Network Project
A Growing Network Model

CID: 00837636

Abstract: This project investigated the degree distribution of networks generated by Price/Barabasi and Albert (BA) model and its two variants. The default preferential attachment, purely random attachment and mixed attachment with probability, q, of preferential attachment are used in these 3 models. The degree distribution from these three models were simulated numerically and compared to theoretical values. The preferential attachment leads to a scale-free network with fat-tailed distribution $p(k) \sim k^{-3}$. Purely random attachment networks are subcritical or not scale free. Mixed attachment method gives scale-free networks in large size but small sized networks would vary. χ^2 test and coefficient of correlation we used to check if the simulation obeys theoretical model estimation for infinite sized networks and they booth give nearly 1.

Word count: 2479 words in report (excluding front page, figure captions, table captions, acknowledgement and bibliography).

0 Introduction

There are many different models for generating graphs. The Price model is the first known scale free network generating model to simulate citation networks. Barabasi and Albert (BA) model could be seen as an undirected version of Price model. The method used to evolve the network can influence the resultant network structure greatly. This project investigated the preferential attachment (default BA model), pure random attachment (a simplified version of BA model) and a mixed attachment method with probability q of preferential attachment. The preferential attachment is the default method used in the BA model and it would lead to a scale free network. Pure random attachment is a method would give exponential degree distribution in theory. The mixed method will just help us understand the model better.

0.1 Definition

The BA model is an algorithm for generating undirected random networks using preferential attachment. It could simulate internet, citation networks and some social networks very well. It uses preferential attachment method to connect new node to existing nodes with a probability proportional to the degree of each node. The attachment method could be change to purely randomly, independent of the degree of existing nodes. And a mix of the 2 can also be used. A scale free network is a network with power law degree distribution. The probability for large degree decays slower than exponential distribution and a long tail appears on the probability distribution graph. Power law distribution is also called fat tail distribution for its long and fat tail on graph.

1 Phase 1: Pure Preferential Attachment

1.1 Implementation

1.1.1 Numerical Implementation

The implementation of preferential attachment would be easy and effective if the graph abstract data structure is well designed. In practise, there are 3 popular implementation of graph, the edge list, adjacency list and adjacency matrix[1]. The adjacency list representation is generally the best implementation considering the time complexity of operations and space complexity for saving data[1]. This representation also fits this programme best. We won't implement a graph data structure by ourself since there are already many libraries exist for this work. In this project the python module networkx is used and it implement graph as adjacency maps according to its documentation. This is a better implementation in python since it optimised the get_edge method to $O(1)$[2] and avoid the extra space needed by python list (python list will allocate more memory than needed for future potential extend or append operation) .

1.1.2 Algorithm

- Initialise: Set up the initial networks which we will evolve from. The initial size of the graph can be set manually in implementation, but it was set to m for simple. The initial graph was a complete graph to maximise the number of edges.
- Drive: Increment time and size of the graph by 1, set the degree of the newly added node to m, which means we need to connect it to m existing nodes.

- Evolve: Connect the new node to m existing nodes by choosing the m nodes randomly with a probability proportional to the degree of the existing nodes.
- Repeat driving and evolving the graph until size reach the required system size, 10^5 in this project.
  The evolve step is hard to implement because we need to select m nodes from existing node by probability proportional to their degree. There are two obvious method to do it: build a list with each node occurs k times and select m different nodes equally randomly or select one edge from the edge list and then choose one node at the end of this edge proportional to their degree. The first method is obviously right and easy to implement method but needs $O(n^2)$ space to maintain the nodes list. The second method might not require extra space depend on how networkx generate edge list and the random select part should be faster despite 2 random number generated because the edge list is only half the size of the nodes list in first method, but I can only describe it intuitively and check on simple cases without strict mathematical proof.

1.1.3 Type of Graph

The graph generated is undirected, unweighted graph with no parallel edges loop from and to the same node allowed. The algorithms also guaranteed that no isolated node exist and so it is also a connected graph. It should also be scale free.

1.1.4 Working Code

Some simple and small size networks were used as test cases. The code works with no unhandled errors. It should work as expected since the logic of the code follows the algorithm strictly and the simple case seems worked as expected (by watching the network evolve). The evolve method used in the code is the first method mentioned above. I cannot prove the second method mathematically and the benefit from saving space seems not exist since every time the Graph.edges method is called, there will be a significant increase of memory usage. A test for complex case would run for a long time and it is hard to notice if it worked as expected by human brain. So, we just assume the model worked as expected from testing simple case.

1.1.5 Parameters

The parameters for the generating function are (N, m, n, seed) where N is the systems size the simulation will stop and m are the edges for each new node. The n is the initial vertices number for the graph and seed is just a seed for random number generator to make the stochastic process reproducible. The number of initial network size affect the average degree $< k >$ of the system.

$$< k > = \frac{1}{n}\sum_i k_i = \frac{2\,Edges}{n} = \frac{2\,n\,(n-1)}{n} = 2(n-1)$$

With constraint that n greater or equals to m we found that this does not affect the initial conditional too much. So, we just set $n = m$ in the simulation.

1.2 Preferential Attachment Degree Distribution Theory

1.2.1 Theoretical Derivation

The evolution of this network is described by master equation. Let the probability of an existing node been attached to a new node, it is proportional to k_i, the degree of this node. E(t) been the number of edges at t, we have:

$$\Pi(k,t) = \frac{k_i}{\sum_i k_i} = \frac{k_i}{2E(t)}$$

The master equation can be written as[3]:

$$\begin{aligned} n(k,t+1) = \quad & n(k,t) && \text{(Start with this number.)} \\ & +m\Pi(k-1,t)n(k-1,t) && \text{(Effect of edges added to vertices with degree } (k-1)) \\ & -m\Pi(k,t)n(k,t) && \text{(Effect of edges added to vertices with degree } k) \\ & +\delta_{k,m} && \text{(The one new node has } m \text{ edges)} \end{aligned}$$

From this master equation, we can define the probability of a node in degree k at time t as

$$p(k,t) = \frac{n(k,t)}{N(k,t)} = \frac{n(k,t)}{\sum_k n(k,t)}$$

$$lim_{t\to\infty} p(k,t) = p_\infty(k)$$

All the equations above together lead to a differential equation

$$p_\infty(k) \approx -\frac{1}{2}\frac{\partial\left(kp_\infty(k)\right)}{\partial k}$$

We can solve it by trial solution and get $p(k) \sim k^{-3}$[4]. Note that for k <m, the probability is 0. This result shows that the distribution is fat tailed for the BA model. The normalisation constraint for probability lead to

$$\int_m^\infty p_\infty(k)dk = 1$$

And thus,

$$p_\infty(k) = 2m^2k^{-3}$$

1.2.2 Theoretical Checks

Without so many approximations, we can get a more exact solution. From the master equations and fractions of nodes with k and probability of connecting to a new node. We can get an equation without many approximation

$$p_\infty(k) = \frac{1}{2}\{(k-1)p_\infty(k-1) - kp_\infty(k)\} + \delta_{k,m}$$

For k<m, the probability would be just 0 as every edge added with degree of m.

For k=m, the equation becomes $2p_\infty(m) = (m-1)0 - mp_\infty(m) + 1$, that is

$$p_\infty(m) = \frac{1}{m+2}$$

For k>m, $\delta_{k,m} = 0$, and rearrange the equation, we get

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2}$$

With the help of gamma function, we can find the form of $p_\infty(k)$

$$p_\infty(k) = Z\frac{\Gamma(k)}{\Gamma(k+3)}$$

Where the Z is normalisation factor. The probability must satisfy

$$\sum_{k=1}^{k\to\infty} p_\infty(k) = \sum_{k=m}^{k\to\infty} p_\infty(k) = \frac{2}{m+2} + \sum_{m+1}^{\infty} Z\frac{1}{k(k+1)(k+2)} = 1$$

$$\sum_{m+1}^{\infty} Z\frac{1}{k(k+1)(k+2)} = Z\sum_{m+1}^{\infty}\left[\frac{1}{2k} - \frac{1}{k+1} + \frac{1}{2(k+2)}\right] = \frac{Z}{2(m+1)} - \frac{Z}{2(m+2)}$$

Thus, we can derive

$$p_\infty(k) = \begin{cases} 0 & for\ k < m \\ \dfrac{2m(m+1)}{k(k+1)(k+2)} & for\ k \geq m \end{cases}$$

This exact solution with few approximations is consistent to the theoretical analysis.

1.3 Preferential Attachment Degree Distribution Numerical Results

1.3.1 Fat-Tail

From the result we get from part 1.2 we found that the degree distribution is exponential and thus, a fat-tailed distribution would occur when we analyse the simulation data. Fat-tailed distribution is not exponentially bounded and thus there are always nodes with large degree appear in the data. Due to the low probability, there are many zeros in the graph of degree distribution vs degree. These properties makes it harder to do statistical analysis on fat-tailed distribution even in log-log plot[3]. There are three ways to deal with fat-tailed distribution: log binning, using cumulative distribution function, and a Zipf plot of rank vs degree. The log binning is used in this project.

1.3.2 Numerical Results

The simulation gives similar graph with different m number, but the cut off degree is increasing (the right most point moves to right when m increase) as well as the size of the

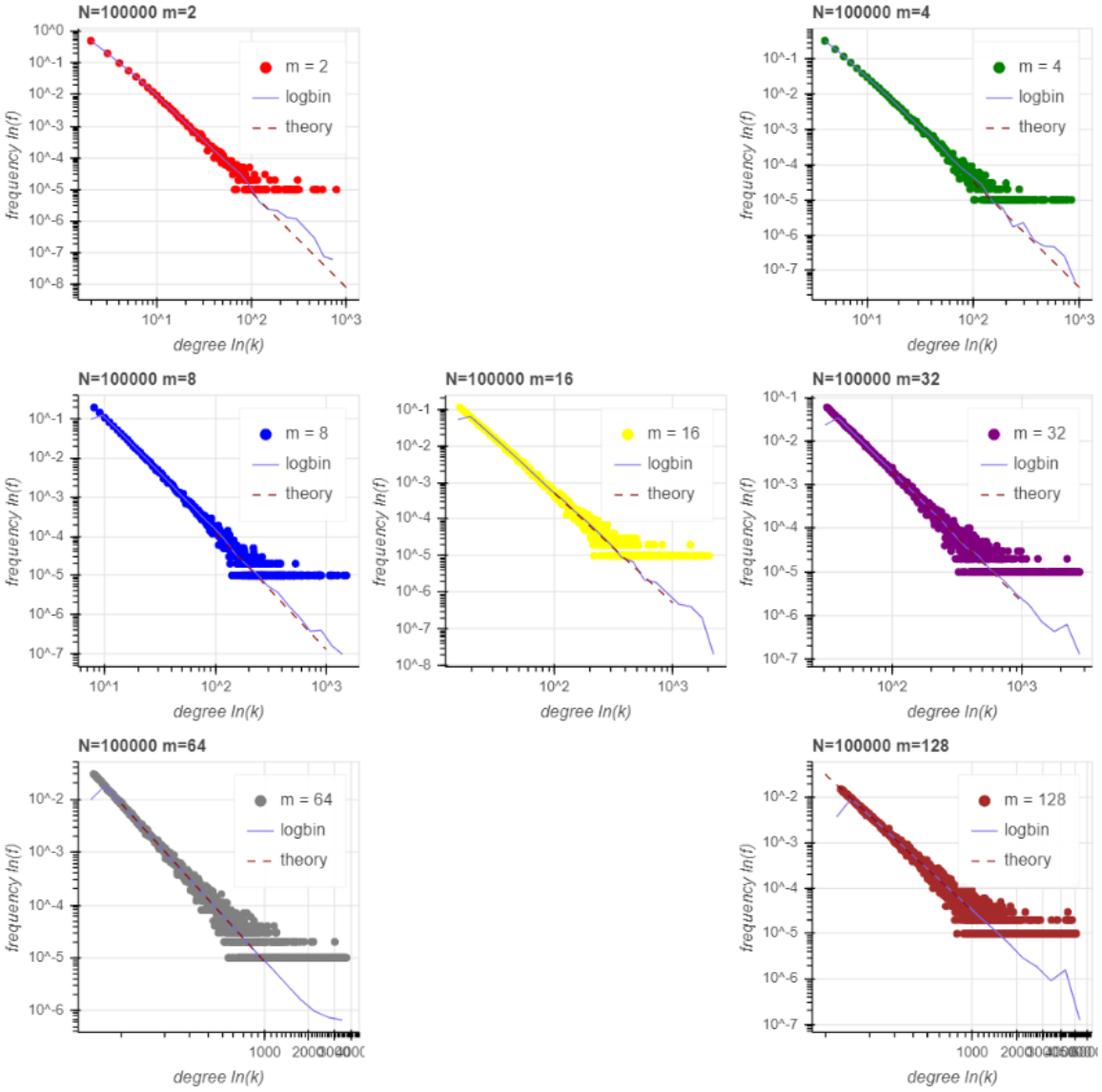

Figure1: the preferential attachment method with N=100000 and varied m

"fat tail" on the graph (shown in figure 1). This is because the probability of having large k is proportional to the square of m while the smallest probability can be represented in the graph is constrained to 1/N. This is kind of statistical noise and data binning method can flatten the noise and give a straight line like curve in the graph. This line generated by log binning can give a better intuitive feeling of the distribution. From figure 1 we can see that visually the log binning lines are good dovetail well to the theoretical value. We can employ χ^2 test for the observed data (fitted data of log binning) and the theoretical value from our model. The p value is always nearly 1 and for most cases it is 1, an odd result. This might be caused by insufficient accuracy when doing the very small floating point number operation or the χ^2 test do not fit this test case because the frequency sequence gives probability and the number varies greatly. Another test use correlation coefficient to determine whether

they fit well. The R^2 is always greater than 0.99. The statistical test showed that the log binning method is very good for fitting the data.

1.4 Preferential Attachment Largest Degree and Data Collapse

1.4.1 Largest Degree Theory

There is no content in this course for the largest degree theory. The only thing I can find related to this topic is the cut off degree in an introduction to scale free networks online[5]. I will assume the cut off degree is equivalent to largest degree. The mathematical definition of the cut off degree is that

$$\sum_{k_{cutoff}}^{\infty} p_\infty(k) \approx \int_{k_{cutoff}}^{N-1} p_\infty(k) = \frac{1}{N}$$

Where N is the size of the system. For the BA model, we use $p_\infty(k) = 2m^2k^{-3}$ and the largest expected degree, so $k_{cutoff} \sim m\sqrt{N}$

1.4.2 Numerical Results for Largest Degree

From figure 1 we can see that when m increase, the tail been noisier. This noise is caused by the size of the system is finite and it could be reduced by increase system size or reduce the m value. But 10^5 is already the greatest scale I can get on my laptop, 10^6 sized system would cause memory error for $m > 32$. So, the finite size effect would be investigated using 16 as m value and system size span from 10^2 to 10^6. The results are shown in figure 2. As the system size increase, the effect of statistical noise decrease, in other words, the tail been thinner comparatively. The hump of log binning curve at the tail is less obvious. This is because cut off grows slower than the system size. The degree distribution is expected to be scale-free and this implies a data collapse similar to self-organized complex system. Thus, we can propose that

$$p(k, L) \sim k^\tau \, \mathcal{G}(\frac{k}{k_{cutoff}})$$

Where $\mathcal{G}\left(\frac{k}{k_{cutoff}}\right)$ is a scaling function. So, we can deduce that

$$\frac{p(k, L)}{k^\tau} = \mathcal{G}(\frac{k}{k_{cutoff}})$$

And this should be a straight in log-log plot.

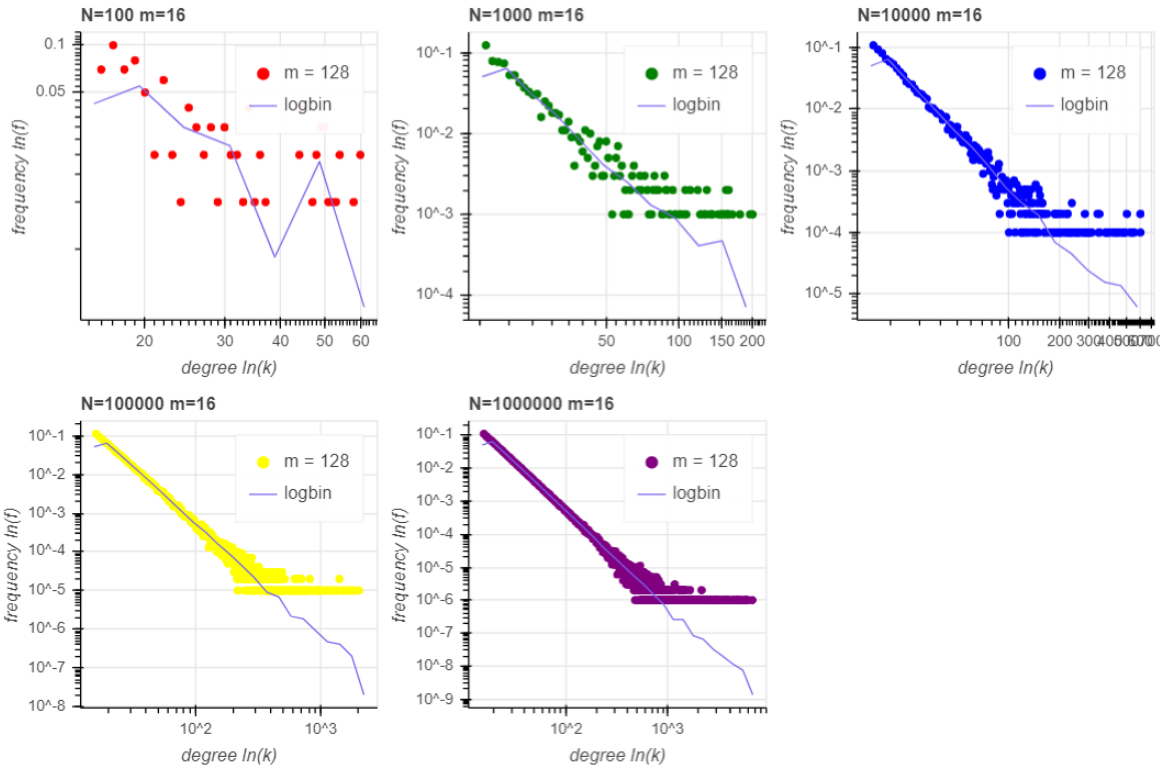

Figure 2: The system with m=16 and varied size.

1.4.3 Data Collapse

The data collapse can be seen in figure 3. They do tend to collapse to the scaling function as they fall into a straight line. They did not fall into the same straight line and there are shifts from each system size because the cut off degree we used is not exact. The square root of the system size cut off degree lacks is just the approximate relationship without a constant. But the shape of the lines implies that the data collapse as expected.

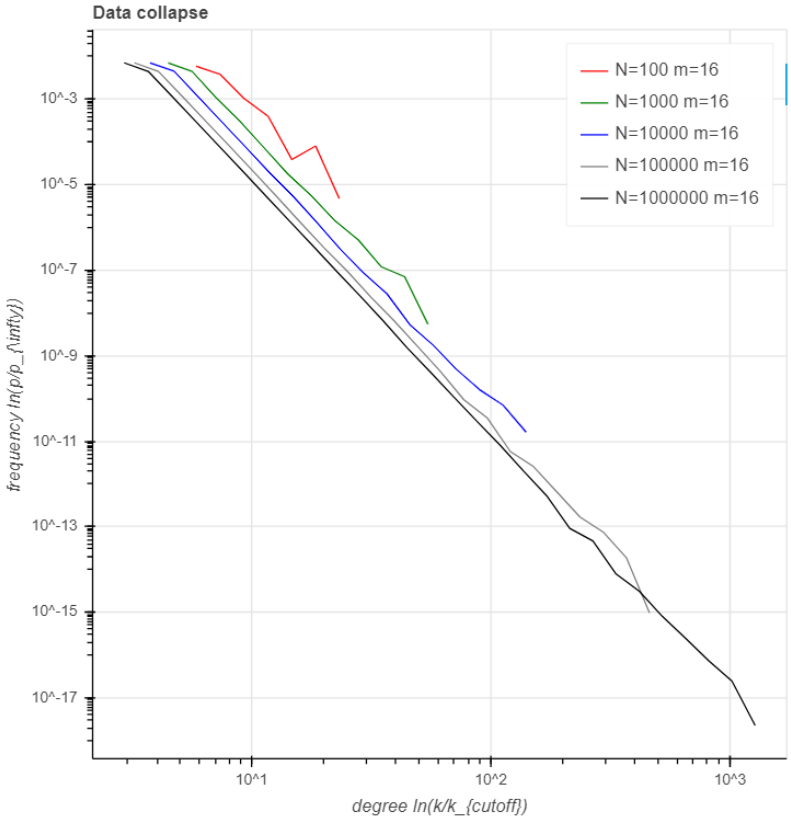

Figure 3: Data collapse of networks with different size and same m.

2 Phase 2: Pure Random Attachment

2.1 Random Attachment Theoretical Derivations

2.1.1 Degree Distribution Theory

Similar to the preferential attachment, but the probability of an existing node will connect to a new node is purely random and independent of its degree. This probability depends only on time (which determines the size of the system). The master equation becomes

$$N(t+1)p(k,t+1) - N(t)p(k,t) = \frac{m}{N(t)}n(k-1,t) - \frac{m}{N(t)}n(k,t) + \delta_{k,m}$$

With the definition of $p(k,t) = \frac{n(k,t)}{N(t)}$, this simplified to

$$(m+1)p_\infty(k) = mp_\infty(k-1) + \delta_{k,m}$$

With k very large, the finite k term can be ignored

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{m}{m+1}$$

Thus, we can get

$$p_\infty(k) = \left(\frac{m}{m+1}\right)^{k-m} p_\infty(m)$$

And normalisation for probability gives

$$p_\infty(k) = \begin{cases} 0 & for\ k < m \\ \frac{1}{m+1}\left(\frac{m}{m+1}\right)^{k-m} & for\ k \geq m \end{cases}$$

This is exponential and thus the network is not scale free and no fat-tail distribution according to the theory.

2.1.2 Largest Degree Theory

The pure random attachment method did not give a scale free network. The degree distribution is not in power, but in exponent. This implies

$$k_{cutoff} \sim \ln(N)$$

We can use the probability above and geometric series summation to $k = \infty$ and get the exact theoretical prediction:

$$k_{cutoff} = m - \frac{\ln(N)}{\ln(m) - \ln(m+1)}$$

2.2 Random Attachment Numerical Results

2.2.1 Degree Distribution Numerical Result

The degree distribution shown in figure 4 is obviously exponential and there is a clearer cut off degree exist on this graph especially on the log binning curve (Figure 4). The theoretical prediction looks quite fit well with the simulation. The p-value and coefficient of correlation are nearly 1 and 0.99 respectively.

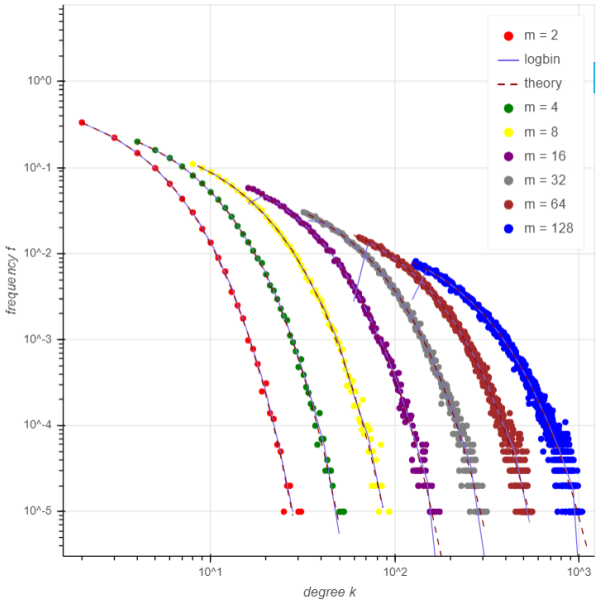

Figure 4: The degree distribution of random attachment

2.2.2 Largest Degree Numerical Results

The degree distribution of observed data and theoretical data are consistent. The cut off degree from theory should also give a good approximation to the observed largest degree. As we have an exact solution to the largest expected degree, we can make directly comparisons. The theoretical predictions lie near the largest observed degree when m is small but much greater than the observed largest degree when m is increasing. The small m cases imply the theory fit well with observations and the large m cases give a worse prediction can be just statistical noise as the model treat N as infinity and $N \gg m$ should be applied when using the model.

3 Phase 3: Mixed Preferential and Random Attachment

3.1 Theoretical Derivations

For mixed preferential and random attachment, the new node would be connected to an exist node each time by stochastic method: by probability q, it would connected to an existing node by preferential attachment and by chance (1-q), the existing node is just selected randomly. This transfers the new probability of adding new nodes to

$$\Pi((k) = q\Pi_{BA_{pa}}(k) + (1-q)\Pi_{random}(k)$$

And the master equation is just a mix of the 2 master equations from the previous 2 model

$$n(k,t+1) = n(k,t) + qm\Pi(k-1,t)n(k-1,t) + \frac{(1-q)mn(k-1,t)}{N(t)} - qm\Pi(k,t)n(k,t) - (1-q)m\frac{1}{N(t)}n(k,t) + \delta_{k,m}$$

And this can be simplified as did in the previous two model

$$2[N(t+1)p(k,t+1) - N(t)p(k,t)] = q(k-1)p(k-1,t) + 2(1-q)mp(k-1,t) - qkp(k,t) - 2(1-q)mp(k,t) + 2\delta_{k,m}$$

Where the q, k, m are constants, so the recurring equation is

$$\frac{p(k)}{p(k-1)} = \frac{k + \frac{2m - 2mq - q}{q}}{k + \frac{2m - 2mq + 2}{q}}$$

This is in the same form as the recurring equation for default BA model. So, the solution of $p_\infty(k)$ follows the power of k and the mixed network should also scale free unless q = 0. The exponent would be affected by q and thus the shape of the "fat tail".

3.2 Numerical results

As shown in figure 5, the degree distribution collapse to exponential distribution when q=0

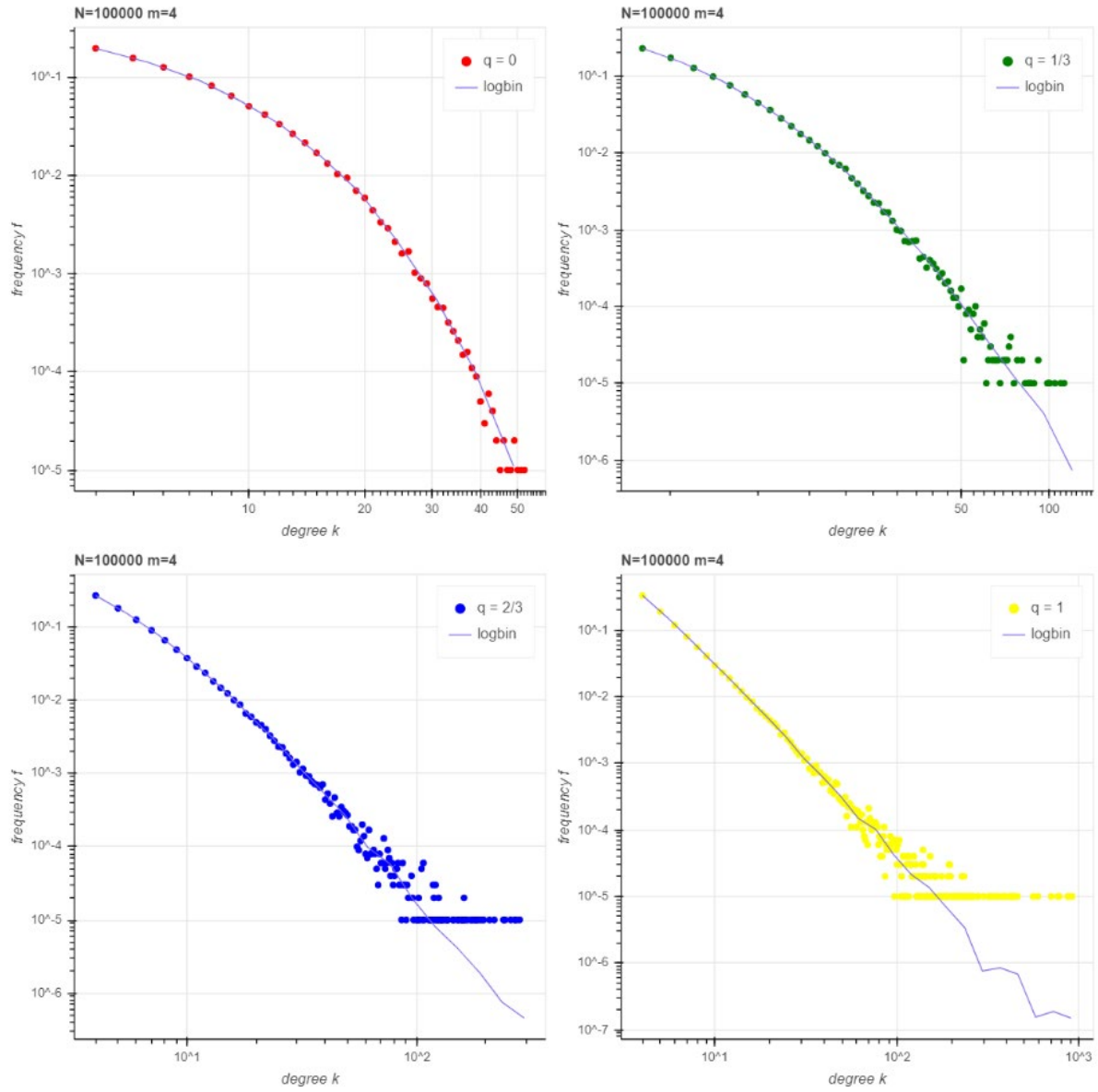

Figure 5: The degree distribution for mixed preferential attachment with different q

and is pure power law distribution when q = 1. Fat tail distribution occurred when q not 0

and the q value would affect the shape of the distribution. The bigger the q is, the "fatter" the tail.

3.3 Discussion of Results

We have already known that for the limiting case (q = 0 or q =1) the degree distribution would reduce to pure random attachment or preferential attachment, thus, exponential and fat tailed distribution. For q values in the middle, it is clearly that the fat tailed distribution occurred while with lower q value the fat tailed distribution is no quite obvious. This effect implies that the probability for large k decrease slower with greater q value. We can deduce that with higher q value, the smaller exponent the fat tail distribution $p(k) \sim k^{-\gamma}$ has.

4 Further Discussions:

This project investigated the BA model and its variants for degree distribution. The initial networks are complete graphs or empty graphs for simplicity. So, the initial conditions are not carefully considered and controlled when generating the network. The m number is fixed during every simulation and this differs from real world networks. A further investigation involving determining the m number randomly at each stage a new node is introduced could be considering. The simulations are all on evolving a small networks (of size comparative to m) to large network, when the initial size is significant to the total number of new nodes we are going to add, the simulation might give different graph.

5 Conclusions

The preferential attachment method used to evolve a network from equal degree on every node. Only a small difference introduced initially is amplified over the whole process and evolve to power distribution. The m value affects the shape of distribution very much. The accumulated advantage plays a central role in this process. In the random attachment case, the degree distribution became exponential and a cut off degree which constrains the expected largest degree is rarely seen in finite sized simulations. Mixed attachment is defined central on the q value and the different shapes of degree distribution could give an intuitive feeling of how the accumulated advantage affects the evolving of a network.

6 References

[1] I. To, *Introduction to Algorithms*, no. June. 2016.

[2] metode penelitian Nursalam, 2016, "Algorithms and data structures in python," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[3] T. S. Evans, "NetworkNotesStudentVersion," vol. 2021, no. February, 2021.

[4] T. Evans, "Complexity & Networks Part II : Networks," no. February, 2021.

[5] N. Science, "The World Wide Web is a network whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to 'surf' with a click from one web document to the other. With an estimated size of over one trillion documents," pp. 1–64, 2018, [Online]. Available: http://networksciencebook.com/chapter/4#hubs.