network was created with the damping factor 0.85. It had 1000 nodes and we chose to run the random walk experiment for 1000 times. In this network, the walker has his own notion of interest, which means the teleportation probability to all nodes is not equal. Here, the teleportation probability is proportional to its pagerank. Figure 26 shows the result of the average page rank for this personalized network.

Figure 27 shows the result when we put two kinds of pagerank together into one graph. From the figure we can see that due to the comparably large damping factor(0.85), there is no prominent difference between them. However, there are still some notable differences. For instance, some nodes have higher visit probability in personalized page rank than in normal page rank because teleportation probability is proportational to the original page rank which enlarge the effect of high page rank.

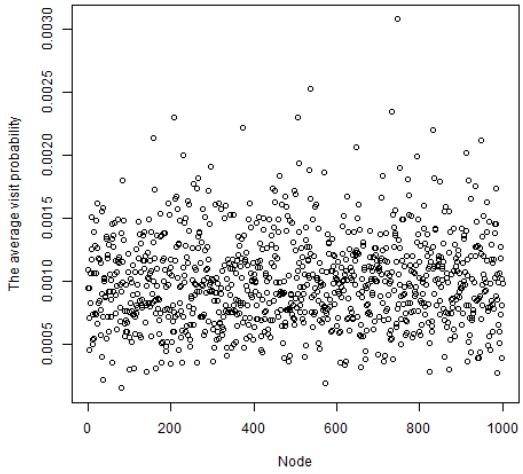

Figure 26: The personalized page rank VS node index

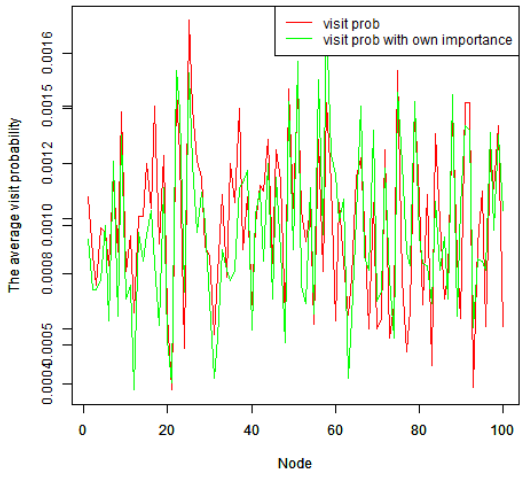

Figure 27: Normal page rank VS Personalized page rank

4.3 Part c

Taking into account the self-enforcement of the page rank, we need to add another factor to represent the users' random behavior. Thus, the page rank equation can be modified as

$$PR(p_i) = d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{C(p_j)} + \frac{1-d}{N}$$

where $M(p_i)$ is the set of nodes that has links to p_i, $C(p_j)$ is the number of output links p_j has, N is the total number of nodes the network has, d is the damping factor. This is the random-surfer model for page rank where the second factor represents the user may jump to a random page with the probability of $1 - d$. It can avoid the problem of dangling page (values dissipates through pages without any links) and crawler trap(pages that just point to each other that all values may accumulate on those pages).