Large Scale Social Network : Models and Algorithms
Project #1

Due on April 19th, 2020 at 11:59pm

Professor Roychowdhury, Vwani

Wang, Yin, He

PART I: GENERATING RANDOM NETWORKS

Question 1

In this part, we are going to study various properties of random networks generated by the ER model. To begin with, the Erdos-Renyi model is a simple model to generate an undirected graph. There are two parameters associated with this model, n (int) is the number of nodes and p (float) is the probability for edge creation. With this notation, we can say that in this model, a graph in G(n,p) has on average $\binom{n}{2}p$ edges, and the distribution of the degree of any particular vertex is binomial:

$$P(deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{1.1}$$

(a) In this part, we first plot the degree distributions for different probability p, and the resulting plots are shown below as Figure 1.

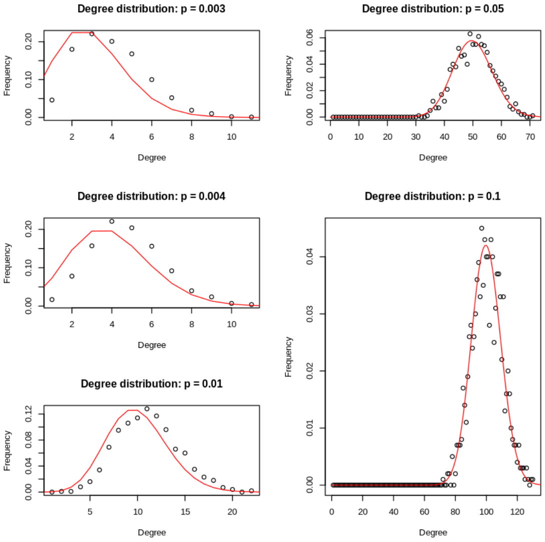

Figure 1: Degree distributions of Erdos-Renyi model for different edge generation probability.

We can see from the plots that the distribution is a binomial distribution. Notice that if np is sufficiently large, the binomial distribution can be seen as the normal distribution with the same variance and expectation, which is emplicitly shows in Figure 1 with $p = 0.05, 0.01$. Next, we print out the mean and variance of each degree distributions and compare them to the theoretical values. Since we observed that it is a binomial distribution, then the theoretical value of mean and variance can be calculated as

$$E\left[deg(V)\right] = np; \; Var(deg(V)) = npq = np(1-p) \tag{1.2}$$

The results are shown in the following table as Table 1.

(b) In this part, in order to numerically estimate the probability that a generated network is connected, we choice to generate the network with specified parameters 500 times in order to calculate the probability. The results are shown in Table 2.

We can see from the results that when $p = 0.003$ and $p = 0.004$, there does not exist any instance that the graph is connected, meaning that the possibility for the connectivity is really small; when $p = 0.01$, there is

Question 1 continued on next page. . .

p-value	theorical mean	empirical mean	theoretical variance	empirical variance
0.003	3	2.964	2.991	2.857
0.004	4	3.79	3.984	3.581
0.01	10	9.962	9.8089	9.904
0.05	50	50.222	48.418	49.614
0.1	100	99.746	91.351	88.429

Table 1: Comparison between theoretical and empirical characteristics for different p.

p-value	Connected Frequency	GCC Diameter
0.003	0	15
0.004	0	10
0.01	0.964	6
0.05	1	N/A
0.1	1	N/A

Table 2: 500 experiments' result on the connectivity and the GCC diameter of Erdos-Renyi model.

a probability of 0.964 that the graph is connected, and when $p = 0.05$ and $p = 0.1$, the graph is always connected, showing that for high p, the possibility of connectivity is so high that can be seen as 1. For the GCC diameter, when $p = 0.003$, it is 15, when $p = 0.004$, it is 10, and when $p = 0.01$, it is 6. Thus, we observed that the probability that the graph is connected increases as p-value increases, and GCC diameter decreases as p-value increases, which makes sense obviously. For completeness, the GCC plots are shown as Figure 2, 3, 4.

(c) According to the instruction, the normalized GCC size is a highly nonlinear function of p, with properties that $p = O(1/n)$ and $p = O(ln(n)/n)$. In this part, for $n = 1000$, we are going to sweep over values of p from 0 to a max value of p that makes the network almost surely connected and create 100 random networks for each p. We need to first empirically estimate the value of p where a GCC starts to emerge. To do this, we make a scatter plot that plots the normalized GCC sizes vs p. The plot is shown as Figure 5.

We can see from the plot that roughly at $p = 0.007$, the slope begins to be flat, meaning that the GCC starts to emerge. Now, according to the theoretical calculation, considering the upper bound, $p = \ln(1000)/1000$, which gives us the result that $p = 0.0069$, which is very close to our empirical result. Also, from the plot, we can see that roughly at p = 0.005, the GCC takes up over 99% of the nodes.

(d) In this part, we first define the average degree of nodes $c = np = 0.5$, and sweep over the number of nodes ranging from 100 to 10000, and plot the expected size of the GCC with n nodes and $p = c/n$ as a function of n. The resulting plot is shown as Figure 6.

We can see from the plot that the slope is extremely large from n = 0 to 2000, and from n = 2000, the slope seems to decrease to a linear fashion, which is very similar to the observation made in the previous question. Now, we repeat the same procedure for $c = 1$, and the result can refer to Figure 7.

Now we see that as c increases, the trend seems to be more linear, but it is still not linear in the region from $n = 0$ to 2000. Next, we repeat the same procedure for $c = 1.1, 1.2, 1.3$, and plot the results in a single plot, which is shown as Figure 8

Question 1 continued on next page. . .

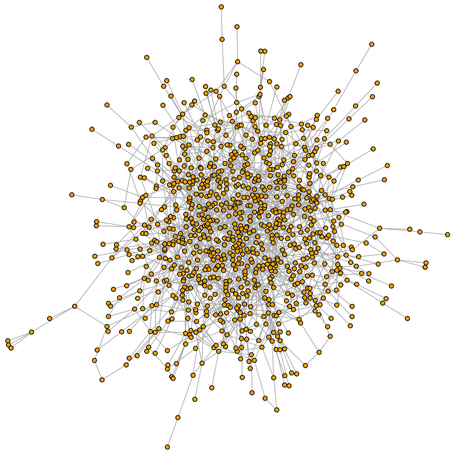

Figure 2: One instance of GCC for Erdos-Renyi model when $p = 0.003$

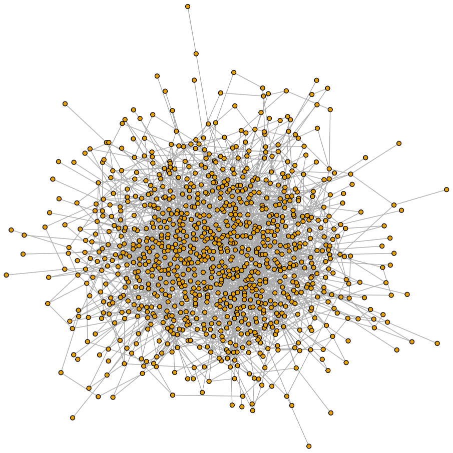

Figure 3: One instance of GCC for Erdos-Renyi model when $p = 0.004$

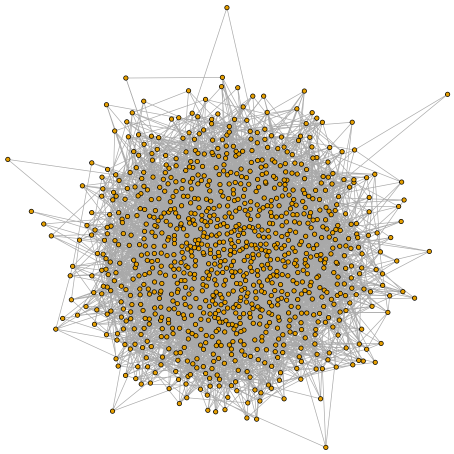

Figure 4: One instance of GCC for Erdos-Renyi model when $p = 0.01$

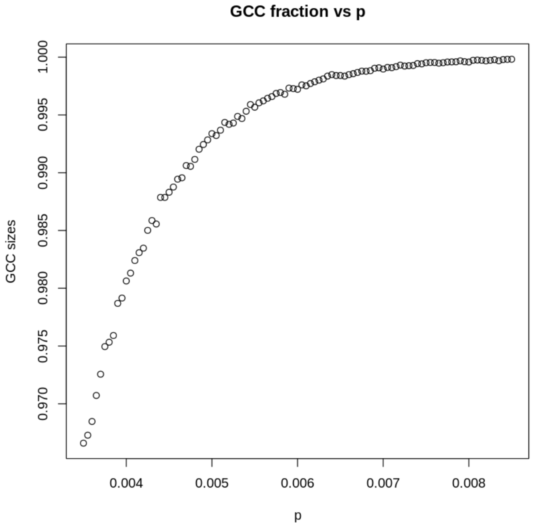

Figure 5: The normalized GCC size against the edge generation probability of Erdos-Renyi model

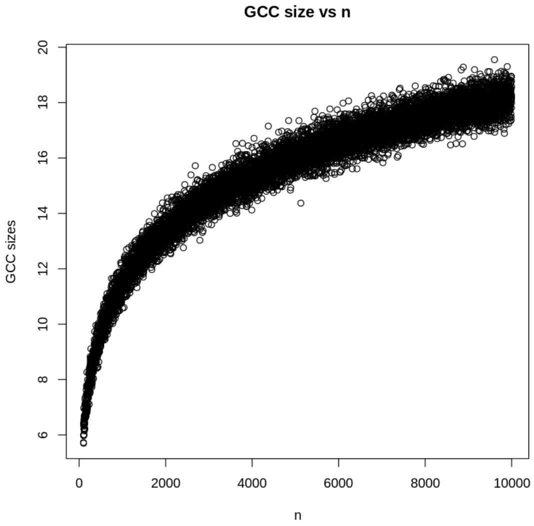

Figure 6: The empirical GCC size against the size of the graph for Erdos-Renyi model with $c = 0.5$.

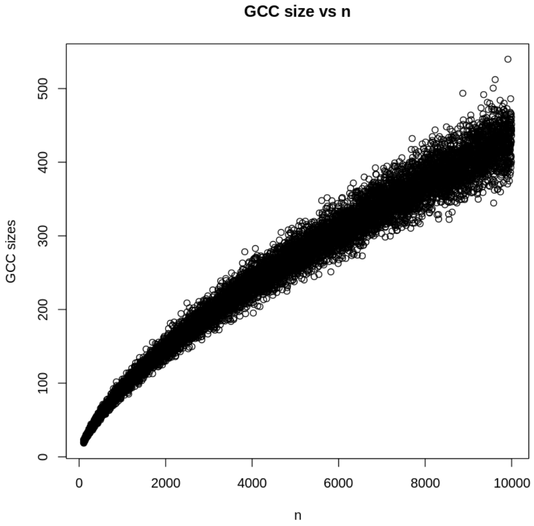

Figure 7: The empirical GCC size against the size of the graph for Erdos-Renyi model with $c = 1$.

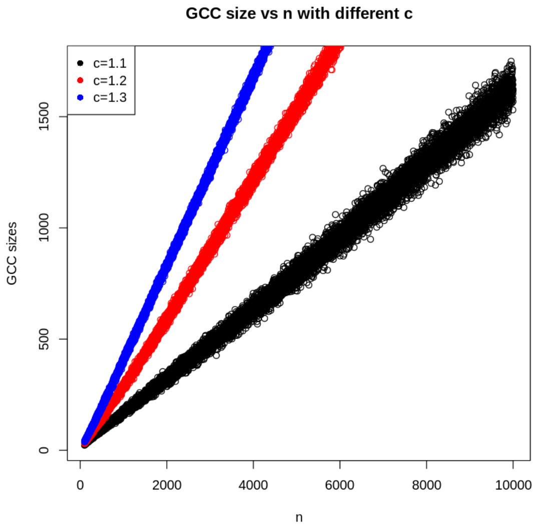

Figure 8: The empirical GCC size against the size of the graph for Erdos-Renyi model with $c > 1$.

From the plot, we can easily see that with higher c-value, which means higher p-value, will produce larger GCC size. And also notice that with higher c-value, the trend seems to be more linear.

Question 2

(a) According to the process of construction of preferential attachment network, the final network is definitely connected. Let us proceed the proof by induction. When $n = 2$, the network derived is connected. Assume that it is still connected when $n = k$. When a new node, V_{new}, is introduced, namely the size of the network is changed to $k + 1$, there will be m edges between the new node and previous nodes. Let us assume one of these edges connects the node V_i with V_{new}. To check the connectivity of new network, we just need to show there always exists path between V_{new} and V_j $(j \leq k)$. This is obvious for $j = i$. For any $j \neq i$, there exists path between V_i and V_j because of the connectivity of network with size k. Hence, there exists a path for any pair of vertices in new constructed network. By induction, we claim that the preferential attachment network is always connected.

(b) Before calculating the modularity, let us articulate the definition of it first. Generally speaking, modularity is a system property which measures the degree to which densely connected components within a system can be decoupled into separate communities or clusters which interact more among themselves rather than other communities. For a given division of the network's vertices into some modules, modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules. That is to say, the larger the modularity is, the betteer the clustering of the network is. Specifically, the modularity is given as below equation

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \frac{s_v s_w + 1}{2} \tag{2.1}$$

where $\left[A_{vw} - \frac{k_v k_w}{2m} \right]$ is the difference between actual number of edges between node v and w and the expected number of edges; s_v, s_w represent the memberships which communities the node v and w belong to respectively.

Using greedy strategy to cluster the preferential attachment model and *modularity* function in *igraph* package to gauge the modularity, we can get the result shown in Table 3.

(c) Using same strategy to cluster the network with 10000 vertices, we can derive the modularity of it shown as Table 3. According to the table, the modularity of larger network is superior to the network with size 1000. This trend is easy to understand by considering the meaning of modularity. As mentioned, the modularity represents the quality of a spericific separation of a network. When the separation method is fixed, to some extent, it can be a measurement to determine the degree to which a network can be clustered. A preferential attachment model with large size is usually easier to separate because of the preference connection between high-degree nodes and new added nodes, making its modularity larger.

(d) The concrete result for preference attachment model with $m = 1$ can refer to Figure 9. According to the annotation of the graph, we can easily derive the slope of two log-log plot as -2.179 and -2.494. According to the theorem, the steady state degree distribution for preferential attachment model can be seen as power-law distribution with $\gamma = 3$. Hence, with more nodes introduced, the slope of linear regression will gradually approach the value of -3.

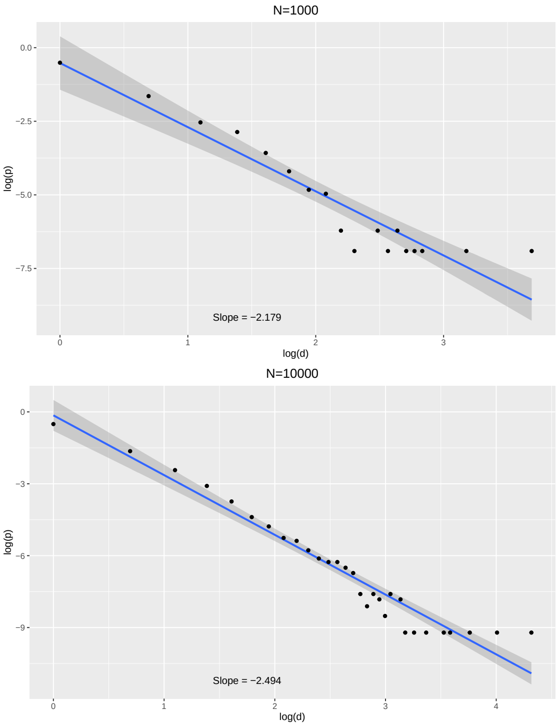

Figure 9: Degree distributions with different numbers of nodes for preferential attachment model $(m = 1)$

(e) One possible way to solve this question is to use Monte Carlo method. If the number of samples is large enough, then the empirical distribution can be extremely close to the true distribution we want to get. However there is another way to directly get the degree distribution of node j that are picked with the process mentioned in the problem. Since the node i and its neighbor node j are both randomly picked, the possibility of choosing V_i and V_j are given below

$$P(V_i\ is\ chosen) = P_i = \frac{1}{|V|};\ P(neighbor\ V_j\ is\ chosen) = P_{ij} = \frac{1}{deg(V_i)} \tag{2.2}$$

Then the possibility of node V_j is picked through mentioned process can be written as

$$P(V_j \ is \ chosen \ finally) = P_{V_j} = \sum_{(V_i,V_j)\in E} \frac{1}{deg(V_i)} \cdot \frac{1}{|V|} \tag{2.3}$$

After getting the probability of a specific node V_j is picked, then degree distribution derived from mentioned process can be represented as

$$P(the \ degree \ of \ randomly \ picked \ node = k) = P_k = \sum_{deg(V_i)=k} P_{V_i} \tag{2.4}$$

The degree distributions derived using above method for network with different sizes are shown as Figure 10. To some extent, the degree distribution can be seen as linear. The estimated slope of the plot using linear regression are -1.052 for network of size 1000 and -1.457 for network of size 10000. The slope of this degree distribution is larger than the original node degree distribution which is easy to be understood. In step one, we are much more likely to pick the nodes with lower degree, for example node with degree 1. However, in step 2, the neighbors of nodes with lower degree usually have high degree. For example, the neighbor of nodes with degree 1 must have at least two edges because of the connectivity of whole network. So in this process, the possibility of choosing high degree nodes improves comparing to initial method, making the slope of linear regression larger.

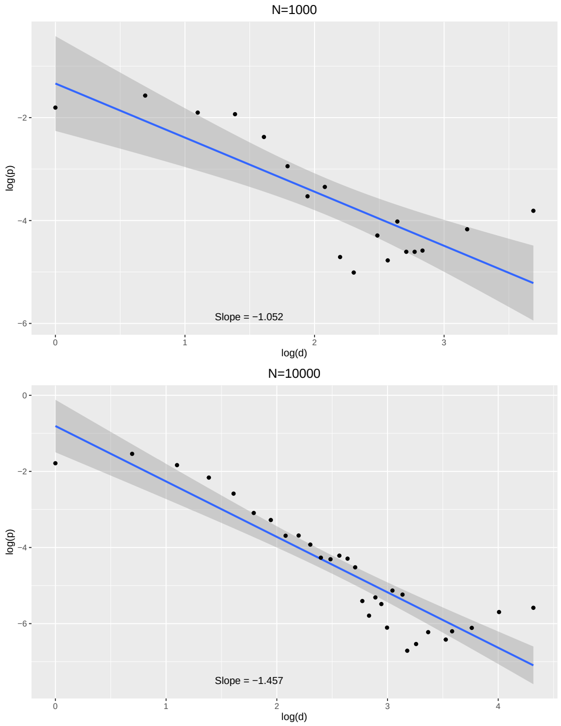

Figure 10: Degree distributions of V_j with different numbers of nodes for preferential attachment model ($m = 1$)

(f) Before plotting the relationship between the age of nodes and their expected degree, we should derive the equation for it. Let us introduce some notations for derivation:

- $k(i,t)$ is the average degree (expected degree) of the i^{th} node (the node added at the i^{th} time step) after time step t.

- m is the number of edges added at each time step.

To derive an expression for $k(i,t)$, let's write the one step forward difference equation

$$k(i,t+1) = k(i,t) + mf_{t+1} \tag{2.5}$$

where f_{t+1} is the possibility that the node added at the i^{th} time step forms an edge with the node added at the $(t+1)^{th}$ time step. The expression for f_{t+1} is given below

$$f_{t+1} = \frac{k(i,t)}{\sum_{V_i \in V} deg(V_i)} = \frac{k(i,t)}{2|E|} = \frac{k(i,t)}{2mt} \tag{2.6}$$

Substituting equation (2.6) in (2.5) and rearranging we have

$$k(i,t+1) - k(i,t) = \frac{1}{2} \times \frac{k(i,t)}{t} \tag{2.7}$$

Notice that for large enough value of t,

$$\lim_{t\to\infty} k(i,t+1) - k(i,t) = \lim_{t\to\infty} \frac{k(i,t+1) - k(i,t)}{t+1-t} = \frac{\partial k(i,t)}{\partial t} \tag{2.8}$$

Therefore, the equation (2.5) can be transformed as a partial differential equation shown below

$$\frac{\partial k(i,t)}{\partial t} = \frac{1}{2} \times \frac{k(i,t)}{t} \tag{2.9}$$

which gives the solution of $k(i,t)$ as

$$k(i,t) = m\left(\frac{t}{i}\right)^{\frac{1}{2}} \tag{2.10}$$

Using above equation, we can plot the relationship between the age of nodes and their expected degree with given m and t. The result is shown as Figure 11.

(g) First of all, the preferential attachment models with different m are always connected since there always exists a path between new added node with previous nodes. For the comparison of two networks of different sizes considering various m, the concrete result is organized as Table 3, where γ represents the negation of the slope of linear regression for the log-log scale plot. There are also some plots showing the results of question (d), (e), (f) for different m. Specifically, Figure 12 illustrates the degree distribution for randomly generated networks of different sizes with $m = 2$ while the Figure 15 shows the result for $m = 5$. Figure 13 and 16 represents the degree distribution of randomly picked node j for $m = 2$ and $m = 5$ respectively. The relationships between the age of a node and its expected degree for $m = 2$ and $m = 5$ can refer to Figure 14 and 17. According to the above visualizations, we can draw the following conclusion:

- The trend that the modularity of a larger preferential model is superior to smaller network is observed for $m = 2$ and $m = 5$ as well. Moreover, the modularity for a preferential attachment model decreases when m increases. This may result from the more edges are introduced during each iteration, the more it is likely to form edges out of communities.

- For $m = 2$ and $m = 5$, the slope of linear regression for the plot of degree distribution in log-log scale still gradually approaches the value of -3 with increasing size. However, the slope of linear regression decreases with increasing m. This might be caused by the intuition that the large m need great number of nodes to achieve the steady state.

- The result for question (f) inherits the features mentioned above, which can be interpreted properly combining the above explanation and the explanation mentioned in (e).

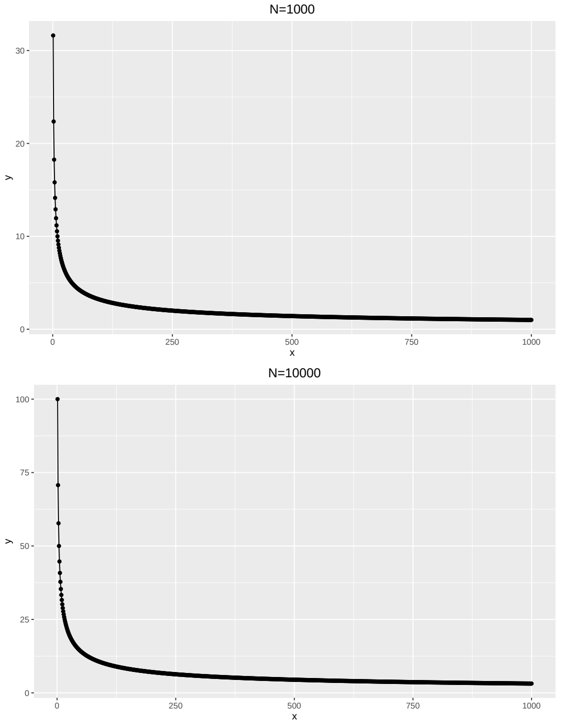

Figure 11: Expected degree of the i^{th} $(1 \leq i \leq 1000)$ node for different t with $m = 1$

m	$\lvert V\rvert = 1000$			$\lvert V\rvert = 10000$		
	Modularity	γ_{node}	$\gamma_{neighbor}$	Modularity	γ_{node}	$\gamma_{neighbor}$
1	0.9321	2.179	1.052	0.9786	2.494	1.457
2	0.5189	2.151	1.113	0.5323	2.233	1.234
5	0.2730	2.003	0.9959	0.2773	2.108	1.114

Table 3: Comparison result for network of size 1000 and 10000 with different m.

Question 2 continued on next page. . .

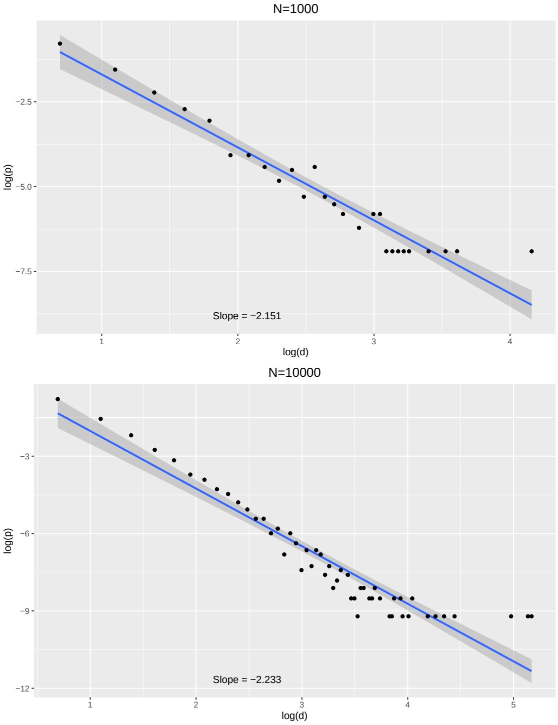

Figure 12: Degree distributions with different numbers of nodes for preferential attachment model ($m = 2$)

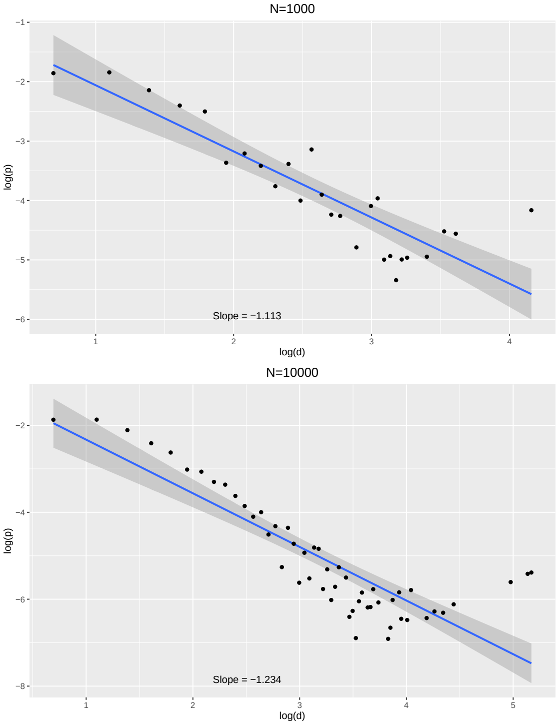

Figure 13: Degree distributions of V_j with different numbers of nodes for preferential attachment model ($m = 2$)

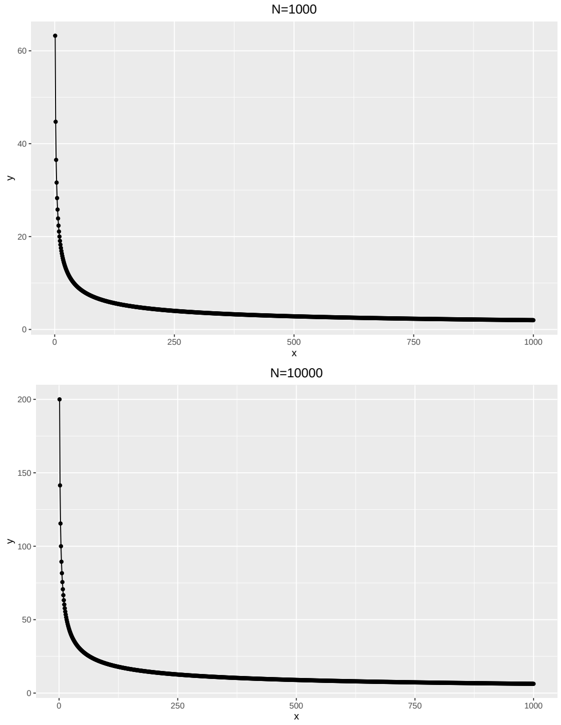

Figure 14: Expected degree of the i^{th} $(1 \leq i \leq 1000)$ node for different t with $m = 2$

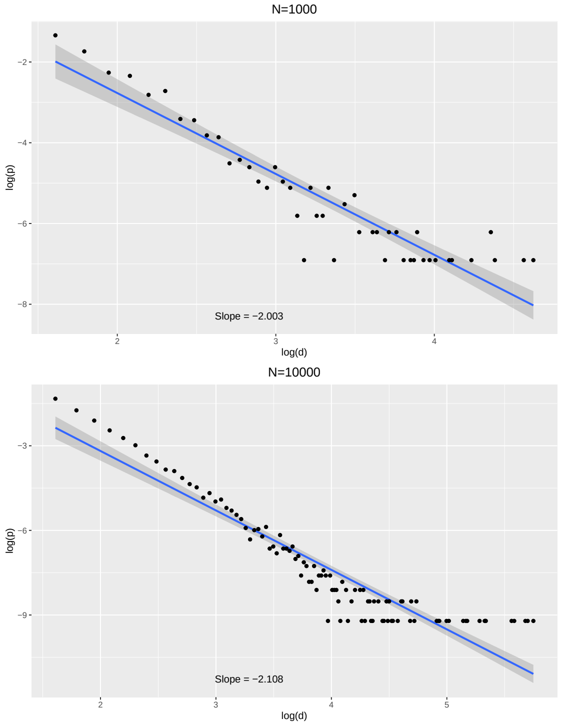

Figure 15: Degree distributions with different numbers of nodes for preferential attachment model $(m = 5)$

Question 2 continued on next page. . .

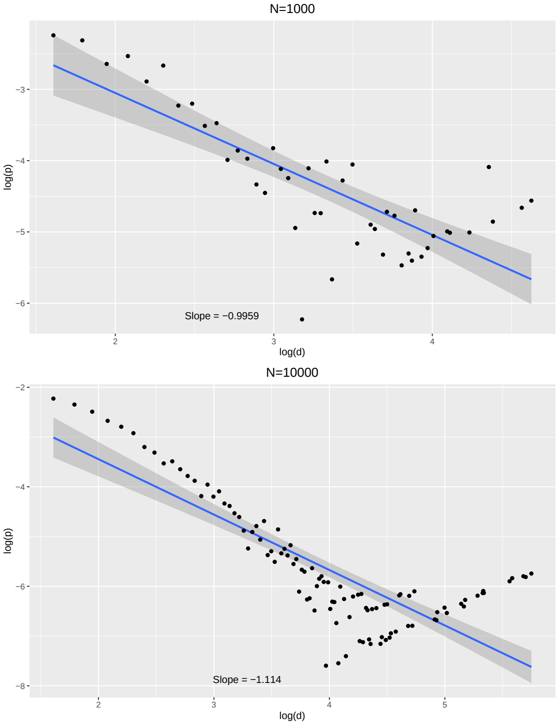

Figure 16: Degree distributions of V_j with different numbers of nodes for preferential attachment model $(m = 5)$

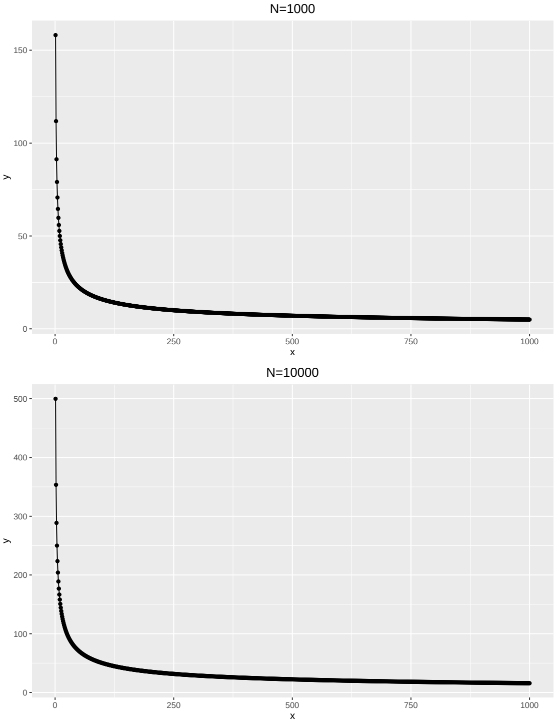

Figure 17: Expected degree of the i^{th} $(1 \leq i \leq 1000)$ node for different t with $m = 5$

- The relationship between the age of nodes and their expected degree with given m and t can be given as $E[deg(V_i)] = m\left(\frac{t}{i}\right)^{\frac{1}{2}}$. Hence, the trend of these plot is totally the same among different m. The only difference between them is the scale of y, namely the scale of expected degree.

(h) According to problem, we are going to use common $cluster\_fast\_greedy$ to separate the communities in original network. However, to correctly detect the communities in rebuilt network, we should use *walktrap.community* for possible multiple edges and self-looping situation in the reconstructed network. The plot of two graphs and corresponding communities in these networks are shown as Figure 18 and 19 by the help of colors of nodes. The modularity for two networks are 0.9331 and 0.7688 respectively. The reconstructed graph has smaller modularity. This may result from the fact that the reconstruction process may introduce problems like self-looping and multiple edges between two vertices, impairing the connectivity of the graph.

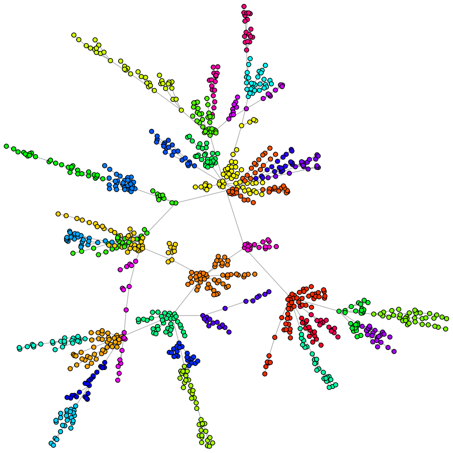

Figure 18: Network of size 1000 constructed by preferential attachment model ($m = 1$)

Question 3

(a) According to the documentation of igraph R package, we just need to set the correct parameter for function $sample\_pa\_age$ to generate a age penalizing preferential attachment model. As mentioned, for the expression of $P[i]$ below

$$P[i] \sim (ck_i^{\alpha} + a)(dl_i^{\beta} + b) \tag{3.1}$$

we are going to set $m = 1$, $\alpha = 1$, $\beta = -1$, and $a = c = d = 1$, $b = 0$. Then the degree distribution for a instance of such preferential attachment model is shown as Figure 20.γ, which is the negation of the slope of linear regression in log-log scale, is 2.637 according to the annotation of the figure.

(b) According to solution of above problems, it is easy to get the modularity of this network. The result is 0.9359.

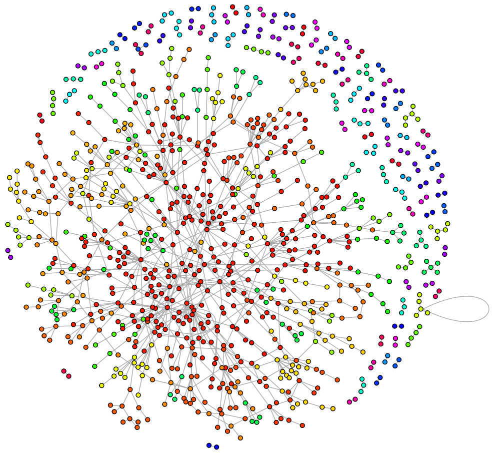

Figure 19: Network of size 1000 constructed using the degree sequence derived and stub connection process.

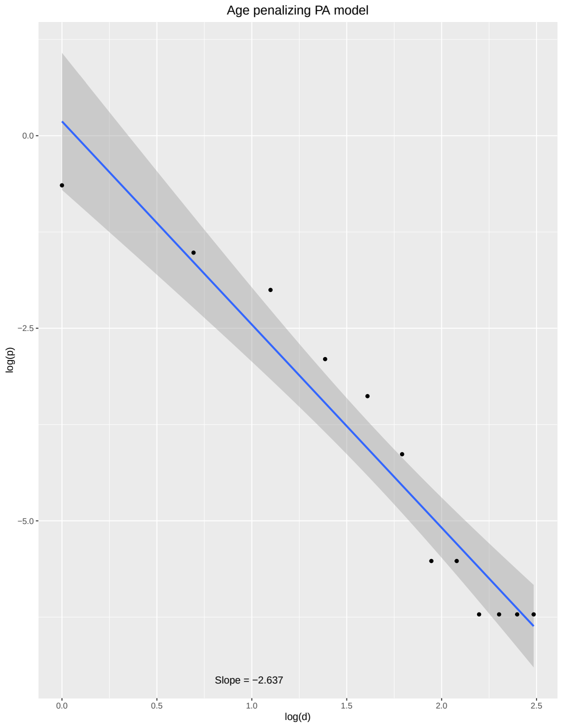

Figure 20: Degree distribution for network constructed by PA model with age penalization with $m = 1$.

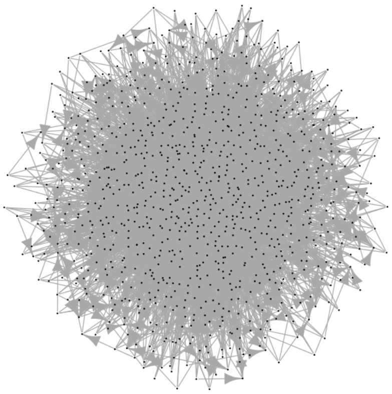

Figure 21: Instance of Erdos-Renyi network of size 1000 with $p = 0.01$

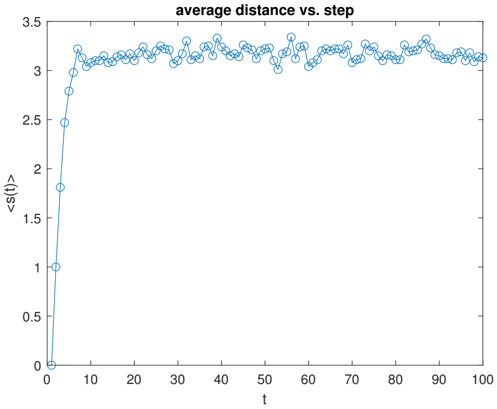

Figure 22: Average distance of the walker against the steps taken for Erdos-Renyi model of size 1000.

PART II: RANDOM WALK ON NETWORKS

Question 1

(a) Just like Part 1, we create a N=1000, p=0.01 Erdos-Renyi network. The network is shown as Figure 21.

(b) We set the step as and the number of loops as 100. The tendency of $\langle s(t) \rangle$ v.s. t and $\sigma^2(t)$ v.s. t is as shown as Figure 22 and Figure 23. In both figures, the curve fluctuated greatly at the beginning. Then as the step increases, it gradually stabilizes. For average distance, it converges on 3.1691, and variance converges on 0.4680.

(c) The degree distribution of the nodes reached at the end of the random walk and degree distribution of graph are shown as Figure 24 and Figure 25. The degree distribution of the nodes arriving at the end of the random walk is very similar to the degree distribution of the graph. They have almost the same mean, but

Question 1 continued on next page. . .

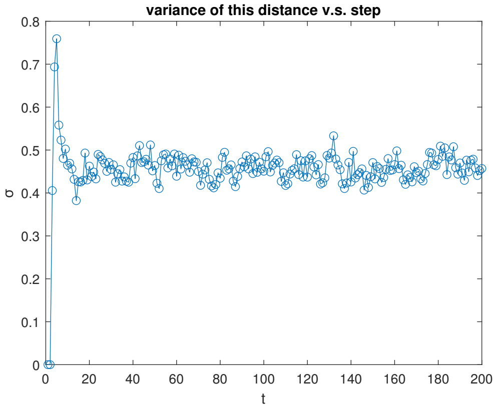

Figure 23: Variance of the distance against the steps taken for Erdos-Renyi model of size 1000.

the variance is different. Overall, both distributions like binomial distributions, as we have discussed above for the Erdos-Renyi network.

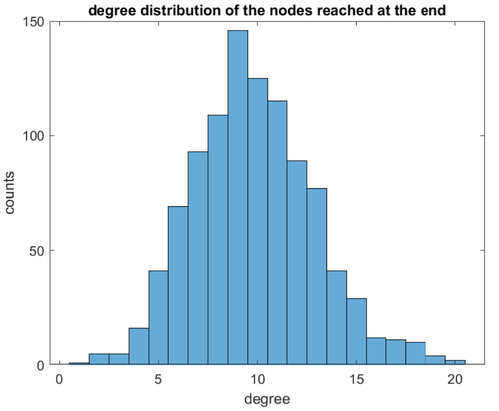

Figure 24: Degree distribution of the nodes reached at the end of the random walk at network of size 1000.

(d) Just like (b), we still set the step as and the number of loops as 100, but Node N = 10000. The tendency of $\langle s(t) \rangle$ v.s. t and $\sigma^2(t)$ v.s. t is as shown as Figure 26 and Figure 27. It quite similar to the Figure 22 and Figure 23. However, for average distance, it converges on 2.3557, and variance converges on 0.2445. Therefore, we can draw conclusions about the effect of size. Contrary to intuition, as more nodes are added, smaller graphs will take longer steps and larger variances to reach steady state. Conversely, larger graphs can reach steady state at shorter step sizes with smaller variances. Our explanation for this is because the larger graph contains more random walk distribution.

Question 2

(a) Once again, we generate an undirected preferential attachment network with 1000 nodes, where each node will be attached to m = 1 old node.

(b) For the trend of $\langle s(t) \rangle$ against t and $\sigma^2(t)$ against t, one can refer to Figure 29 and 30.We can see that the biggest difference between the Barabasi-Albert network and the Erdos-Renyi network is that the average

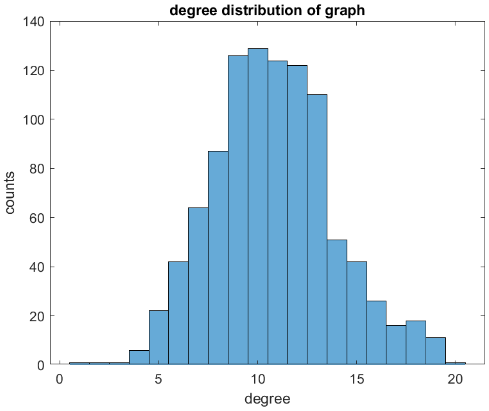

Figure 25: Degree distribution of the graph generated by Erdos-Renyi model with 1000 vertices.

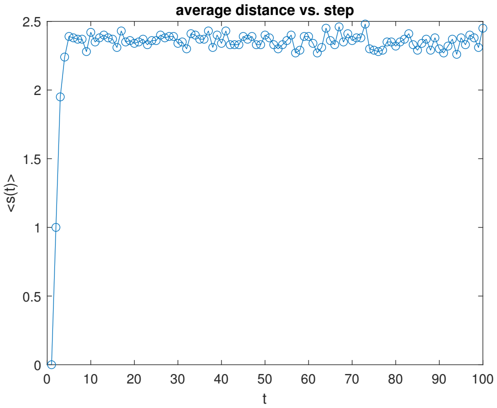

Figure 26: Average distance of the walker against the steps taken for Erdos-Renyi model of size 10000.

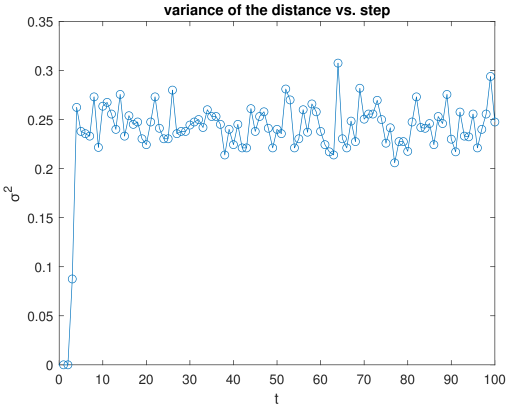

Figure 27: Variance of the distance against the steps taken for Erdos-Renyi model of size 10000.

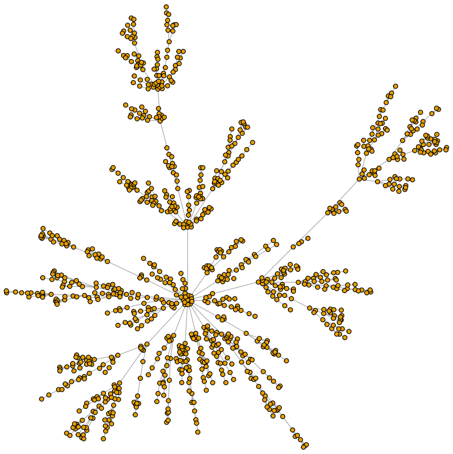

Figure 28: One instance of preferential attachment model with $n = 1000, m = 1$.

shortest path and variance do not seem to tend to be fixed. According to the principle, we can know that the Barabasi-Albert network follows the power law distribution, and the outgoing degree of each node is set to 1. In this test, even after 100 steps, the average short distance and variance are still increasing, and the fluctuation range is larger than the Erdos-Renyi network.

(c) Comparing the two histograms of Figure 31 and Figure 32, the conclusion is the same as 1(c): the degree distribution of the nodes arriving at the end of the random walk is highly dependent on the degree distribution of the original network. Both distributions follow the power law distribution.

(d) All relevant graphs are shown below. Specifically, the trend of average distance and its variance aganist the number of steps taken can refer to Figure 33 and 34 while the results about network of size 10000 can be found as Figure 35 and 36. This time we can find that for smaller graphs (100 nodes), the mean and standard deviation are relatively stable and fast. In addition, the small image also has a relatively small standard deviation, which also means that it is more stable than the other two large images.

This conclusion is different from the Erdos?Renyi network. This is because the degree distribution of the two networks is different. The larger the diameter and the larger the graph is, the less information is included in the degree distribution. This means that random walkers at larger graphs will use more steps to reach a steady state. Similarly, after reaching a relatively stable area, the deviation is also larger than the smaller chart.

Question 3

(a) We get the figure of probability distribution of the walker visits each node as Figure 37. However, this result is not straightforward to draw a conclusion. Hence, we will change the x-axis to the degree of the

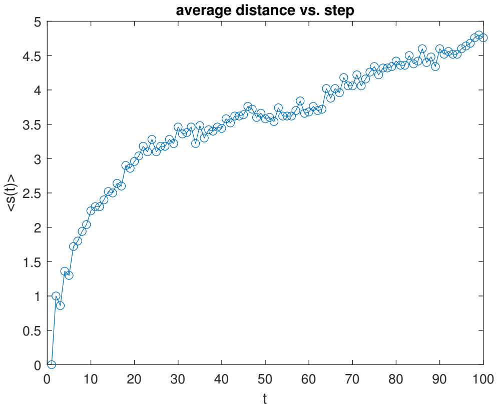

Figure 29: Average distance of the walker against the steps taken for PA model of size 1000.

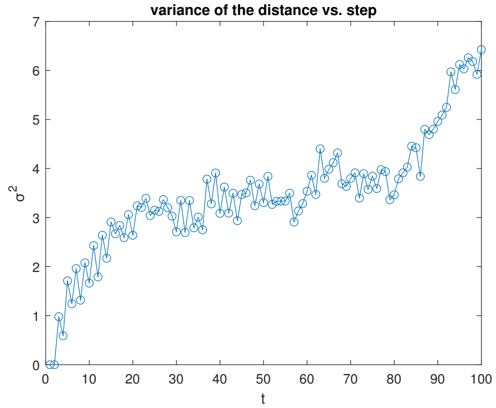

Figure 30: Variance of the distance against the steps taken for PA model of size 1000.

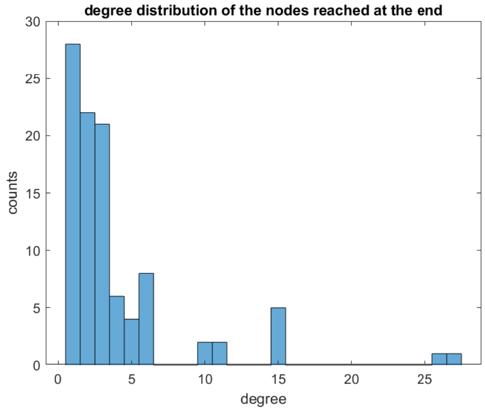

Figure 31: Degree distribution of the end node for network with fat-tailed degree distribution of size 1000.

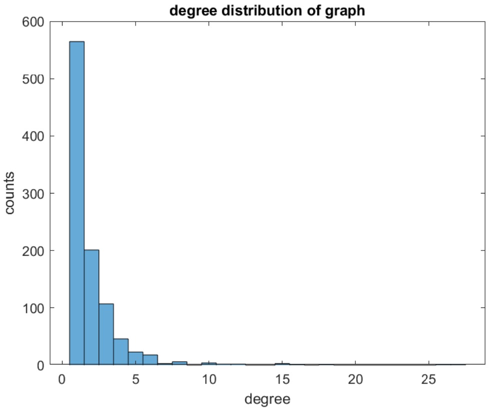

Figure 32: Degree distribution of the graph generated by PA model with 1000 vertices.

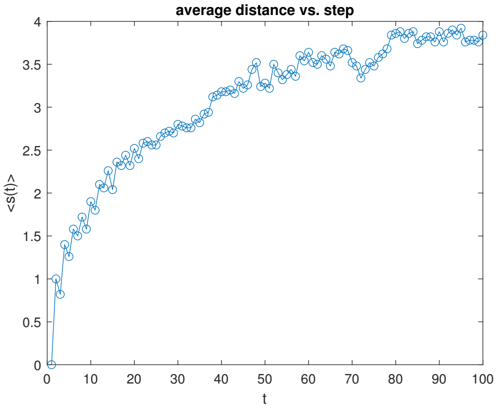

Figure 33: Average distance against steps for fat-tailed degree distribution network of size 100.

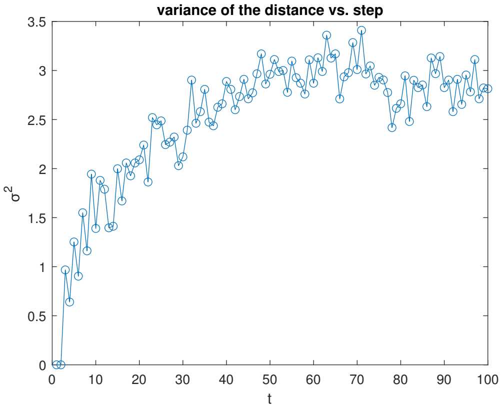

Figure 34: The variance of distance against steps for fat-tailed degree distribution network of size 100.

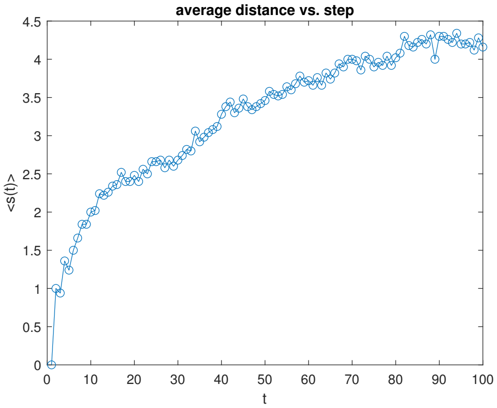

Figure 35: Average distance against steps for fat-tailed degree distribution network of size 10000.

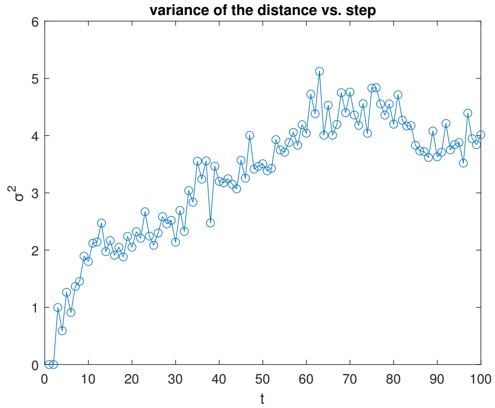

Figure 36: The variance of distance against steps for fat-tailed degree distribution network of size 10000.

nodes and get Figure 38. We can find that the probability is approximately linear with the degree of the node. To prove this, we measured the Pearson correlation coefficient of the graph and reported the value $R = 0.9019$. This satisfied our expectations, because nodes with higher degrees have more access methods than nodes with lower degrees. We also used a linear regression model and reported the fitting result as $y = 1.64 \times 10^{-4}x - 1.68 \times 10^{-3}$.

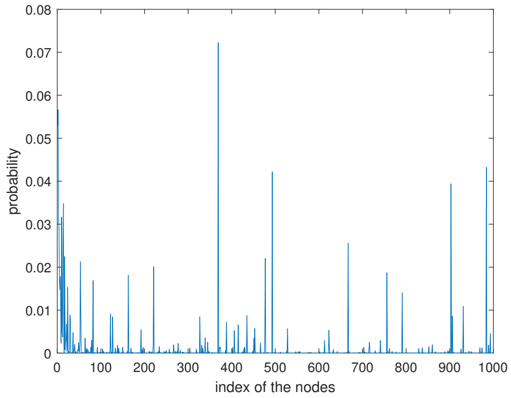

Figure 37: Probability that the walker visits each node.

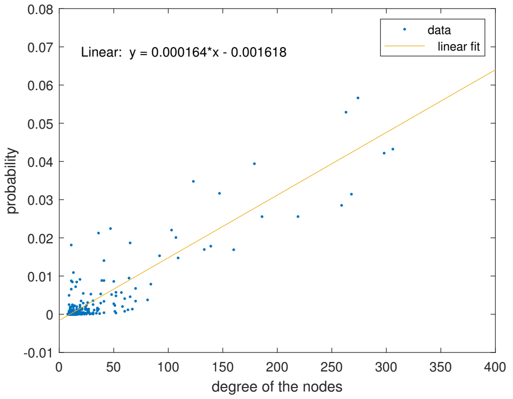

Figure 38: Linear fitting of visiting probability and the degree of nodes

(b) The same as 3(a), but this time we set a teleportation probability of ? $= 0.15$. We get the two similar figures as Figure 39 and 40. The Pearson correlation coefficient of the graph and the reported value is $R = 0.9281$, which is even more higher that (a). The result of linear regression is $y = 1.332 \times 10^{-4}x - 1.125 \times 10^{-3}$. A smaller slope means that low-degree nodes now have greater access opportunities. Therefore, compared with random walk without long-distance transmission, the probability of high access is not high. In summary, random walks (with and without uniform teleportation) all show strong linearity, while teleportation leads to a lower slope.

Question 4

(a) This time, by using the same function but changed the teleportation mechanism, we got R as 0.9132, which is between the result of (a) and (b) in Question 3. The plot illustrating results can refer to Figure 41

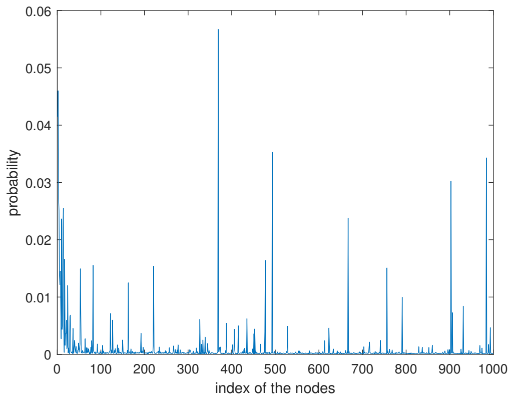

Figure 39: Probability that the walker visits each node with teleportation.

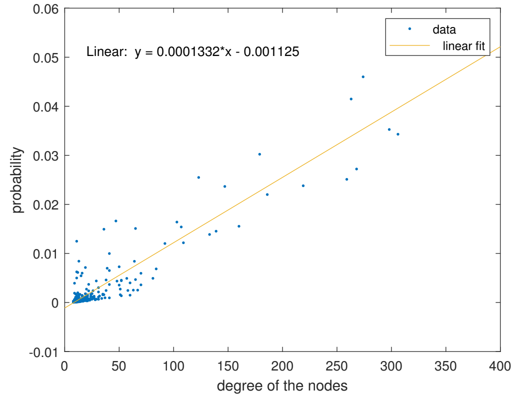

Figure 40: Linear fitting of visiting probability and the degree of nodes with teleportation.

and Figure 42

With uniform teleportation, the number of visits will move more towards nodes with a lower order, because lower-order nodes have more ways to reach them. Use the teleportation based on PageRank, since the probability of teleportation to a node is proportional to PageRank, the teleportation will not increase the access to low-level nodes.

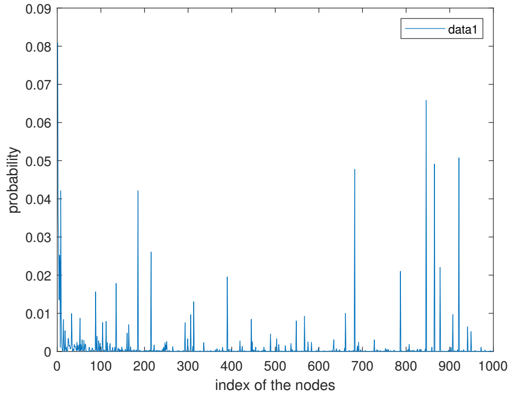

Figure 41: Probability that the walker visits each node for Personalized PageRank.

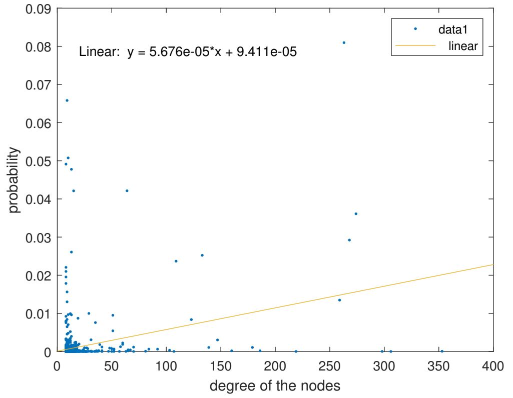

Figure 42: Linear fitting of visiting probability and the degree of nodes for Personalized PageRank.

(b) The two outliers in the upper left corner represent nodes with a median PageRank, which serve as the only invisible destinations. The Pearson correlation coefficient is lower than all previous experiments, with a magnitude of 0.7135287. This weaker linear relationship can be expected because nodes with lower degrees (two medians) are more likely to be visited suddenly.

It is obvious from the visualization, which is shown as Figure 43, 44 and 45 that for a small number of nodes, PageRank increases significantly. Two nodes with a median PageRank (500 marks on the x-axis) show that PageRank has increased dramatically because they are the only transfer destination. The other peaks may be nodes closer to the two median PageRank nodes.

(c) In this part, we combine 4(a) and 4(b) and change the invisible transmission probability to cover the PageRanks value and the influence of trusted web pages.

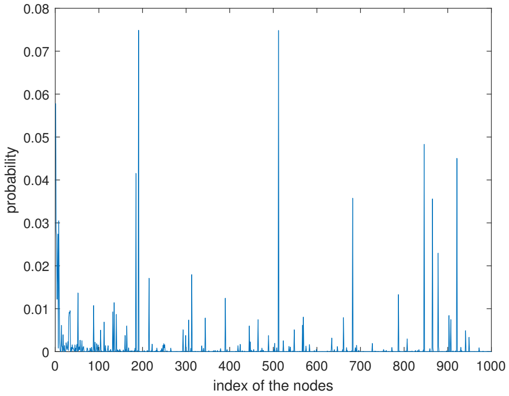

Figure 43: Probability that the walker visits each node with median PageRank.

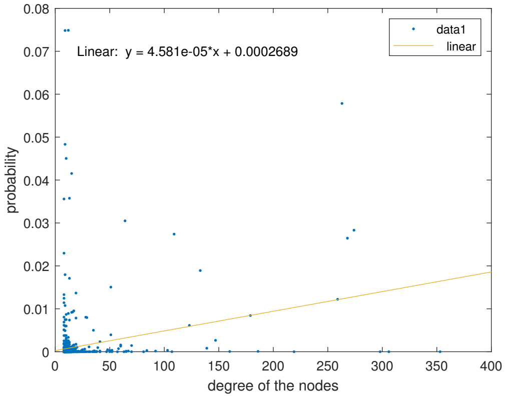

Figure 44: Linear fitting of visiting probability and the degree of nodes with median PageRank.

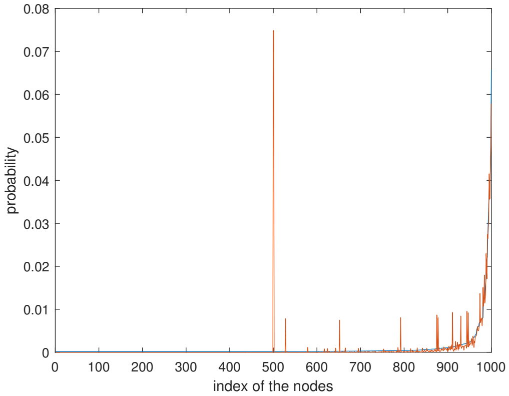

Figure 45: PageRank with probabilities with median PageRanks.

Therefore, we set the median PageRanks to $\frac{\theta}{2}$, and adjust the probability of $\frac{\theta}{2}$ nodes, while the remaining nodes follow the PageRanks normal distribution with a weight of $1 - \theta$. In this way, the probability vector will be changed to

$$P = r \times (1 - \theta) + \beta \times \frac{\theta}{2} \tag{4.1}$$

where r stands for the PageRank, for nodes with median PageRanks.