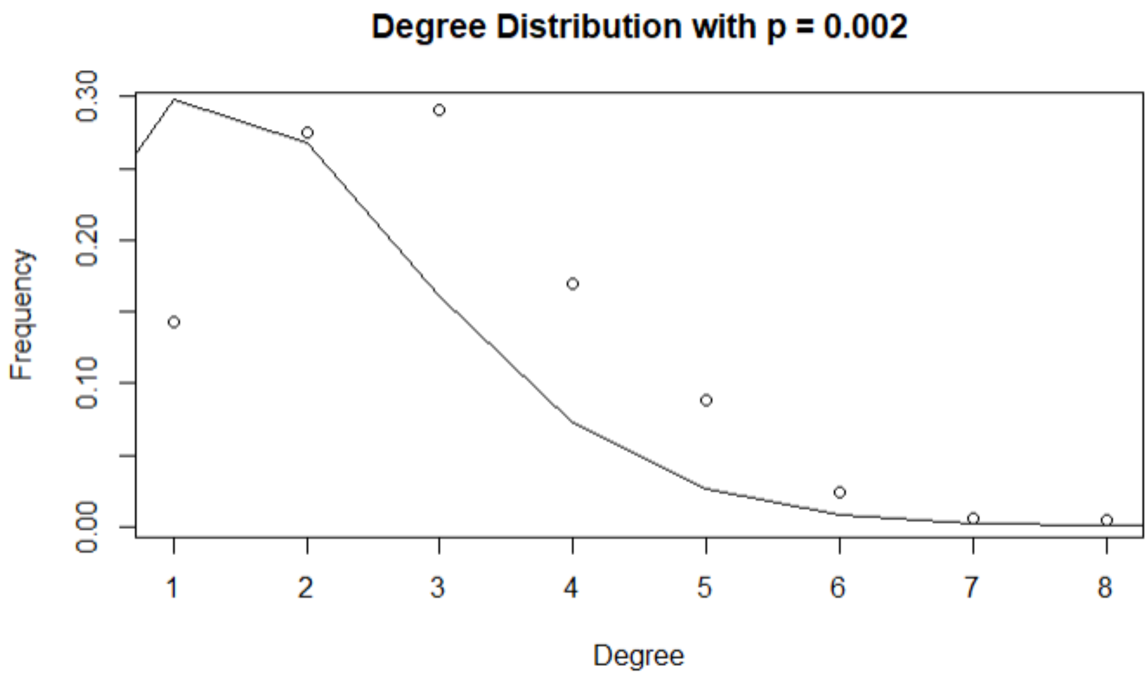
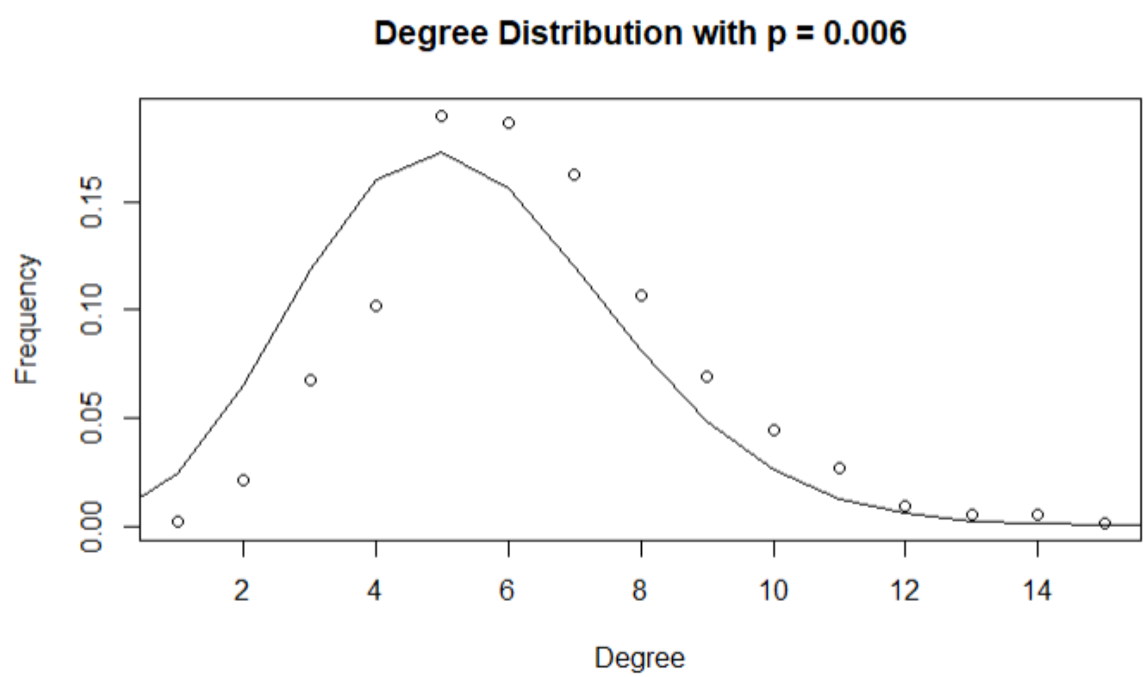
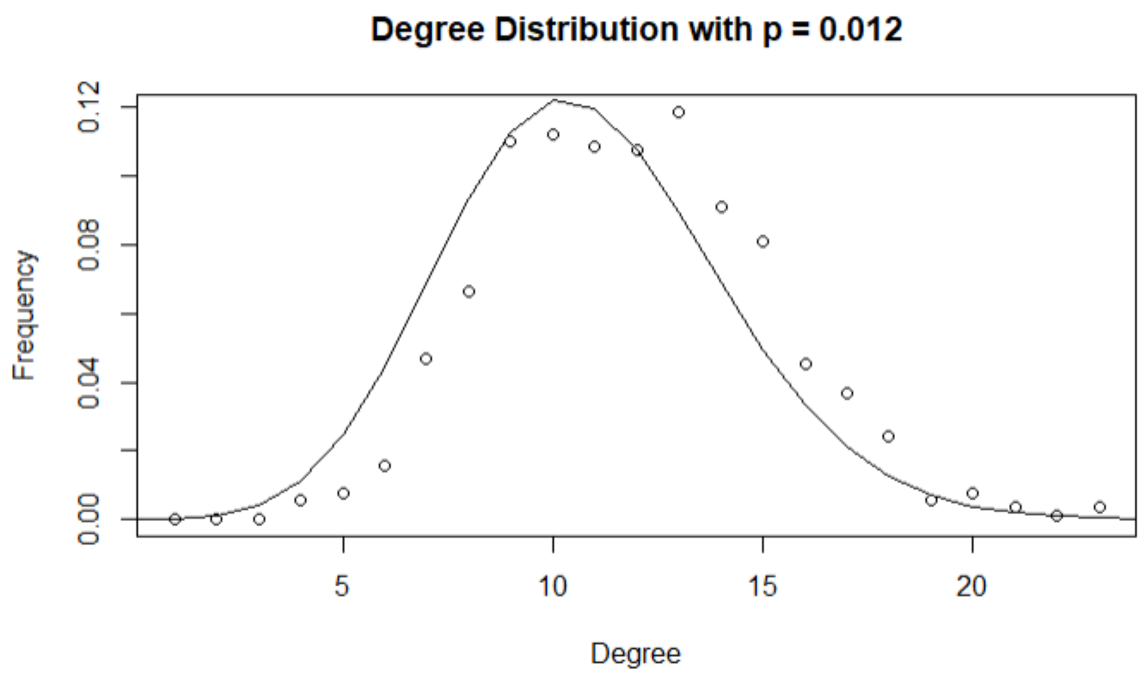
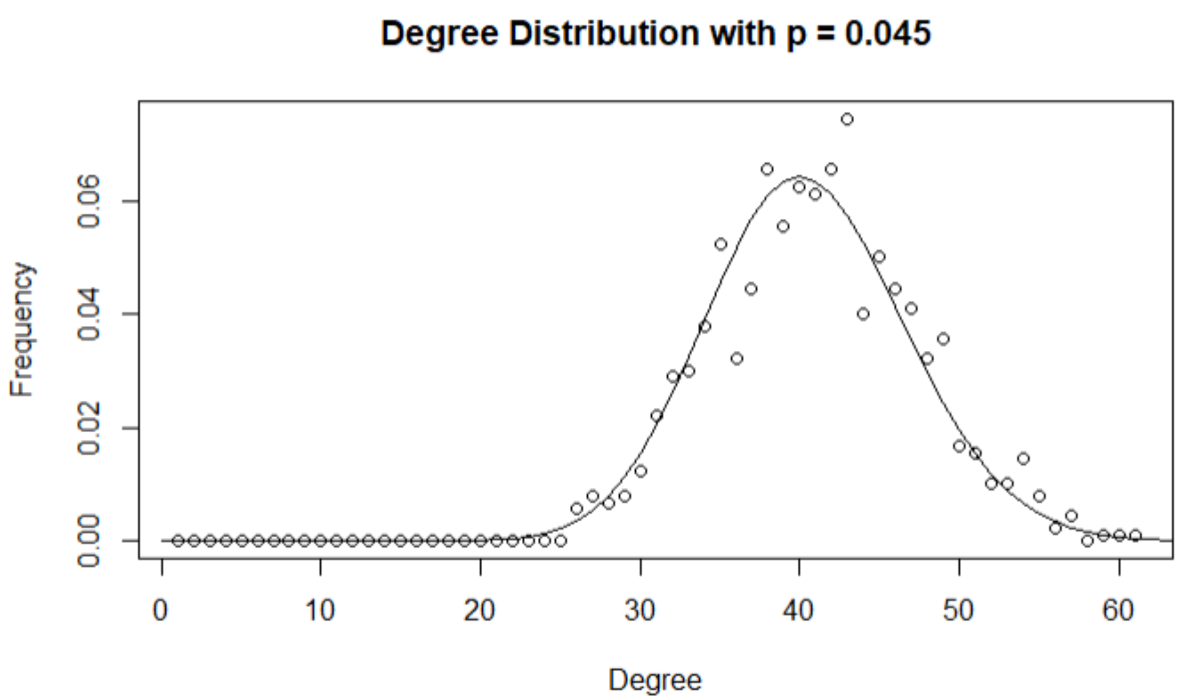
Project 1_ Report (Random Graphs and Random Walks)
Haoting Ni (905545789), Yikai Wang (905522085), Yuanxuan Fang (005949389)

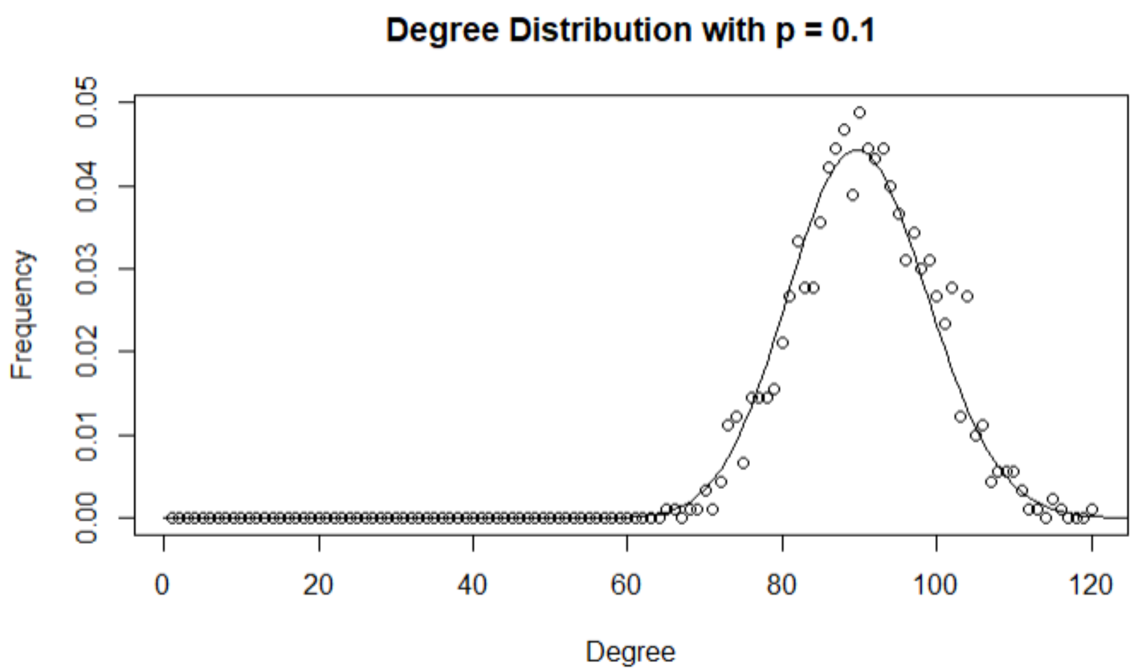
1.1.a






Degree Distribution with p = 0.012
Frequency
0.00
0.04
0.08
0.12
5
10
15
20
Degree


Degree Distribution with p = 0.045
Frequency
0.00
0.02
0.04
0.06
0
10
20
30
40
50
60
Degree




```
[1] "With p = 0.002, Mean=1.902222, Variance=1.776858"
[1] "With p = 0.006, Mean=5.346667, Variance=5.112169"
[1] "With p = 0.012, Mean=10.880000, Variance=10.368231"
[1] "With p = 0.045, Mean=39.997778, Variance=39.481641"
[1] "With p = 0.1, Mean=89.722222, Variance=77.477815"
```

The distribution for above plots with n=900 and p = 0.002, 0.006, 0.012, 0.045, 0.1 is binomial distribution. Also, as p increases and n is sufficiently large, we can consider the part of binomial distribution as the normal distribution.

Theoretical Values:
Mean = np; Variance = np(1-p)

```
[1] "Theoretical Values: With p = 0.002, Mean=1.800000, Variance=1.796400"
[1] "Theoretical Values: With p = 0.006, Mean=5.400000, Variance=5.367600"
[1] "Theoretical Values: With p = 0.012, Mean=10.800000, Variance=10.670400"
[1] "Theoretical Values: With p = 0.045, Mean=40.500000, Variance=38.677500"
[1] "Theoretical Values: With p = 0.1, Mean=90.000000, Variance=81.000000"
```

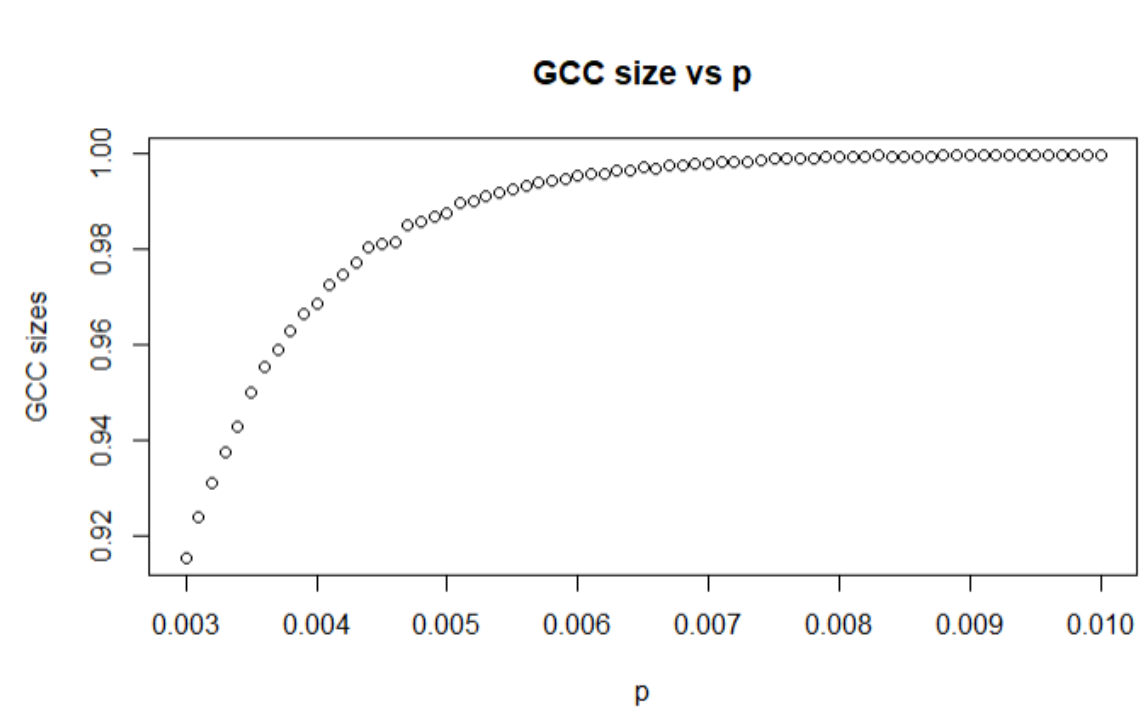
Compared to the theoretical values, the mean and variance of the degree distribution are very close to the theoretical values.

1.1.b

```
[1] "Probability of connectedness for p = 0.002: 0.000000"
[1] "Diameter of GCC for a random non-connected network for p = 0.002: 26.000000"
[1] "**************************************************************"
[1] "Probability of connectedness for p = 0.006: 0.000000"
[1] "Diameter of GCC for a random non-connected network for p = 0.006: 8.000000"
[1] "**************************************************************"
[1] "Probability of connectedness for p = 0.012: 0.003333"
[1] "Diameter of GCC for a random non-connected network for p = 0.012: 5.000000"
[1] "**************************************************************"
[1] "The network is fully connected, no Diameter of GCC with p = 0.045"
[1] "The network is fully connected, no Diameter of GCC with p = 0.1"
```

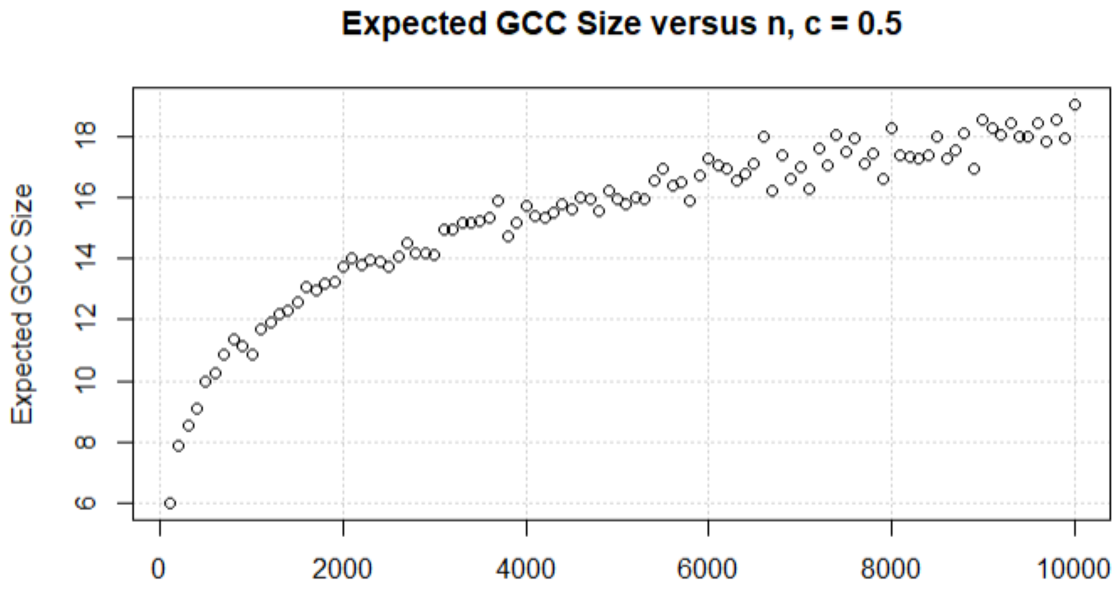
When p = 0.002, 0.006, and 0.012, the probability of connectedness is really small, which means the realizations of the ER network are not connected. Also, the diameter of GCC will decrease as the value of p increases. When p = 0.045 and 0.1, the realizations are always connected.

1.1.c

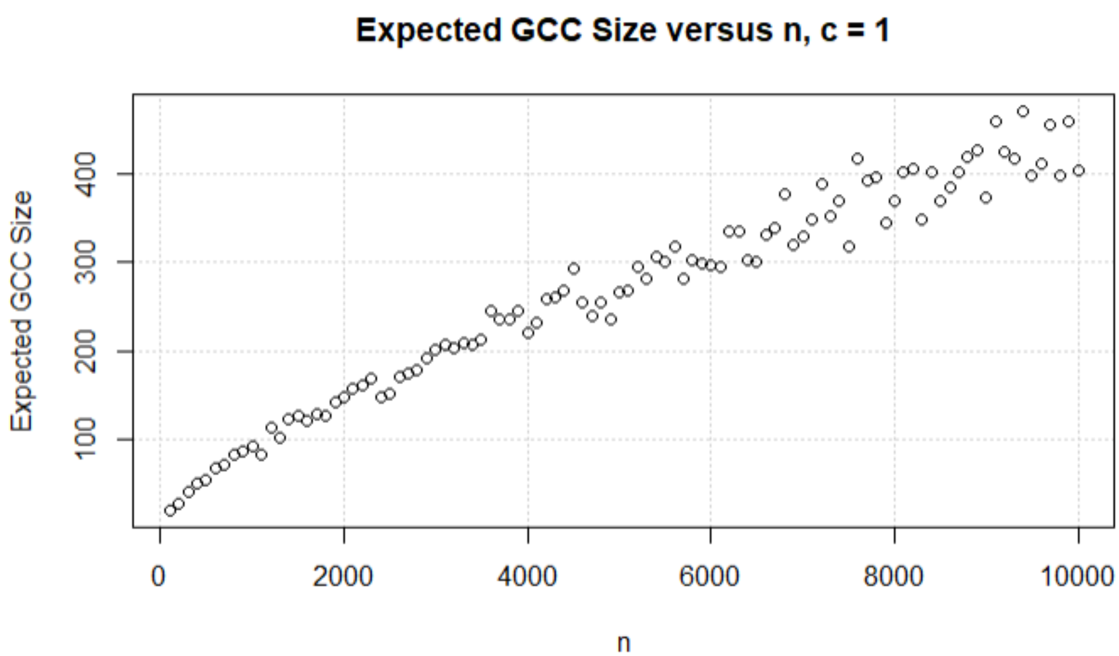


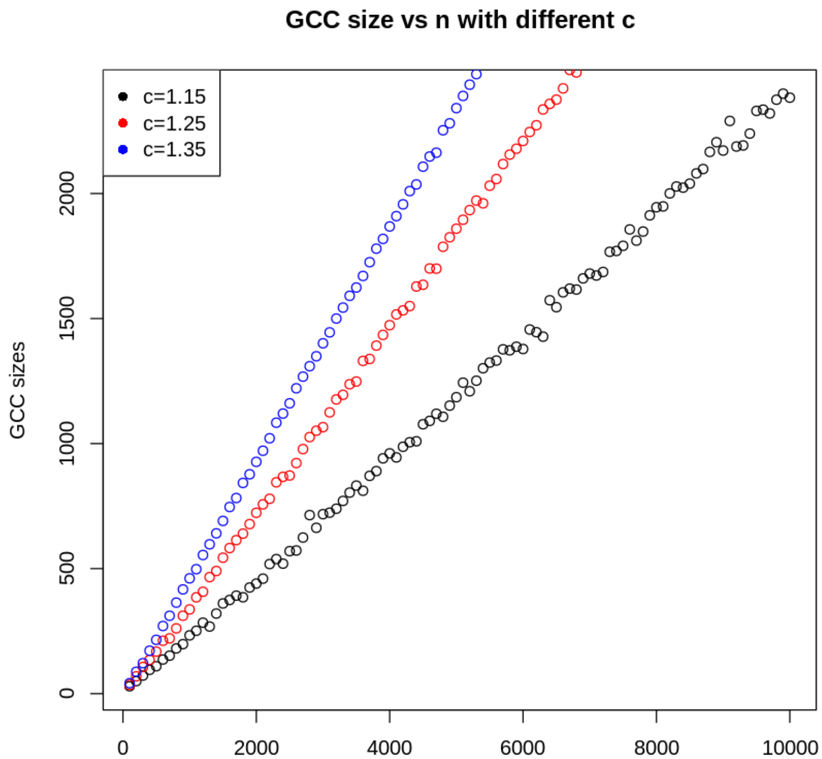
From the plot, we can find that when p starts with 0.0007, the line of the plot begins to merge to 1. Also, by calculating the theoretical value, the upper bound is p = ln(n)/n = ln(900)/900 = 0.0075 which is closed to the 0.0007. In addition, we can estimate the GCC takes up over 99% if the nodes at the value of p = 0.006.

1.1.d



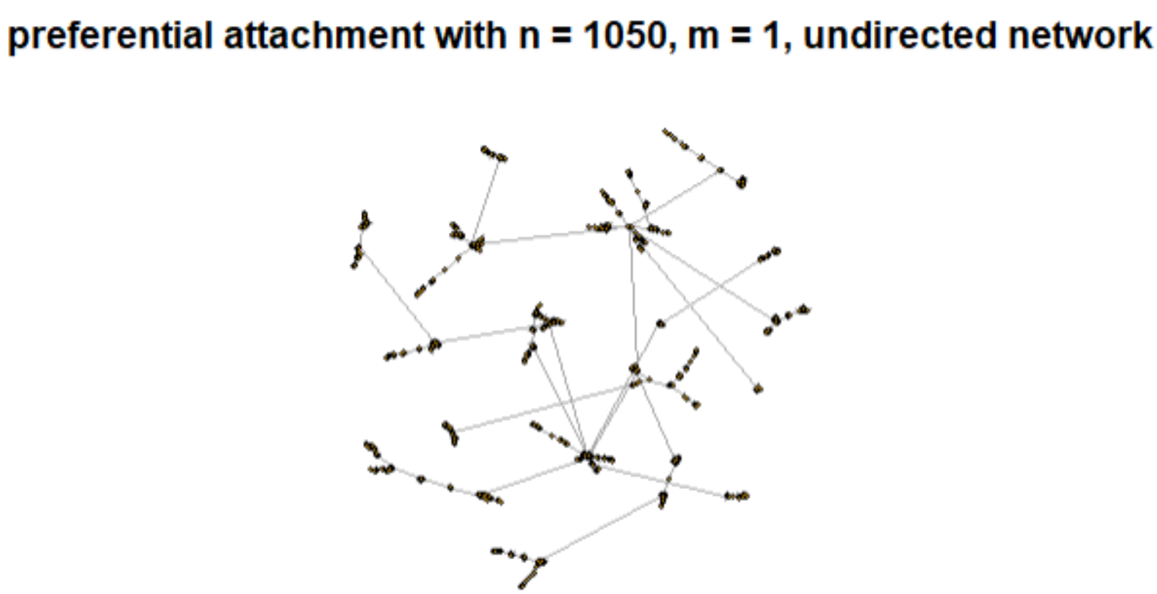
From the plot, the whole trend is increasing, and the slope is decreasing as n increase.





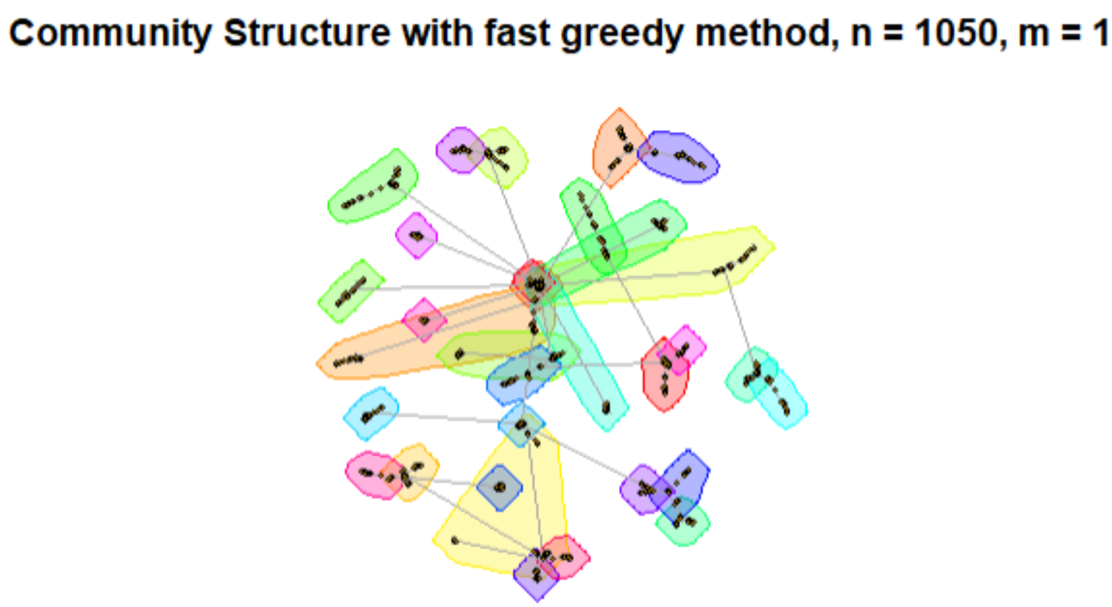
As the c increases, the trend of GCC sizes will be more linear and the slope of the line will increase.

1.2.a



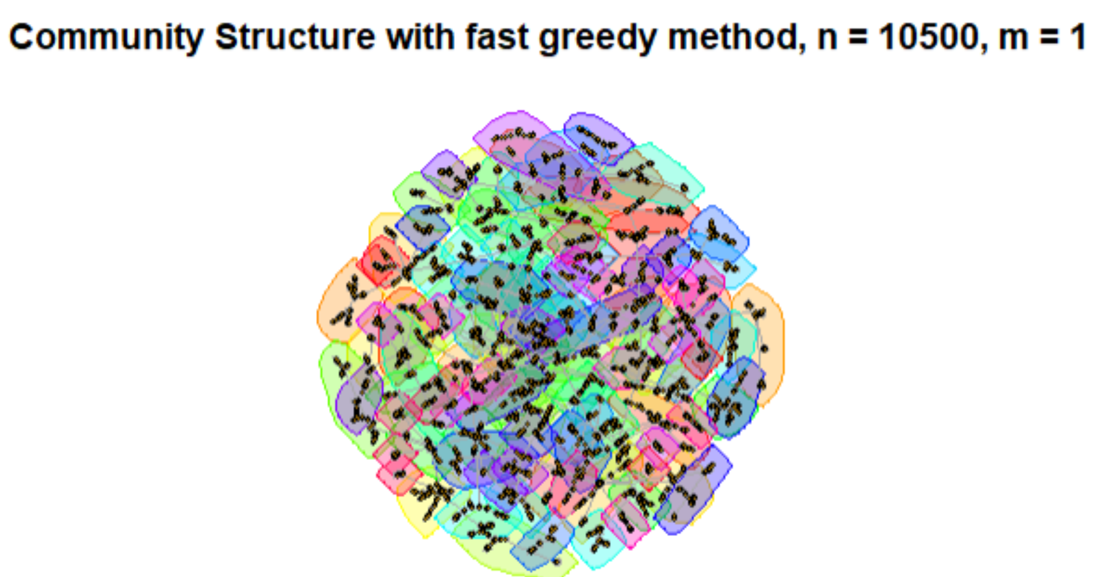
Yes, the undirected network with n= 1050 nodes and m = 1 old nodes is always connected.

1.2.b



Modularity: `0.931872`
Assortativity: `0.056679`

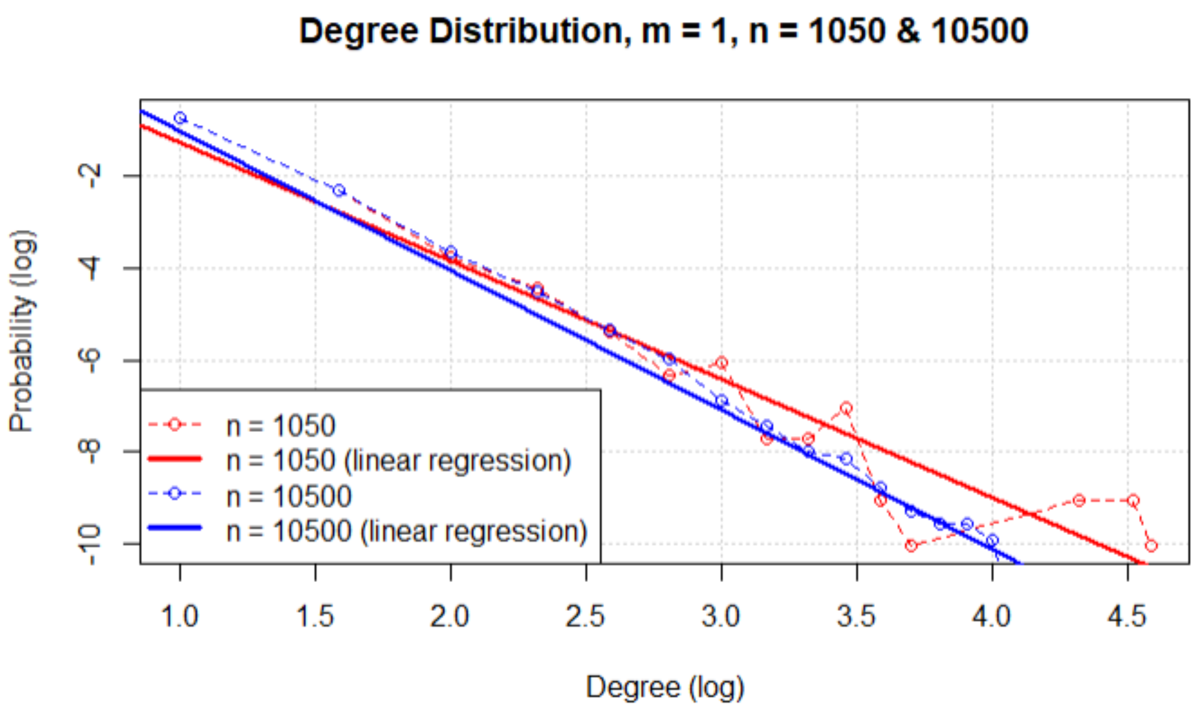
1.2.c



Modularity:`0.978946`

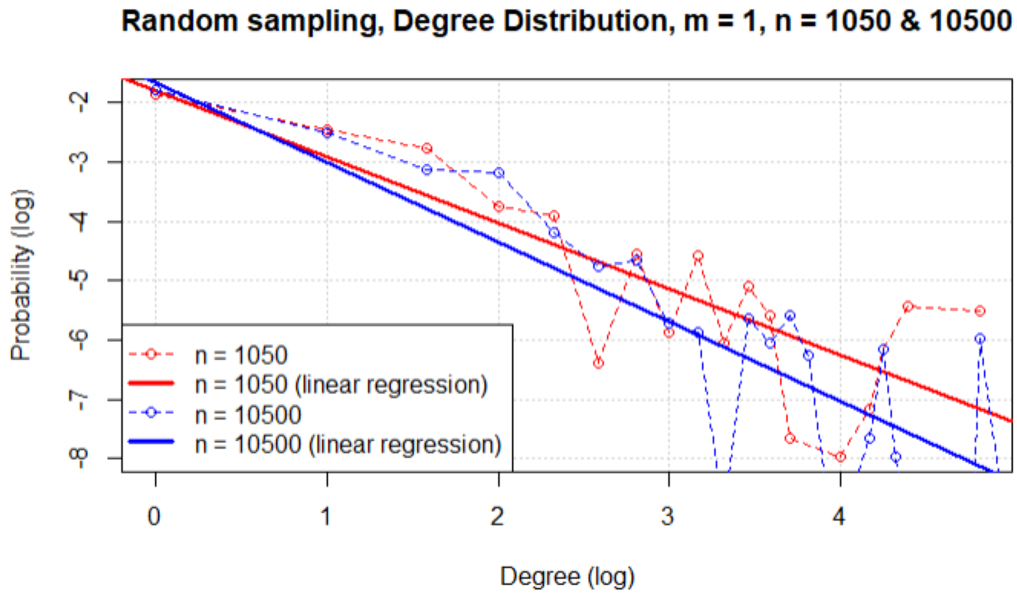
Assortativity: `-0.047238`
As we see, there are more communities in the network with the larger number of nodes compared to the smaller number of nodes. In addition, the value of modularity will be larger in the network with the larger number of nodes.

1.2.d




```
"slope with m = 1, n = 1050 is :-2.579081"
"slope with m = 1, n = 10500 is :-3.028255"
```

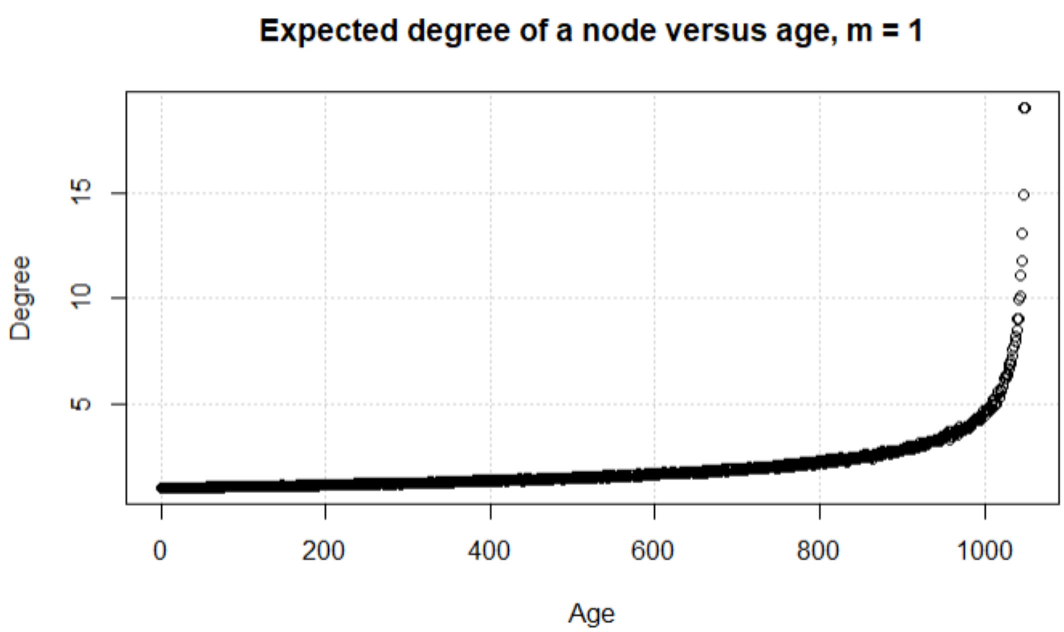
1.2.e




```
"slope with m = 1, n = 1050 is :-1.116781"
"slope with m = 1, n = 10500 is :-1.344294"
```

The distribution in the log-log scale is roughly linear with larger variance. Alos, it has a larger slope than that in the node degree distribution because in the log-log scale, we chose the higher degree nodes which leads to the larger slope of linear.

1.2.f

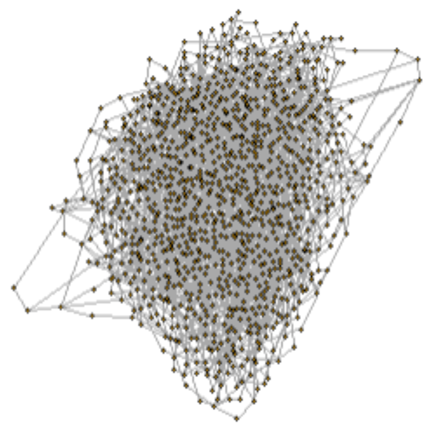


As the age increases, the expected degrees of a node increase.

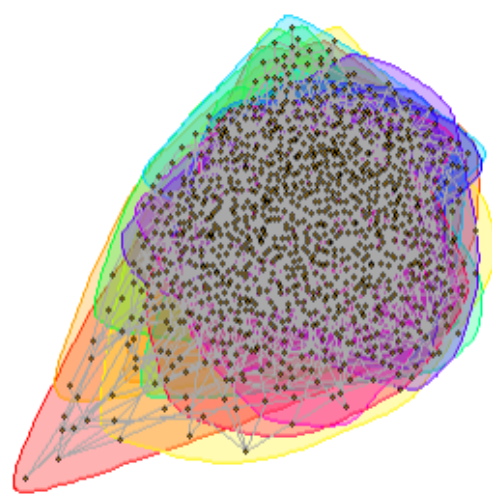
1.2.g

m= 2:

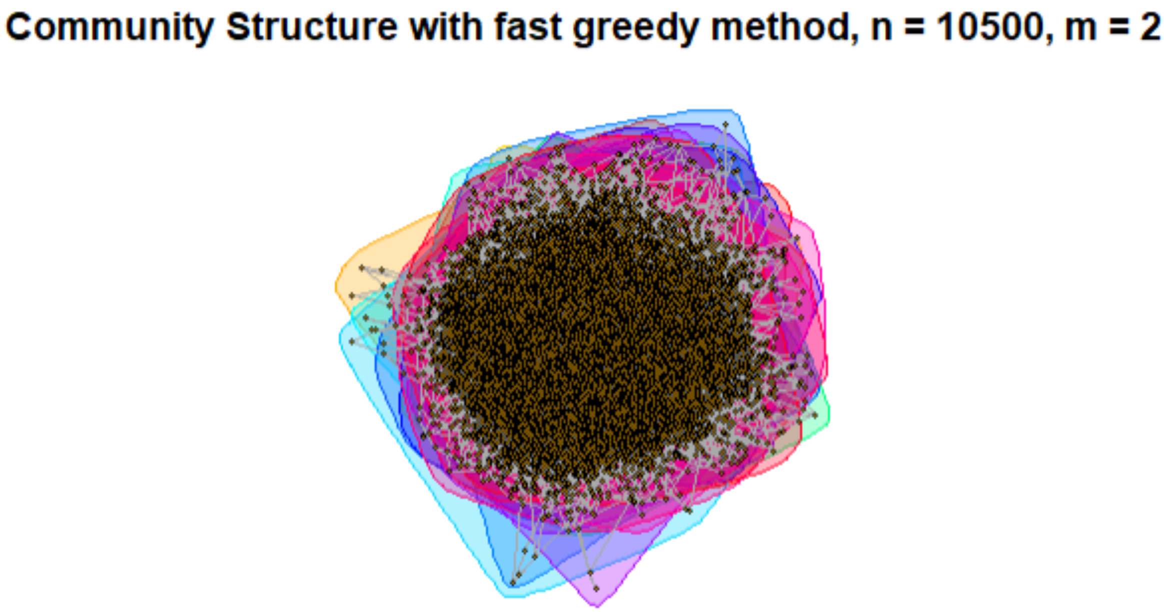
preferential attachment with n = 1050, m = 2, undirected network



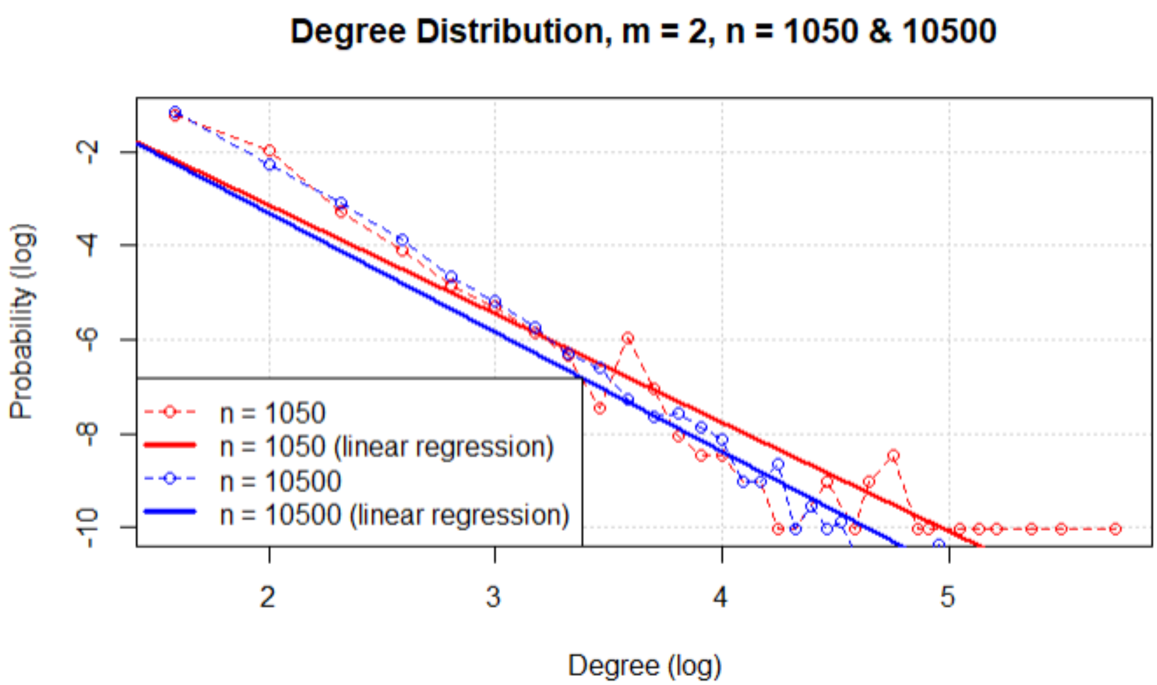
Community Structure with fast greedy method, n = 1050, m = 2



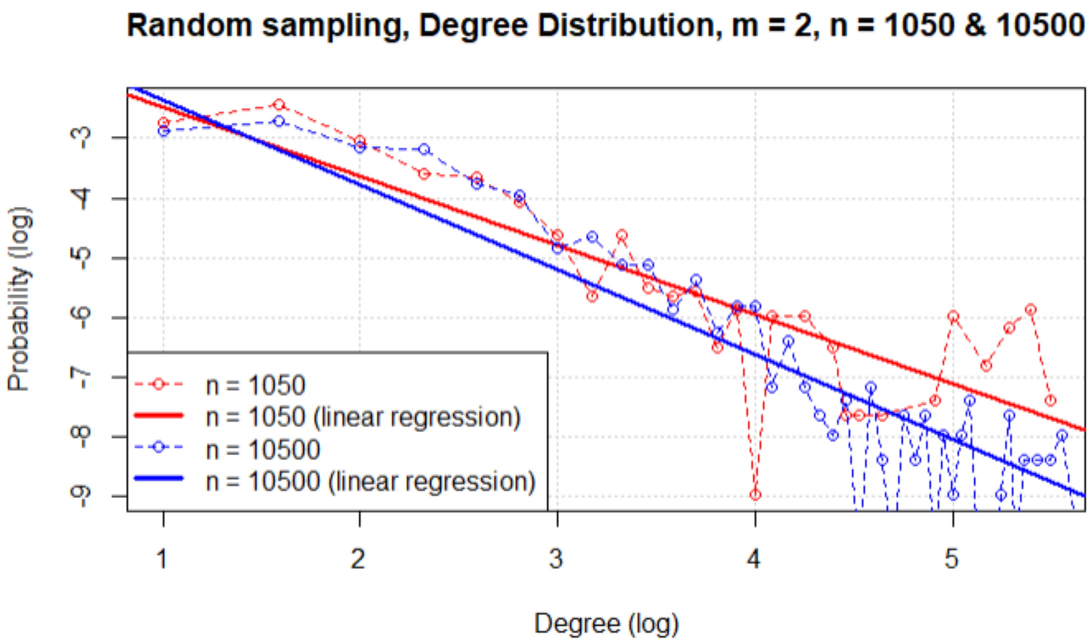
```
"Modularity is : 0.528910"
"Assortativity is :-0.029883"
```



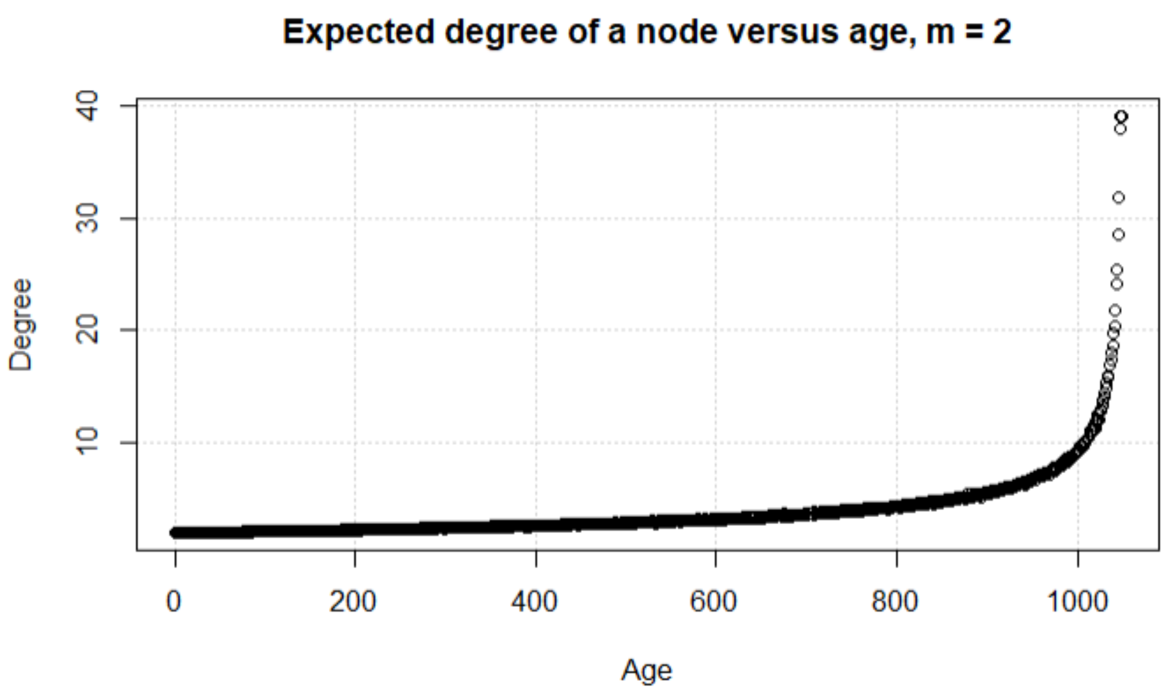

```
"Modularity with m = 2 is: 0.532819"
"Assortativity with m = 2 is:-0.008083"
```




```
"slope with m = 2, n = 1050 is :-2.322557"
"slope with m = 2, n = 10500 is :-2.551360"
```

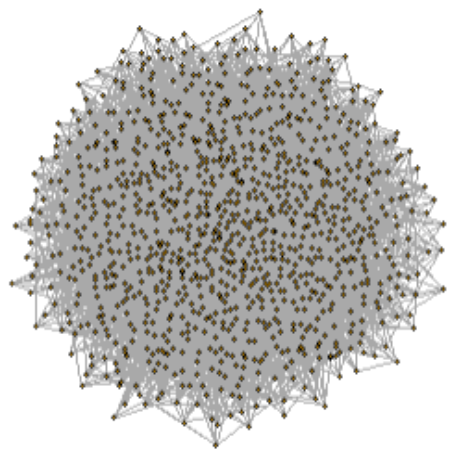



```
"slope with m = 2, n = 1050 is :-1.155003"
"slope with m = 2, n = 10500 is :-1.418722"
```

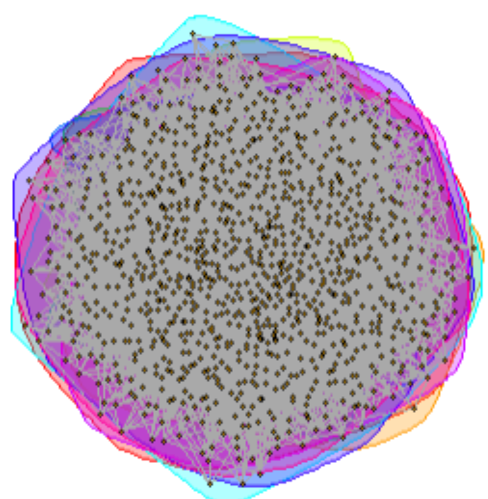


m = 6:

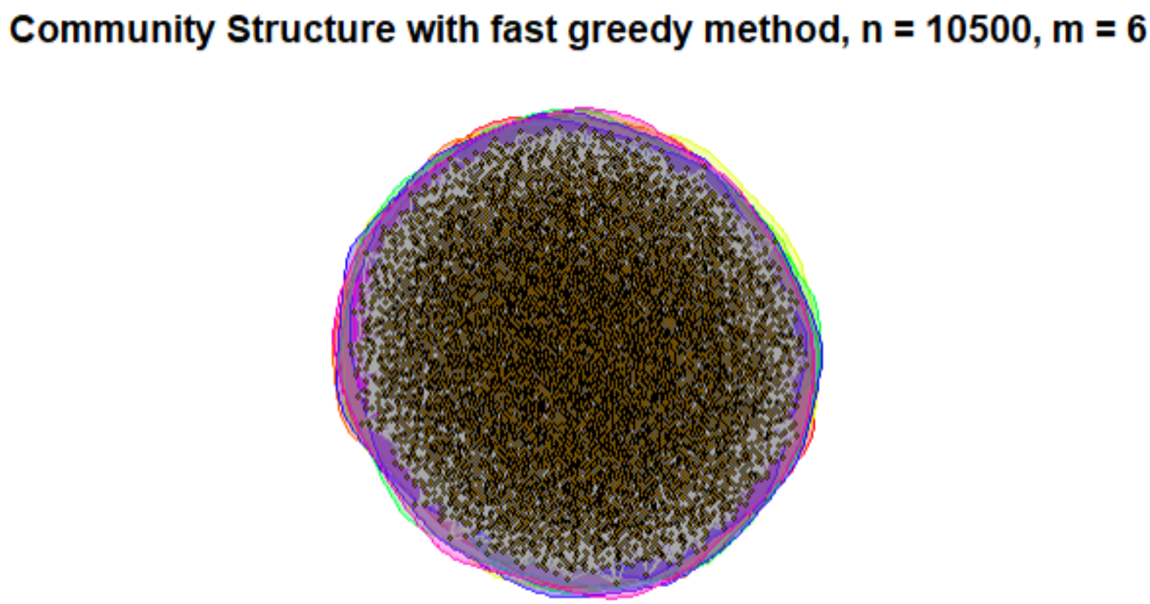
preferential attachment with n = 1050, m = 6, undirected network



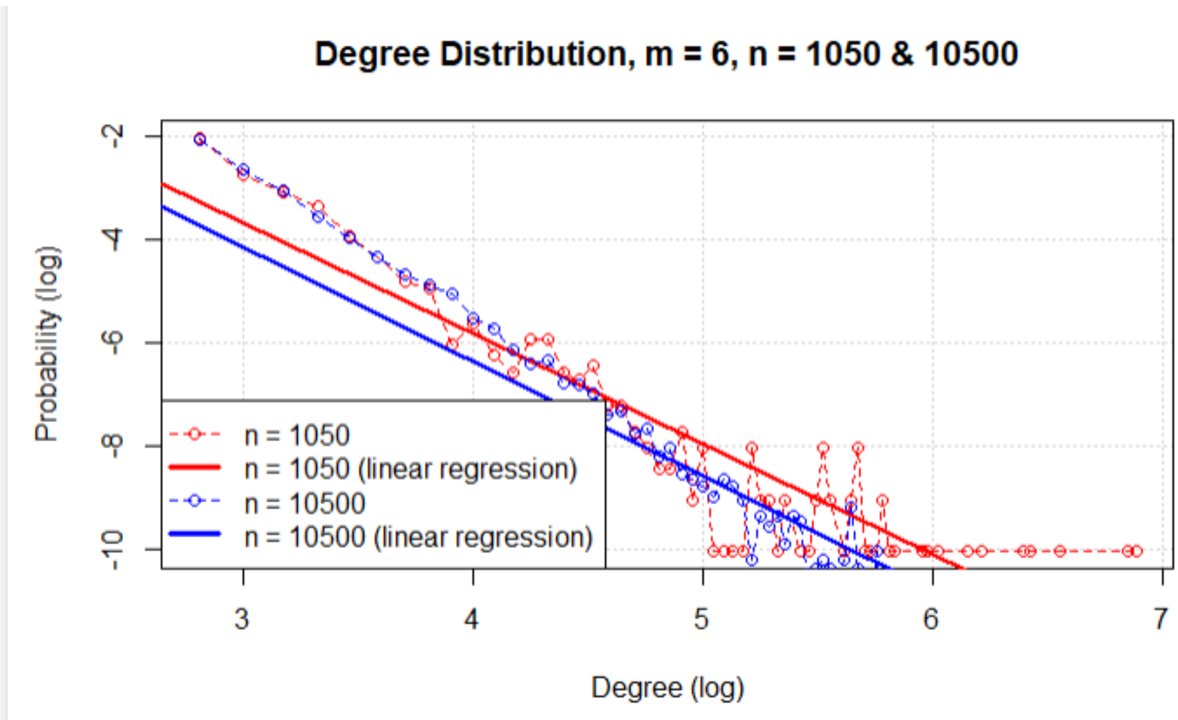
Community Structure with fast greedy method, n = 1050, m = 6



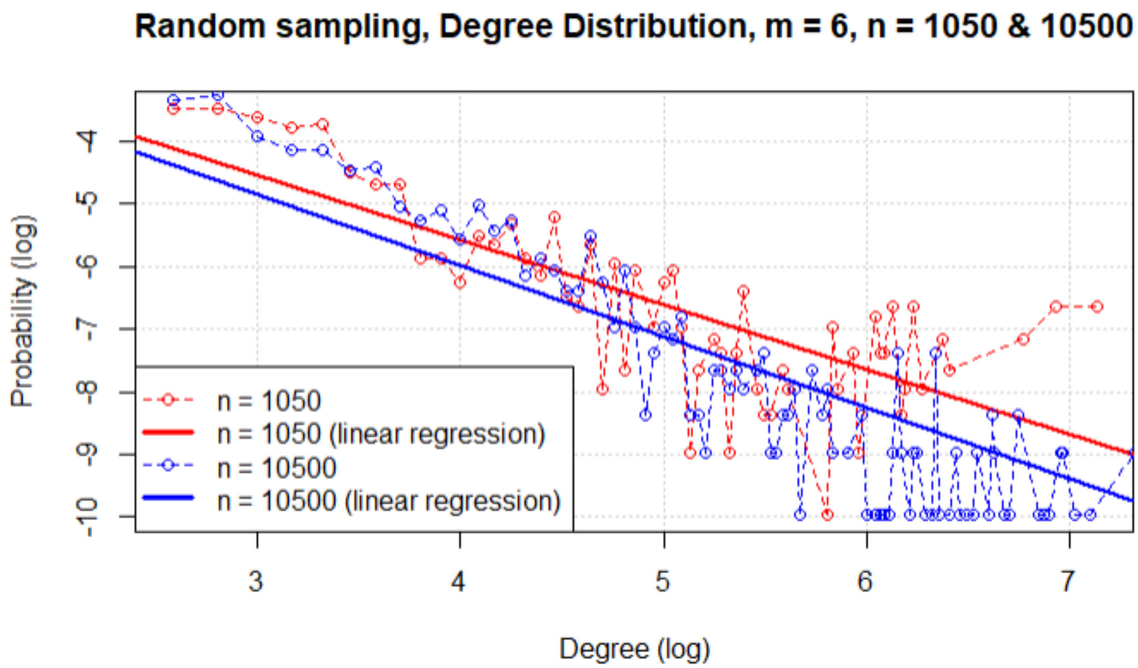
```
"Modularity is : 0.248476"
"Assortativity is :-0.019060"
```



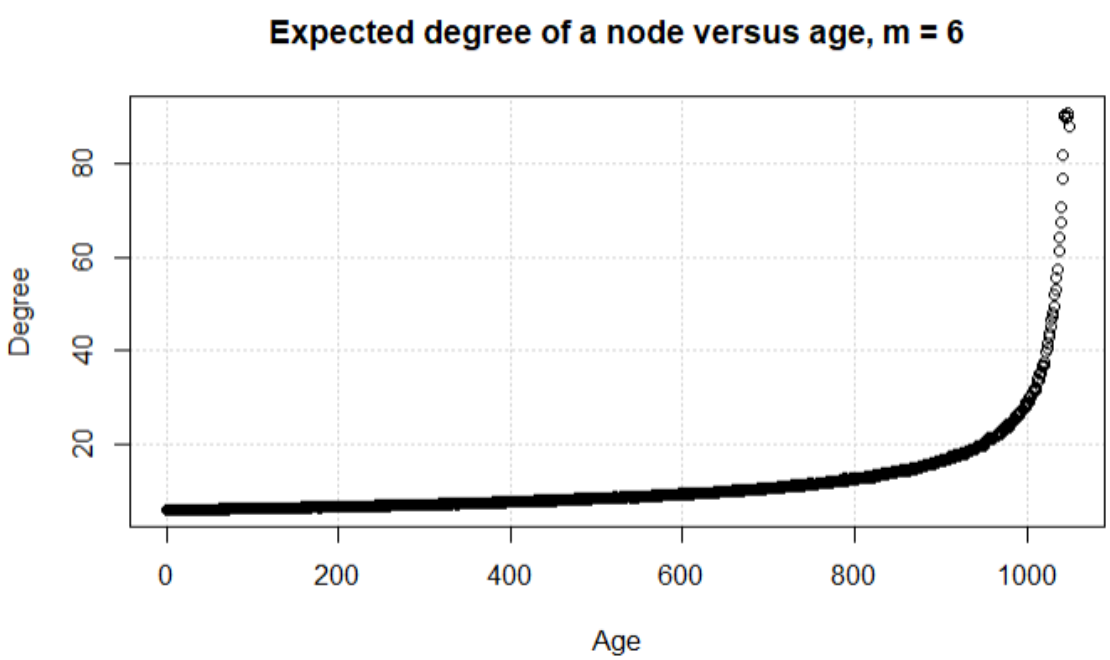
"Modularity with m = 6 is:0.246149"
"Assortativity with m = 6 is:-0.001978"



"slope with m = 6, n = 1050 is :-2.141910"
"slope with m = 6, n = 10500 is :-2.216066"

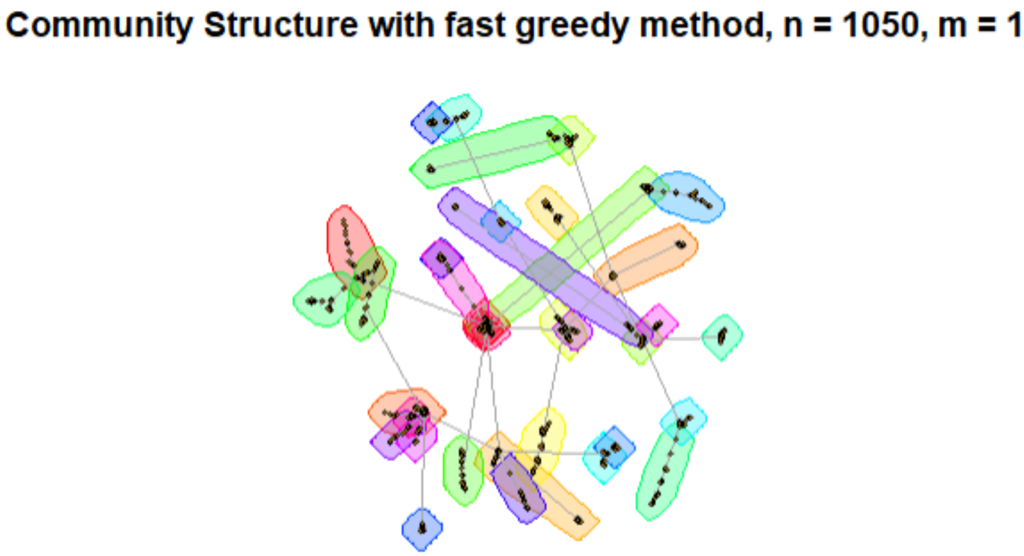



```
"slope with m = 6, n = 1050 is :-1.038103"
"slope with m = 6, n = 10500 is :-1.136519"
```

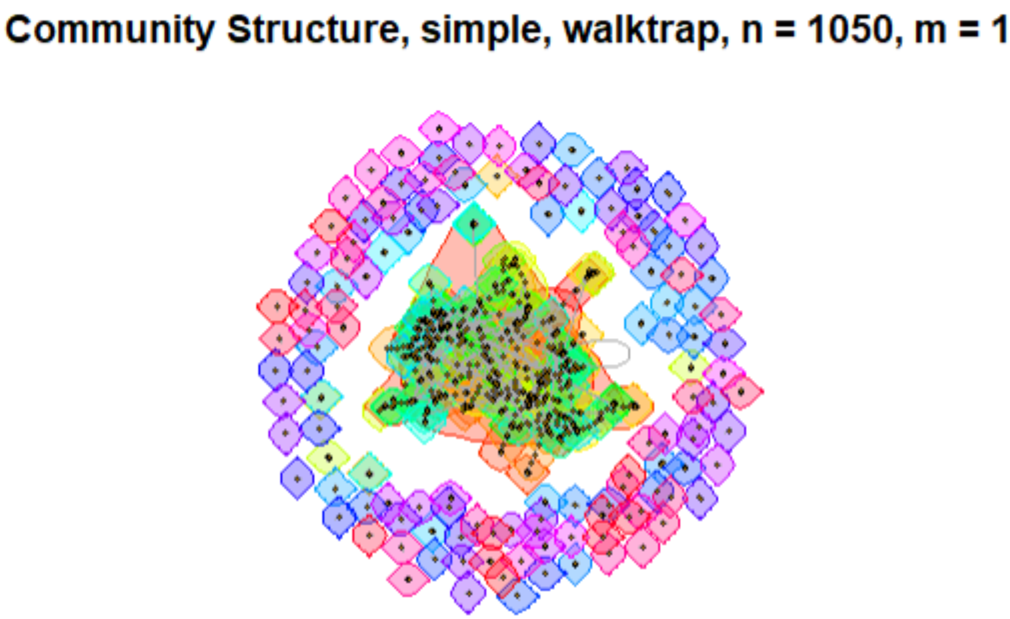


From above plots, we can find the following conclusions:
Modularity: The value of Modularity for the community structure will decrease when the value of m increases.
Slope: In the log-log scale, the slope of linear regression for the plot of degree distribution will increase when m incareses.
Relation between the age of nodes and expected degree will not change with increase of m.

1.2.h



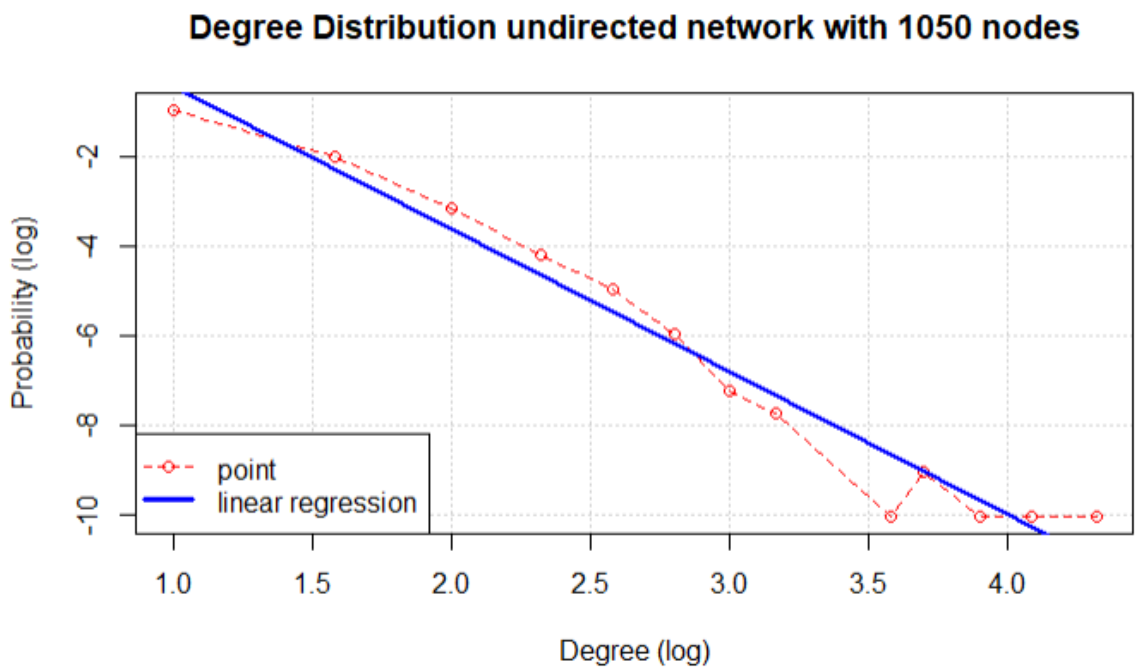
"Modularity is : 0.930233"



"Modularity is : 0.727700"

Using stub-matching can help to create networks with customized degree distributions, but it will have sparse connections between nodes with lower modularity. Thus, while stub-matching is advantageous for degree control, it can make community detection difficult.

1.3.a



SInce `"slope undirected network with 1050 nodes is :-3.188906"`, the power law exponent is 3.188906.

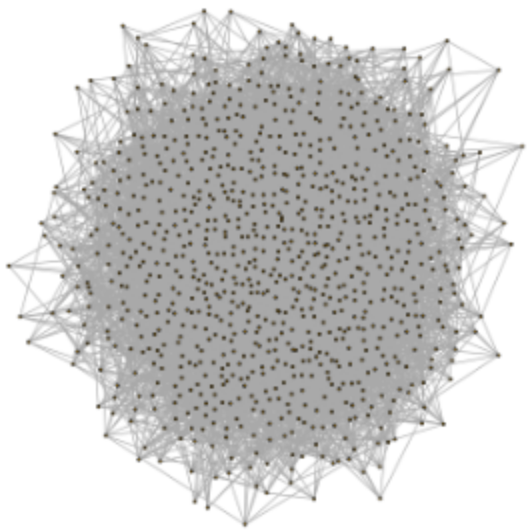
1.3.b

Community Structure with fast greedy method

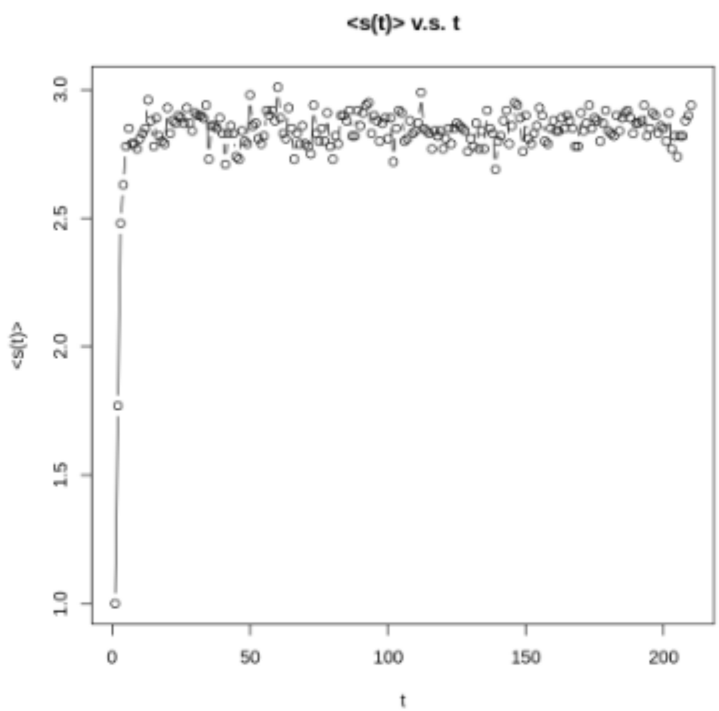
`"Modularity is:0.936141"`

2.1.a

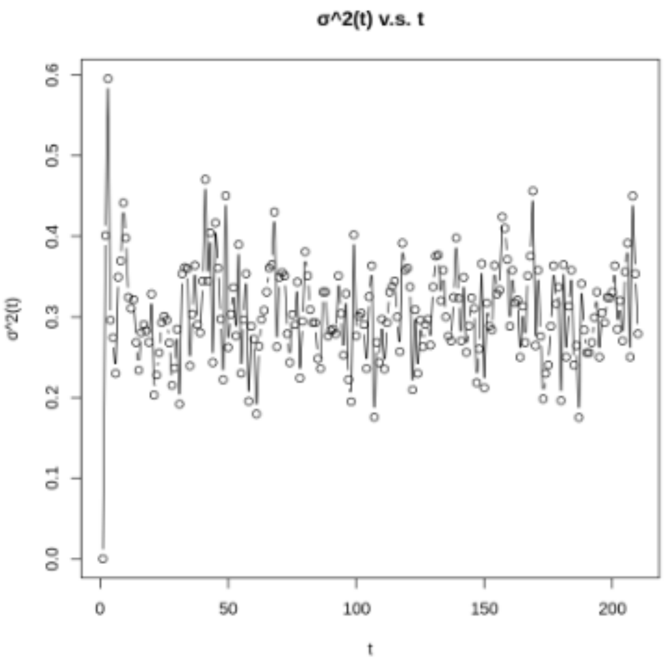
Undirected random network with n = 900, p = 0.015, diameter = 4



2.1.b

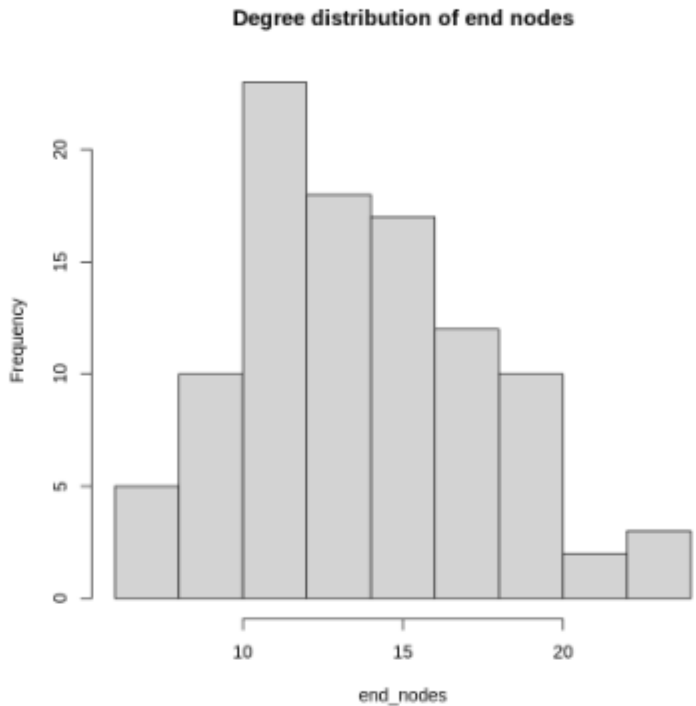


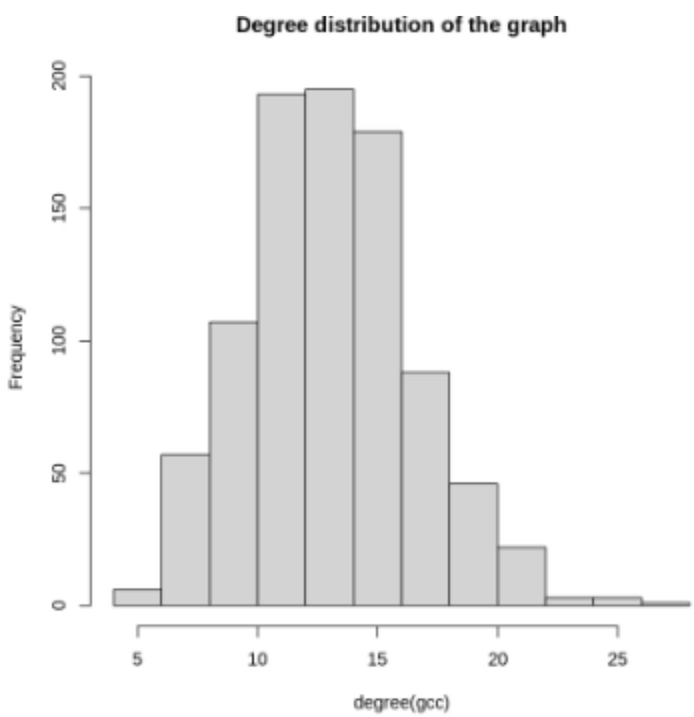
The average distance increases rapidly before t = 10, but after that, the average distance starts to converge to the 2.9.



The variance of this distance has large variation at the beginning, but after t =10, the value of variance starts to close to 0.3.

2.1.c

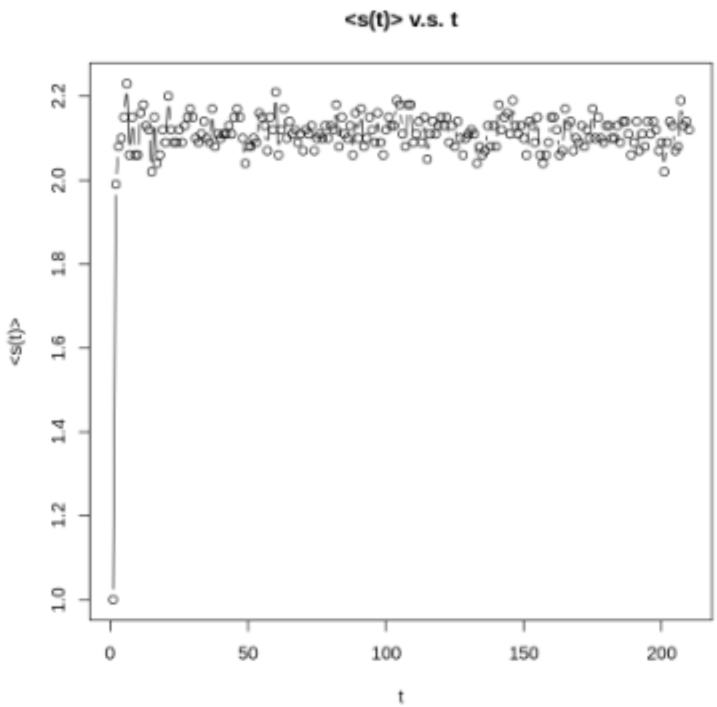


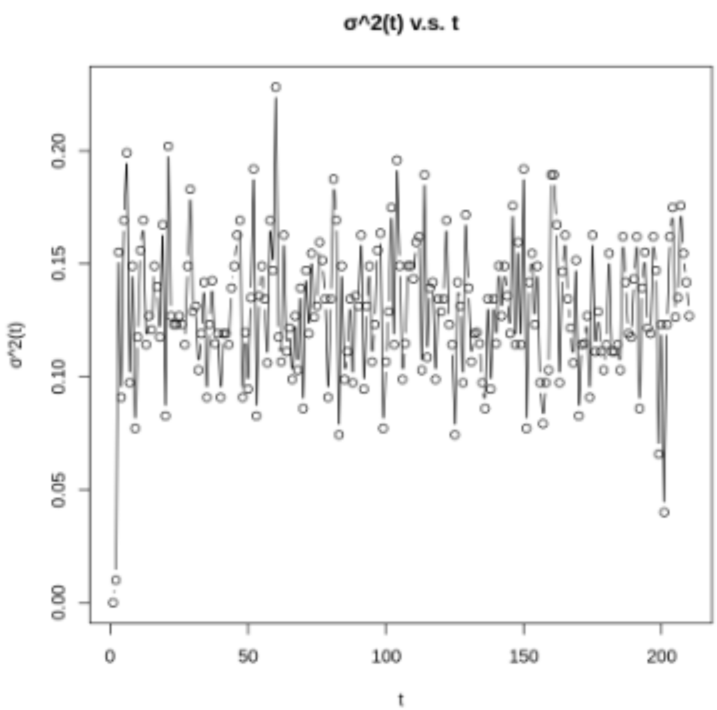


We observe that the degree distribution of the end nodes follows the degree distribution of the graph with similar mean and variance.

2.1.d

Degree distribution of the nodes with n = 9000, p =0.015, diameter = 3



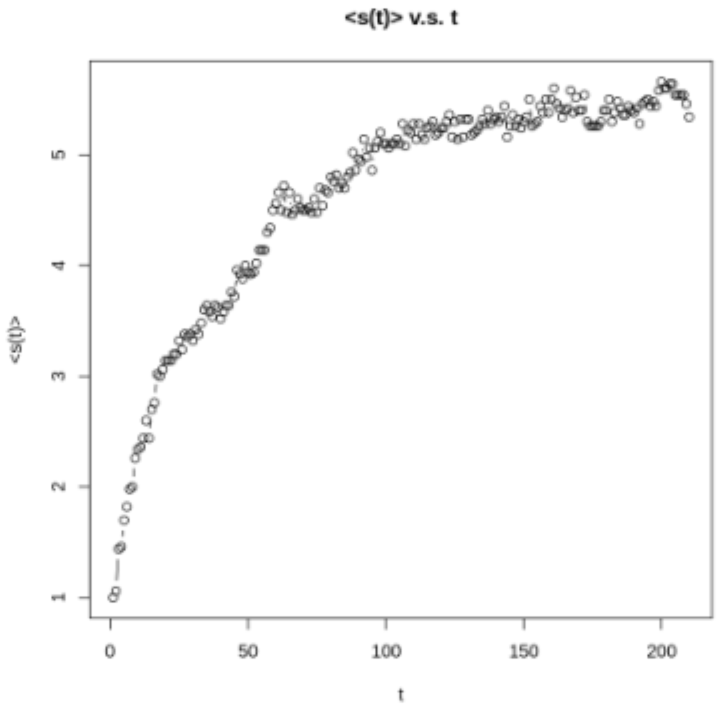


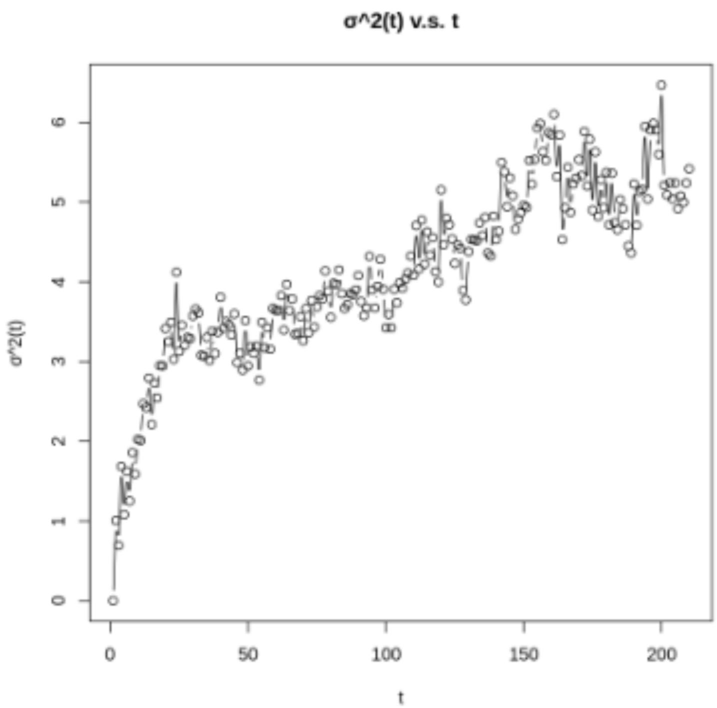
The diameter of this network is 3 and the diameter of the network with n= 900 is 4. Also, we can observe that both the convergence of the average distance and variance are less than that in the network with n=900. Thus, the lower diameter of the network can lead to the smaller the convergence of the average distance and variance.

2.2.a

Undirected preferential attachment network, n=900, m=1, diameter = 23

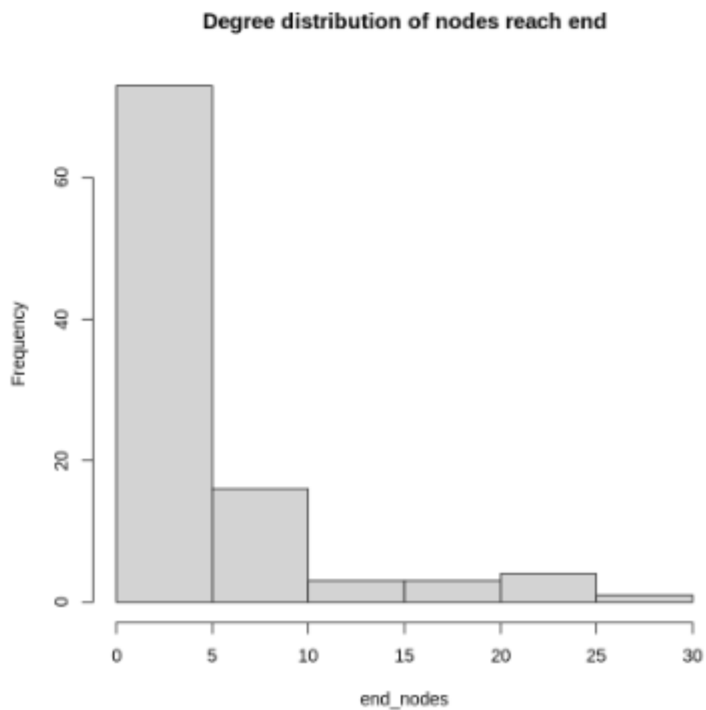
2.2.b

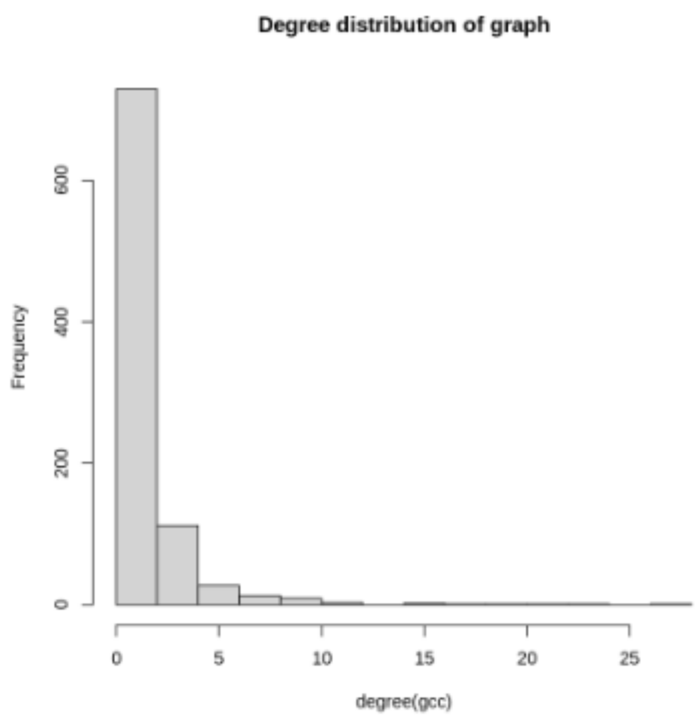




Both the average distance and the variance will increase as the t increases, and the slope of the plot decreases with t increase. In addition, the value of variance has more variation than the value of average distance.

2.2.c

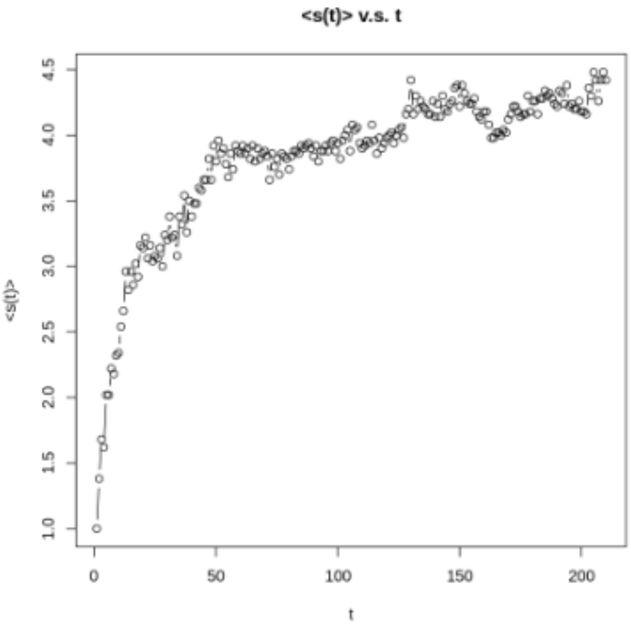


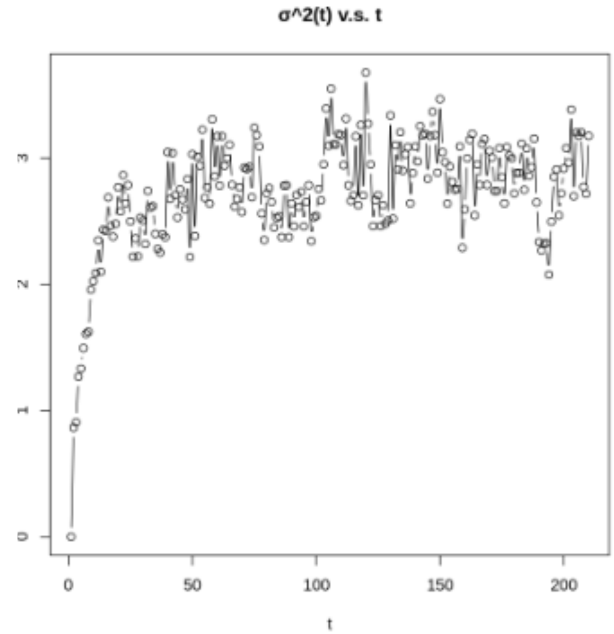


We observe that the degree distribution of the end nodes follows the degree distribution of the graph with similar mean and variance.

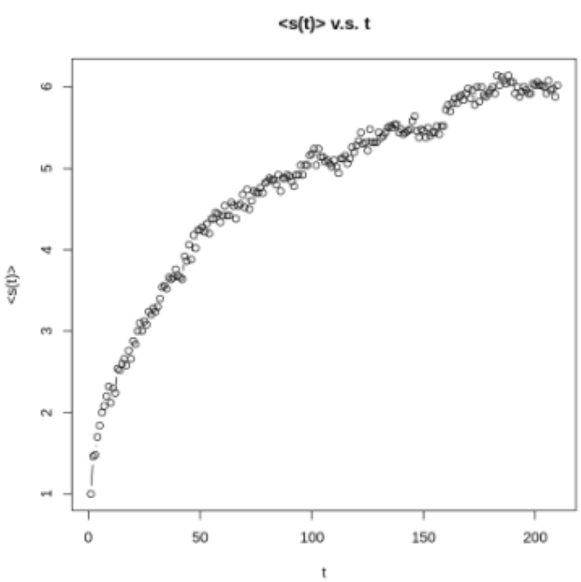
2.2.d

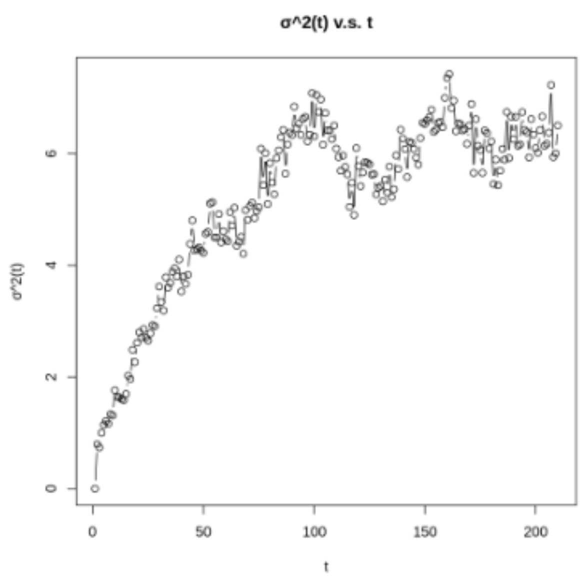
Undirected preferential attachment network, n=90, m=1, diameter = 11





Undirected preferential attachment network, n=9000, m=1, diameter = 27





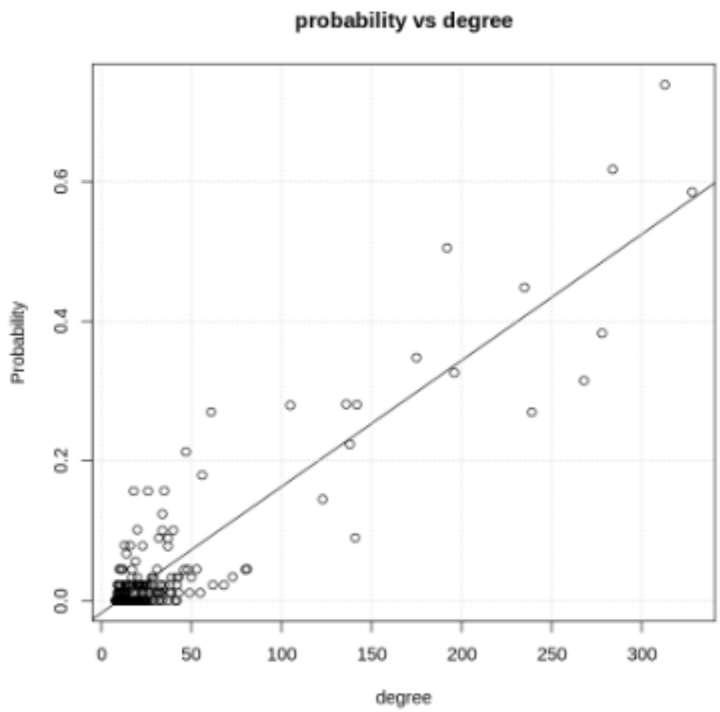
The diameter of the different network:

n = 90: 11
n = 900: 23
n = 9000: 29

From these plots, we can observe that as the diameter of the network increases, the maximum of average distance and variance will be larger. Take variance as an example: maximum of variance is around 4.5 when n=90 with diameter is 11, but when n = 9000 with diameter of network is 29, the maximum of variance will be around 6.

2.3.a

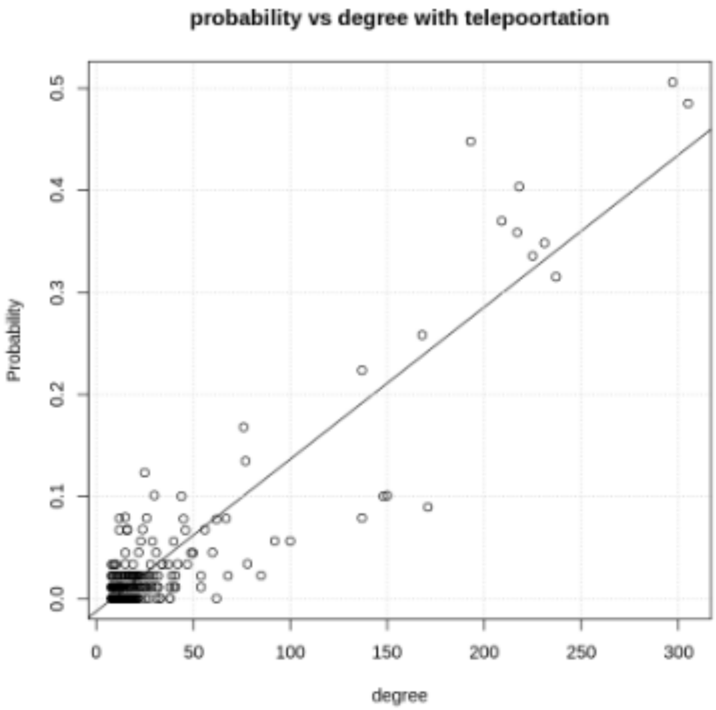
```
(Intercept)
-0.01760351
degree(combined_g)
       0.001806689
```



The plot shows that there is a positive linear relationship between probability and the degree of the nodes. This means the nodes with higher degree will have higher probability to be passed by the random walker.

2.3.b

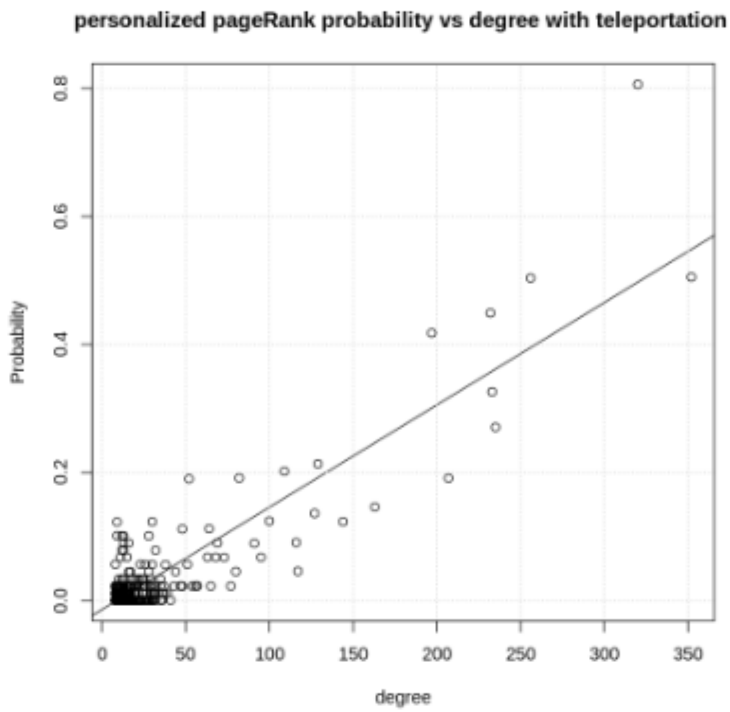
```
(Intercept)
-0.01256602
degree(combined_g)
       0.001490969
```



We chose teleportation probability of alpha = 0.2, then there still is a positive linear relationship between probability and the degree of the nodes. Compared to the previous question, we can find if we use a teleportation probability of alpha = 0.2, the slope of the linear regression will be lower than the previous question without teleportation probability.

2.4.a

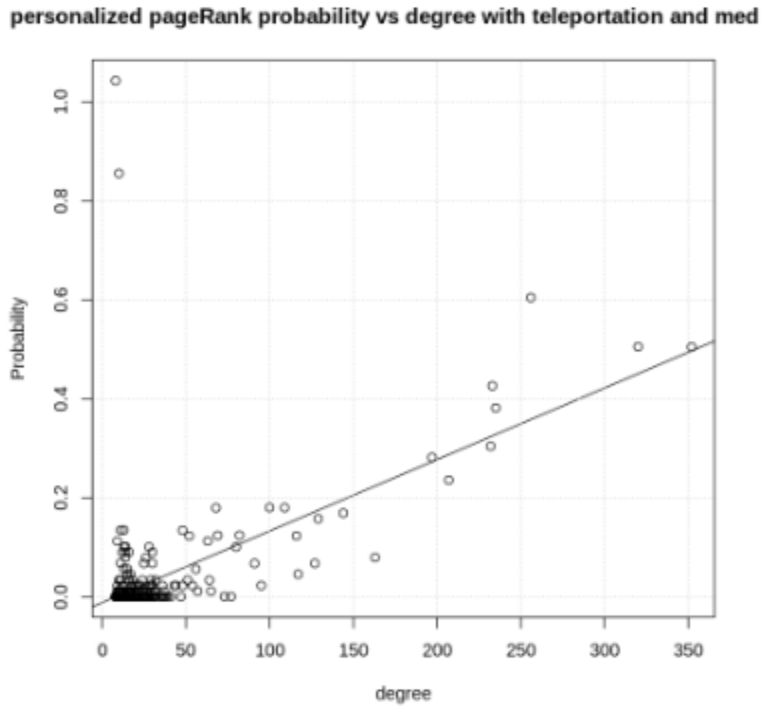
```
(Intercept)
-0.01428475
degree(combined_g)
       0.001598689
```



From the plot, we can observe that there is a positive linear relationship between the probability and the degree of the nodes. Compared to the slope of the line for regular pagerank without teleportation, the personalized pagerank with teleportation has a lower slope of line. This means the teleportation will reduce the probability of passing the connected nodes.

2.4.b

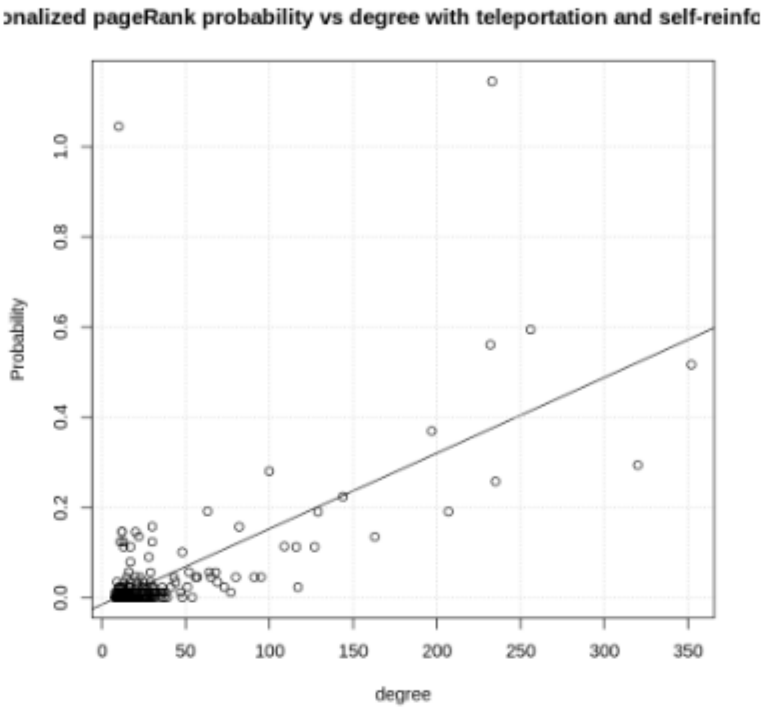
```
(Intercept)
-0.01182569
degree(combined_g)
        0.00144457
```



Since we only chose two nodes which is a small number, this will lead to the pageranks increasing rapidly. Then, the nodes with lower degree will be more likely to be passed. Thus, even though we can find a positive linear relationship between the probability and the degree of nodes, it is a weaker linear relationship.

2.4.c

```
(Intercept)
-0.01556122
degree(combined_g)
        0.00167869
```



Because normally, we need to consider the pagerank. By Taking this into account, we set the probability of teleportation lands on the two nodes we get from part(b) to be 0.7. And the probability of rest of the nodes to be 0.3. And according to the pagerank we get from part(a), we can easilty spread out the 0.3 probability to rest of the node. And from the graph, the slope we can see it gets bigger. Thus by taking this way, it impacts pagerank and median of pagerank.