Network Project

John Norrie

March 26, 2017

Abstract

Networks are an important modelling tool in many fields of science and social science. From modelling the internet to the spread of epidemics, networks provide an intuitive and ... approach to concepts so popularity and contact.

In this paper will be presented theoretical and numerical analysis of three separate models. The first is a Barabasi Albert Model. This model uses preferential attachment based on the degree size. The second is concerned with random attachment, and the third uses the concept of random walks on the graph to determine which nodes are used in attachment.

There are two key areas which will be analysed in detail; the first of which is the degree probability distribution, and the second is the highest recorded degree. In both of these areas, theoretical derivations will be provided, and supported or criticised using numerical data.

1 Implementation of the BA Model

The Barabasi Albert model(BA) is alogirthm used to create freescale networks which rely on preferential attachment when adding new nodes. This means that nodes with a high degree are more likely to be attached to be new nodes. The algorithm works as follows:

1. Create an initial graph, $\mathcal{G}_0$ at t_0
2. Increment $t \to t+1$
3. Add one new vertex
4. Add m new edges from the new vertex to vertices in the network, each with a probability of attachment $\Pi \propto k$ where k is the degree of each vertex.
5. Repeat from step 2 until the network contains N vertices.

The biggest factors when implementing this model is speed. The size of the networks we wish to analyse mean that a fast program is essential. The programme I developed creates a list of numbers, each couplet defining an edge. At the k^{th} time step, a number is chosen at random (uniformly distributed) from this sequnce (x). then the number k and x are appended to the end of the list. This is carried out for N time steps, after which the frequency of numbers are counted, and the outcome is the degree of each node. An example is shown below.

1. -$G = [0, 1, 0, 2, 1, 2]$ is our starting graph, $N = 3$, $m = 3$
2. Add 2 new edges
3. 0 and 2 chossen randomly
4. $\Rightarrow G = [0, 1, 0, 2, 1, 2, 3, 0, 3, 2]$
5. Ending the iteration the counts are sorted, giving degree distribution:
   $deg(G) = [3, 2, 3, 1]$

This forces prereferntial attachement on the system, as if number $i < k$ appears more times in the sequence(e.i has more egdes attached), it is more likely to be chosen.

1.1 Initial Graph

There are a few points of ambiguity in this model. The first of which is with respect to $\mathcal{G}_0$. There is no explicit guidance on how to choose $\mathcal{G}_0$, however the choice of starting graph does have an affect. When deriving the master equation for the network, we will use the approximation that $E(t) = mN(t)$ (where $E(t)$ and $N(t)$ are the number of edges and nodes in the graph at time t respectively.) for larger t. However we can make this approximation exact by choosing an $\mathcal{G}_0$ such that $E(0) = mN(0)$.
There are many graphs that satisfy this relation, however I would like the starting graph to be as small as possible. Doing this reduces the influence at has on the actual model. Therefore I used a complete graph.
The number of edges in a complete graph is,

$$E = \sum_{n=1}^{N} n - 1 = \frac{N(N-1)}{2} \tag{1}$$

Therefore, applying the conditon,

$$E(0) = mN(0) \Rightarrow \frac{N(0)(N(0)-1)}{2} = mN(0) \tag{2}$$

$$\Rightarrow N(0)^2 - (2m+1)N = 0 \tag{3}$$

$$\Rightarrow N = 0(\text{trivial}) \text{ and } \quad N = 2m+1 \tag{4}$$

Therefore choosing $\mathcal{G}_0$ to be a complete graph with $2m+1$ nodes is sufficient for the condition $E(0) = mN(0)$.

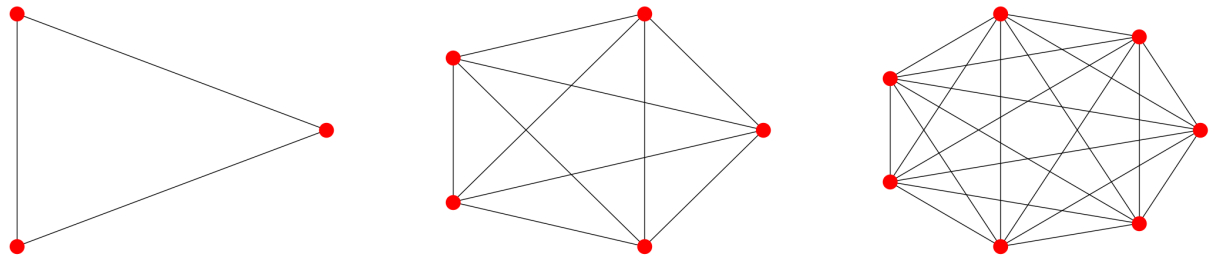


Figure 1: $\mathcal{G}_0$ *for m=1,2,3 respectively.*

1.2 Double Edges

It is also uncear in this model whether we allow more than one edge connecting two nodes. How this might affect the outcome is uncertain. If there are nodes with large degree, according to preferential attachment, they are more likely to be choosen again when attaching more edges, so they may be very common. On th other hand, the probability of a node being chosen twice in a large network is very small, and so double edges may be very are. In testing the latter case was proven to be correct(figure 2). Although this is compelling evidence, I decided to work with a model which excludeds double edges, as it is more applicable in physical interpretation. For example, when modelling the links between websites or friendships, double undirected edges are nonsensical.

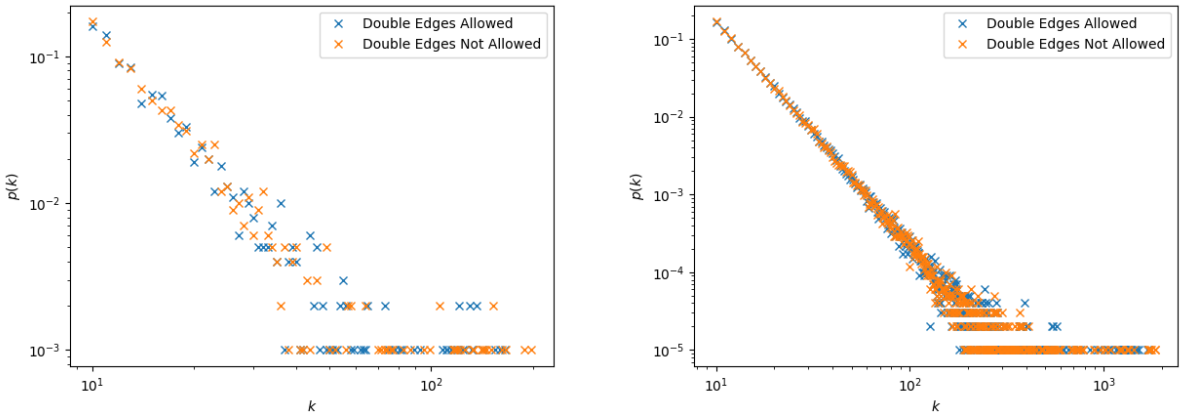
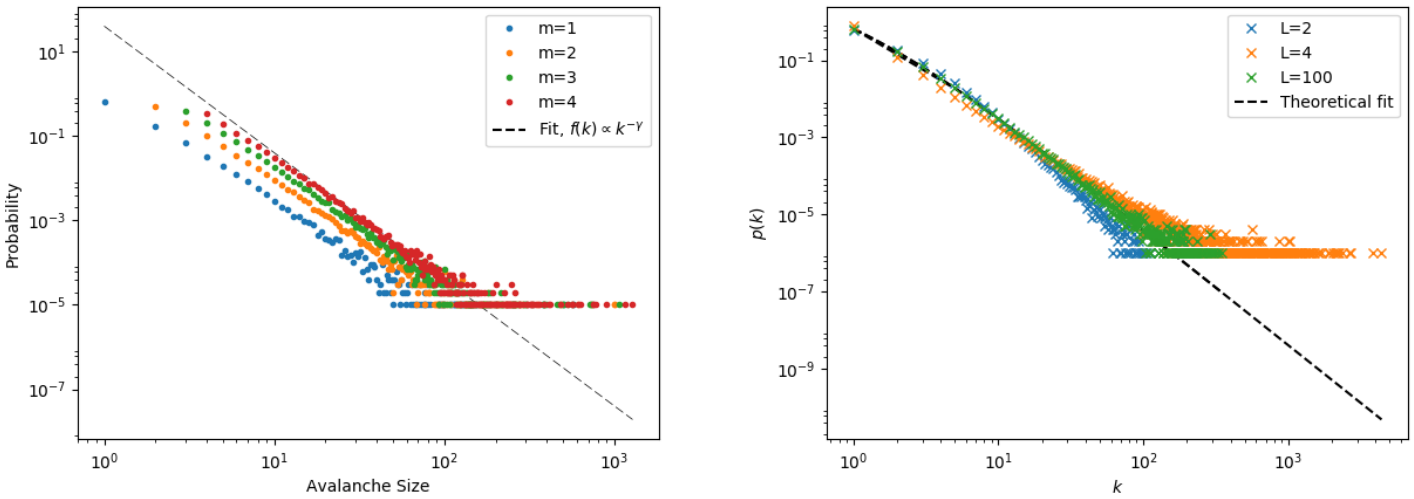


Figure 2: *Degree distribution for m=10, N=1000 and N=100,000 respectively. There is little diference between between the distribution where double edges were allowed, and where double edges weren't allowed.*

1.3 Testing

A way to test whether the programme is working is to analyse the degree distribution(figure 3) so see whether it is what we would expect from a BA model.



(a) (b)

Figure 3: (a) shows the degree distributions of a single trial for $N = 10^5$ and $m = 1, 2, 3, 4$. They exhibits many features of the BA model. Firstly there is a prominent 'fat tail' to our distribution. This relates to having single nodes with high degrees(hubs), however the majority having fairly low degree.
Secondly the figure shows that there distributions follow a powerlaw, $k^{-\gamma}$, where γ will be derived later. This is a definition of a scale free network.
(b) shows the degree distributions of 1000 trails for $N = 10^4$ and $m = 1, 2, 3, 4$. The fat tail of the distribution is a lot more defined, since there is more data. I shall use this method of rapeating trials throughout my report to obtain better statistics and understanding.

2 Degree Distribution of the BA model

2.1 Theoretical Derviation

There are a few ways of approximating the degree distribution $p(k)$, all of which use the master equation. The master equation of a system, desrcibes how the system changes with time evolution. When building the network, for the BA model the master equation is,

$$n(k, t+1) = n(k, t) + m\Pi(k-1, t)n(k-1, t) - m\Pi(k, t)n(k, t) + \delta_{k,m} \tag{5}$$

Where $\Pi(k, t)$ is the probaility of an edge being attached to a particular node of degree k. Since we are taking $\Pi(k, t) \propto k$, and that the probabilities are normalised, way get that:

$$\Pi(k, t) = \frac{k}{\sum_{k=1}^{\infty} kn(k, t)} \tag{6}$$

Where $n(k, t)$ is the number of nodes of degree k. Also, each edge is reponsible for two degrees, so,

$$\Pi(k, t) = \frac{k}{2E(t)} \tag{7}$$

I have already discussed that $E(t) = mN(t)$ using the initial conditions chosen, and so substituting $E(t)$ for $N(t)$, the probability becomes $\Pi(k, t) = \frac{k}{2mN(t)}$. Applying this to (1):

$$n(k, t+1) = n(k, t) + \frac{(k-1)n(k-1, t)}{2N(t)} - \frac{kn(k, t)}{2N(t)} + \delta_{k,m} \tag{8}$$

Now we define the probability of choosing any node of degree k at time t:

$$p(k,t) = \frac{n(k,t)}{N(t)} \tag{9}$$

$$(1) \Rightarrow \quad N(t+1)p(k,t+1) - N(t)p(k,t) = \frac{(k-1)}{2}p(k-1,t) - \frac{k}{2}p(k,t) + \delta_k, m \tag{10}$$

In order to go further, we assume that $p(k,t)$ converges for large t. That is that $p_\infty = lim_{t\to\infty}p(k,t)$
Applying this to (6) the final form of our master equation becomes:

$$(N(t+1) - N(t))p_\infty(k) = -\frac{(k-1)}{2}p_\infty(k-1) - \frac{k}{2}p_\infty(k) + \delta_k, m \tag{11}$$

We note that $N(t) = t$ and so we find the final form of the master equation:

$$p_\infty(k) = \frac{1}{2}((k-1)p_\infty(k-1) - kp_\infty(k)) + \delta_{k,m} \tag{12}$$

2.2 Continuous Approximation

Equation (12) can be used to find the degree distribution for the model. An approximation of this distribution can be found the continuous case,$k+1 \to k+\Delta k$. (12) becomes:

$$p(k) \approx \lim_{\Delta k\to 0} \frac{-\frac{1}{2}((k-\Delta k)p_\infty(k-\Delta k) - kp_\infty(k)) + \delta_{k,m}}{\Delta k} \tag{13}$$

$$\Rightarrow p(k) \approx \frac{\partial k p_\infty(k)}{\partial k} \tag{14}$$

By inspection we find that $p(k) \propto k^{-3}$ is a solution. This means that the degree distribution follows a power law, and so implies that the BA model is scale free.
As k grows the continous case becomes more accurate, as relatively the difference between k and $k-1$ becomes less. So we would expect this case to be true as $m \to \infty$, as this forces higher degrees on the system.

2.3 Difference Derivation

It is possible to derive a solution from the difference equation. First we look at $k > m$ and rearrange (7):

$$\frac{p_\infty(k)}{p_\infty(k-1)} = -\frac{k-1}{2(k+1)} \tag{15}$$

We use an identity of the Gamma function. The equation:

$$\frac{f(z)}{f(z-1)} = \frac{z+a}{z+b} \tag{16}$$

Has solution,

$$f(z) = A\frac{\Gamma(z+1+a)}{\Gamma(z+1+b)} \tag{17}$$

Therefore our difference equation has solution

$$p_\infty(k) = A\frac{\Gamma(k)}{\Gamma(k+2)} \tag{18}$$

Using the identity $\Gamma(n) = (n-1)!$ for $n \in \mathbf{N}_0$, the solution becomes:

$$p_\infty(k) = \frac{A}{k(k+1)(k+2)} \tag{19}$$

The constant A can be found by looking at the boundary case ($k = m$) of (7),

$$p_\infty(m) = -\frac{m}{2}p_\infty(m) + 1 \tag{20}$$

$$\Rightarrow p_\infty(m) = \frac{1}{m+2} \tag{21}$$

This boundary conditon implies that,

$$A = 2m(m+1) \tag{22}$$

Thus we derive the solution to the difference equation as:

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{23}$$

I expect this distribution to be more accurate than that obtain by the continuous derivation, as I have made less assumptions and approximations.
Checking that this equation is normalised,

$$\begin{aligned}
&\sum_{k=0}^{\infty} p_\infty(k) = 2m(m+1)\left\{\sum_{k=0}^{\infty}\frac{1}{k(k+1)(k+2)}\right\} \\
&\Rightarrow= 2m(m+1)\left\{\sum_{k=0}^{\infty}\frac{1}{2k} - \frac{1}{k+1} + \frac{1}{2(k+2)}\right\} \\
&\Rightarrow= 2m(m+1)\left\{\frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{1}{k} - \frac{1}{k+1}\right) + \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{1}{(k+2)} - \frac{1}{k+1}\right)\right\}
\end{aligned}$$

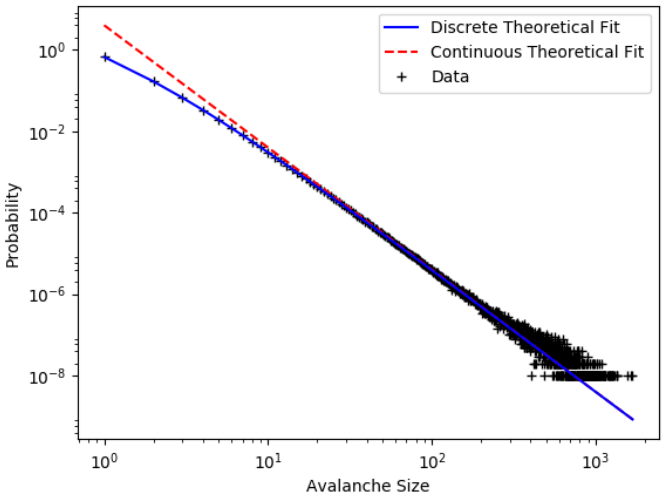
Expanding these sums for the first few terms, it is obvious that only two terms survive,

$$\begin{aligned}
\Rightarrow 2m(m+1)\frac{1}{2}&\left(\frac{1}{m} - \frac{1}{m+1} + \frac{1}{m+1} - \frac{1}{m+2} + ...\right) \\
&+ \frac{1}{2}\left(\frac{1}{m+2} - \frac{1}{m+1} + \frac{1}{m+3} - \frac{1}{m+2} + ...\right) \\
\Rightarrow 2m(m+1)&\left\{\frac{1}{2m} - \frac{1}{2(m+1)}\right\} = \frac{2m(m+1)}{2m(m+1)} = 1
\end{aligned} \tag{24}$$

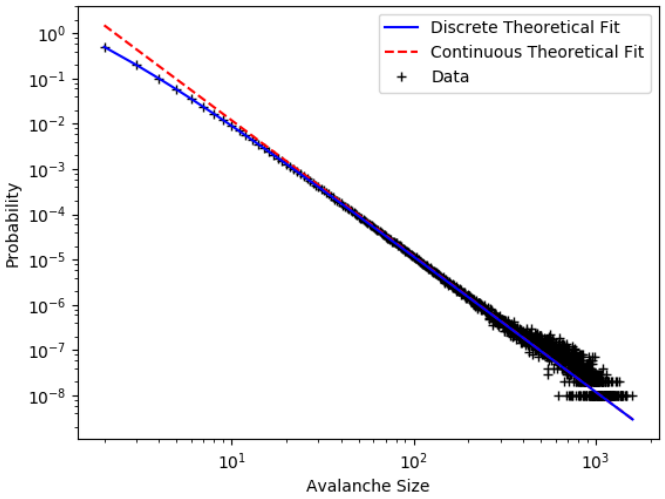
Therefore the probabilities are normalised.

2.4 Comparison with Real Data

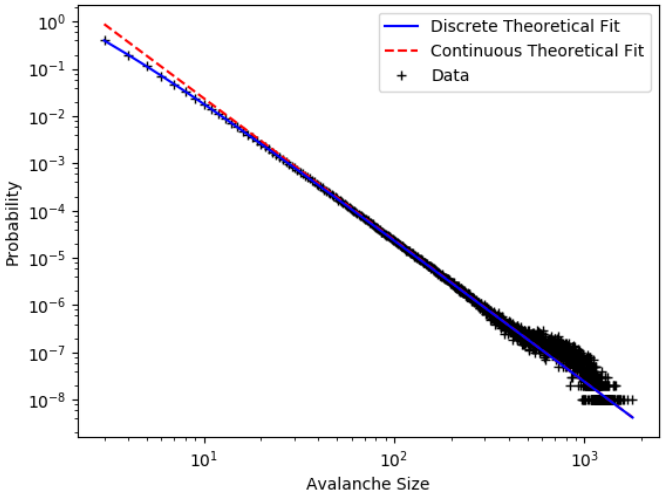
Now I wish to compare these theoretical plots with the actual data captured by my model.
I shall run my programme for $m = 1, 2, 3, 4$ and for graphs of 10^5 nodes. I believe this is a large enough network to assume that any affects cause by size, such as correction to scaling, are negligible. Also in order to derive the degree distributions, we required $\lim_{t\to\infty} p(k,t) = p_\infty(k)$, and so choosing large N, means that we have$t >> 1$.
A key characteristic of the model is that as one increases the number of nodes in the graph (N), the maximum dregee k_1 observed also increase, which means no matter big the graph, the data towards the larger degrees will always be sparse. As already mentioned, to combat this I ran the same experiment 1000 times in order to build up a enough observatinos for large degrees. Figure 4 shows the continuous and discrete fits derived earlier. Visually, one can see from that for small values of k, the probability distribution follows the discrete fit perfectly. This is because there are a lot more nodes with degree small k, and so a lot more data is available, thus the distirbution is prominent. However, for large k we have fewer and fewar nodes per degree, as predicted. This creates the 'fat tail' affect present on all three figures as seen in .
Notice as well that as m increase, our theoretical, and practical data moves closer to the line $p(k) = k^{-3}$. This verfifies the effect I would expect to see for large m.



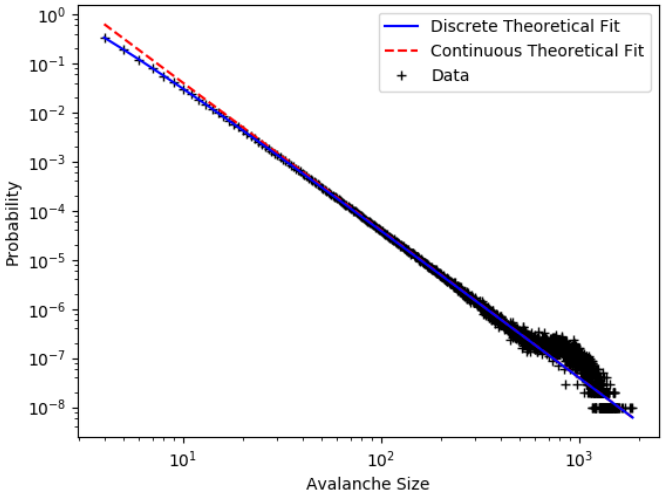
(a) $m = 1$



(b) $m = 2$



(c) $m = 3$



(d) $m = 4$

Figure 4: The loglog plots of the raw data porbability distribuiton. This data was captured from networks of size 10^5, and over 10^3 traisl. a,b,c,d show the outcome for m=1,2,3,4 repectively.

2.5 Statistical Approach

I wish to analyse the model statistically. However, the 'fat tail' characteristic of the data means that the majority of our points lie near the origin. Therefore any statistical method I use will be dominated by the 'fat tail', and any conclusion I draw will be obscured. Log binning is way of mimising the 'fat tail, while keeping the necessary characteristics of the probability distribution.
When creating probability distributions from samples, data is put into bins, the frequency recorded and then normalised. In most cases we have a bin lengths constant, $b_{n+1} - b_n =: \Delta$. However in a log bin proccess, the bins have a relation $\frac{b_{n+1}}{b_n} = \Delta$. This means that the bins increase exponentially as the data grows larger. This means for small k, where the data are plentiful, the bins are small to capture as much information as possible. However for large k, where our data is sparse, the bins are large, meaning the a lot of data is grouped together to help gain insight into the behviour. The geometric mean of each bin is then plotted. Figure 4 shows how the log bin processes the raw data.

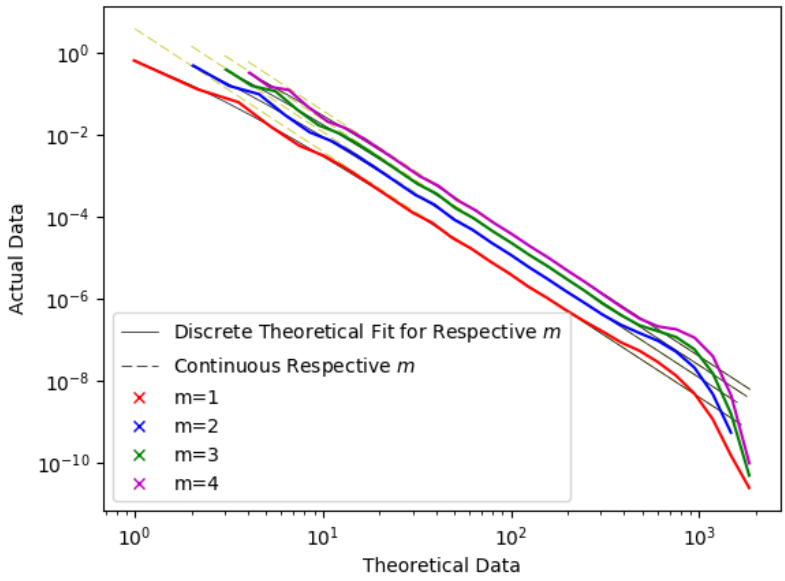
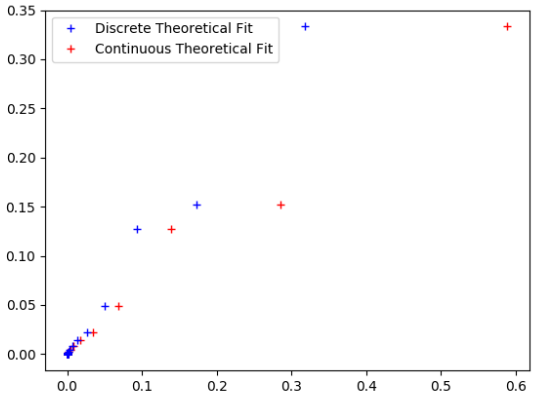


Figure 5: The degree distribution for random attachment. Note the distinctive exponential cut off and lack of a fat tail. This data was taken from graphs of N=10,000, over 100 trails, for m=1,2,3,4.

To statisitically analyse this data I plotted the theoretical fit and naive fit vs. the actual logbinned data. Assuming that my two fits and the actuall data are related, one would expect to see a linear regression between the variables. Figure 5 shows an example of this.



	Linear Regression for Log Binned Data					
	Naive Fit			Theoretical Fit		
	R^2	p-value	Slope	R^2	p-vaule	Slope
m=1	0.992	2.71e-25	0.1628	0.999	2.87e-40	0.9824
m=2	0.993	2.71e-25	0.3637	0.999	2.86e-40	1.0444
m=3	0.993	2.72e-25	0.4886	0.996	2.69e-28	1.0394
m=4	0.991	2.35e-24	0.5719	0.994	2.72e-26	1.0312

Figure 6: The theoretical model(red) and the naive model(black) vs. the log binned, for $m = 4$.

These relation can be explored further by calculating the least squares regression statistics(R^2). I used the Pearson R^2 statistic as the visual evidence from figrue 5 implies that our relation is linear. Using these I tested with the null hypothesis that there is no correlation vs. the hypothesis that they are correlated. The R^2 statistic, the p-values for this hypotesis, and the least regression slope can be found in figure 5b).

The p-statistic is very small in all cases. Infact we can reject the null hypothesis with a confidence level of 1.e-25 for each fit. This impies that there is definitly correlation, and therefore both models would suffice. To analyse which model is better one needs to examine the R^2 terms and slopes. For all m, the R^2 statistic is greater for the theoretical fit, than that of the naive fit. This implies that the theoretical fit is better.
Analysing the slope, one would expect the slope to be close to 1, if the theoretical explained the actual data well. Refering to figure b) shows that is is true for the theoretical fit.
Note that the difference R^2 statistics for the two fits shrinks. This is more evidence that points towards the phenomena that as m increases our two fits converge.

3 Finite Size Effect

3.1 Theoretical Derivation

I wish to find how the greatest degree k_1. First I assume that there is only one node with degree k_1. And therefore:

$$\sum_{k1}^{\infty} p_\infty(k) = \frac{1}{N(t)} \tag{25}$$

From the partial fraction form of the degree distribution, and applying the same logic as I did for normalisation, we have

$$\sum_{k_1}^{\infty} \frac{2m(m+1)}{k(k+1)(k+2)} = \frac{2m(m+1)}{2k_1(k_1+1)} = \frac{1}{N(t)} \tag{26}$$

Rearranging this formula gives a quadratic in k_1,

$$2k_1(k_1+1) = 2Nm(m+1) \Rightarrow k_1^2 + k_1 - Nm(m+1) = 0 \tag{27}$$

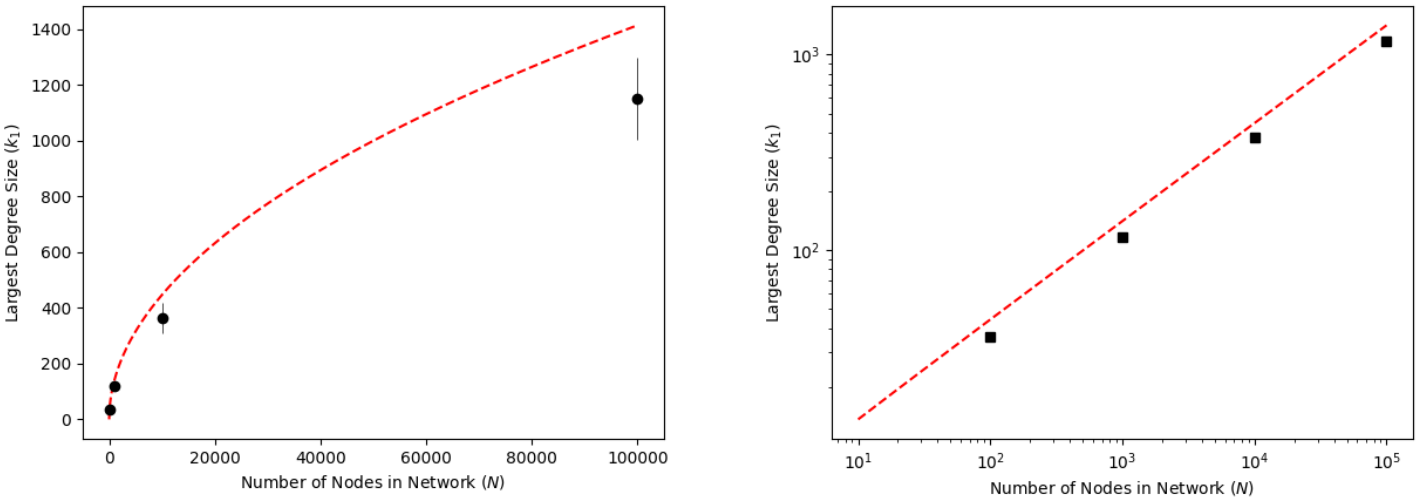
The positive solution gives,

$$k_1 = \frac{1}{2}(-1 + \sqrt{1 + 4Nm(m+1)}) \tag{28}$$

Therefore, for large N and m, we expect, $k_1 \propto \sqrt{N}$.

3.2 Experimental Data

To test how k_1 scales N, I ran my program for $N = 10^2, 10^3, 10^4, 10^5$, collecting the maximum degrees. I focus on $m = 4$, the maximum m I recorded. THis is because I need to fulfill the $Nm >> 1$ condition in order to observe $k_1 \propto \sqrt{N}$. We are interested in these larger degrees, as they obviously depend a lot more on the finite size of the graph. The fact that this is an extreme statistic means that it is very intermittent, and therefore one would expect to see a lot of variance between trials. Therefore I repeated this experiment 100 times, taking the mean and variance to gain a better statistic.



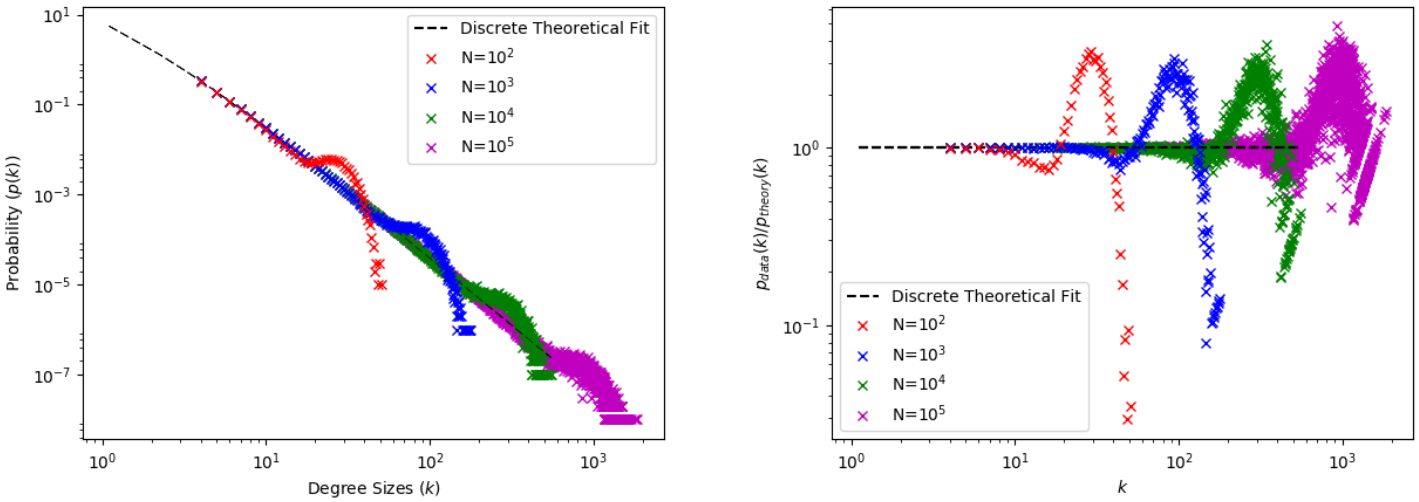
(a) (b)

Figure 7: The mean k_1 for $N = 10^2, 10^3, 10^4, 10^5$. The fit(red line) is the one described by equation (20). a) Shows the plot, and b) shows the log-log plot.

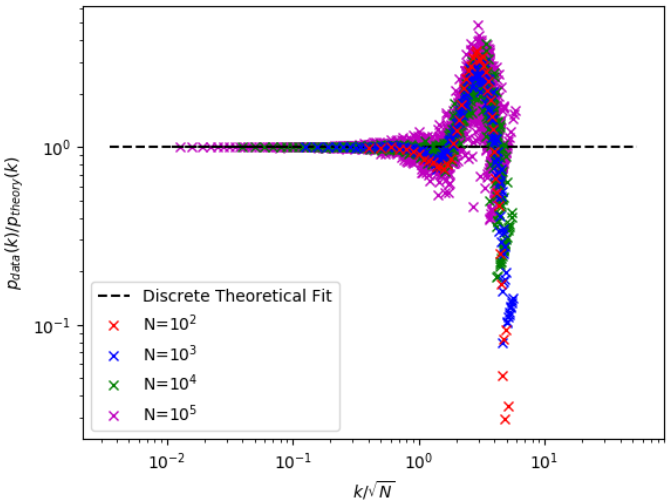
Figure 6 (a) shows that the fit my not entirely describe the dependence of k_1, however in figure 6 (b), the fit and the data are parellel, in log log space. This implies that k_1 does scale with $\sqrt{N}$.

3.3 Data Collapse

Since we now know how the degree distribution depends on k, m and N, I can perform a data collapse. I concentrate on m=4, as it gives me better statistical data for large degree sizes. I vary $N = 10^2, 10^3, 10^4, 10^5$. Figure 6 shows a plot of how the uncollapsed degree distributions correlate.



(a) (b)



(c) $m = 3$

Figure 8: a) Shows the uncollapsed probability distributions for $N = 10^2, 10^3, 10^4, 10^5$, $m = 4$. b) Shows the collapse when dividing through by the theoretical distribution.

First I collapse the probability distribution in k by dividing through by my theoretical prediction:

$$c_k(N) = \frac{p_{data}(k)}{p_{theory}(k)} \tag{29}$$

Figure 5 shows the outcome of this proccedure. Now I collapse in N. To do this I use the relation I earlier derived of how the k_1 scales with N. I stretched the degrees observed by a factor of $N^{-\frac{1}{2}}$. Figure 7 shows the full collapse.
In both cases, instead of using errorbars, I have opted to just plot the raw data to show the spread of the data. For a signle trail, this is not good measure, as extreme statiscics are few and far between. However over a large number of trails (1000 in this case), the structure of the data becomes very apparent. For $N = 10^5$ one can see that there are still a few outlying points, implying that more trail maybe necessary to get an exact idea of the spread, however I still feel the behaviour of the data is still caputured fairly well.

4 Random Attachement

It is also useful, to see how preferential attachement compares with a graph with random attachment. In this network has the same probability of being having an edge attached, or more mathematically $\Pi(k,t) = \frac{1}{N(t)}$, where $N(t)$ normalizes the probability.

4.1 Degree Distribution

The master equation will be the same as when considering pereferential attachment.

$$n(k,t+1) = n(k,t) + m\Pi(k-1,t)n(k-1,t) - m\Pi(k,t)n(k,t) + \delta_{k,m} \tag{30}$$

Using the new probabilities and that $p(k,t) = n(k,t)/N(t)$:

$$N(k,t+1)p(k,t+1) - N(k,t)p(k,t) = mp(k-1,t) - mp(k,t) + \delta_{k,m} \tag{31}$$

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

$$p_\infty(k) = mp_\infty(k-1) - mp_\infty(k) + \delta_{k,m} \tag{32}$$

Concetrating on the case $k = m$:

$$p_\infty(m) = -mp_\infty(m) + 1 \quad \Rightarrow \quad p_\infty(m) = \frac{1}{m+1} \tag{33}$$

Using this probability to the case k=m+1:

$$p_\infty(m+1) = \frac{m}{m+1} - mp_\infty(m+1) \quad \Rightarrow \quad p_\infty(m+1) = \frac{m}{(m+1)^2} \tag{34}$$

Therefore I make the assumption that $p_\infty(k) = (\frac{1}{m+1})(\frac{m}{m+1})^{k-m}$ it true for all k. Using the case $k = m$ as an anchor step, and proving the relation from $k > m$ and $k+1$:

$$\begin{aligned} &p_\infty(k+1) = mp_\infty(k) - mp_\infty(k+1) \\ &\Rightarrow p_\infty(k+1) = m\left(\frac{1}{m+1}\right)\left(\frac{m}{m+1}\right)^{k-m} - mp_\infty(k+1) \\ &\Rightarrow p_\infty(k+1) = \left(\frac{1}{m+1}\right)\left(\frac{m}{m+1}\right)^{(k+1)-m} \end{aligned} \tag{35}$$

Therefore by induction the above assumption is true. To check that this equation is normalised:

$$\begin{aligned} &\sum_{k=m}^{\infty} p_\infty(k) = \left(\frac{1}{m+1}\right)\sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k-m} \\ &\Rightarrow= \left(\frac{1}{m+1}\right)\left(\frac{m+1}{m}\right)^{m}\sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k} \\ &\Rightarrow \left(\frac{1}{m+1}\right)\left(\frac{m+1}{m}\right)^{m}\left(\frac{\left(\frac{m}{m+1}\right)^{m}}{1-\frac{m}{m+1}}\right) = 1 \end{aligned} \tag{36}$$

The numerical degree distribution for the random attachment follows the theoretical distributions derived(figure ..4..). Unlike the preferential attachement, the random data does not have a fat tail, and appears to follow and exponetial cut-off instead. This is to be expected, as it is a lot less likely to find big hubs in a this model, than that ofpreferential attachment. This implies that this model is does not follow a powerlaw and is therefore nota scale free network.

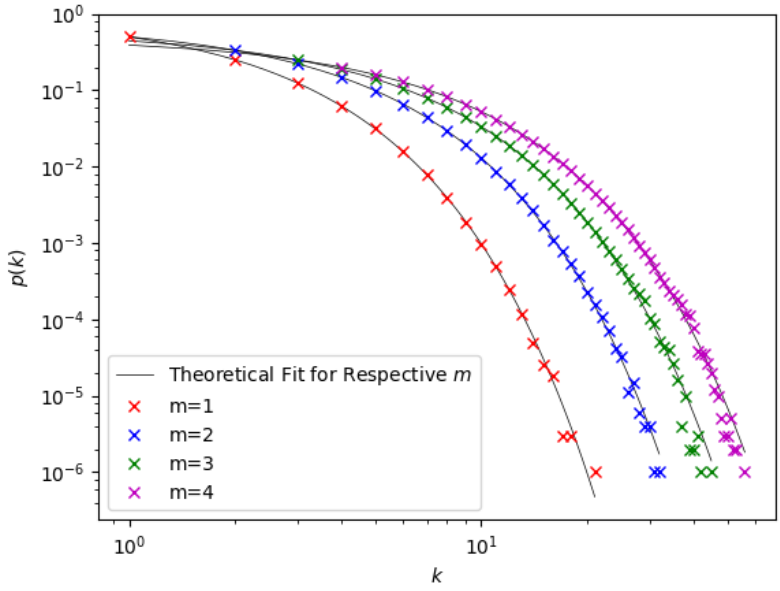


Figure 9: The degree distribution for random attachment. Note the distinctive exponential cut off and lack of a fat tail. This data was taken from graphs of N=100,000, over 100 trails, for m=1,2,3,4.

4.2 Theoretical Derviation of Largest Degree

Using the same logic as with the preferential largest degree,

$$\left(\frac{1}{m+1}\right)\sum_{k_1}^{\infty}\left(\frac{m}{m+1}\right)^{k-m} = \frac{1}{N} \tag{37}$$

Changing the index to $k = k_1 + k'$, we have,

$$\left(\frac{m}{m+1}\right)^{k_1-m}\sum_{k'=0}^{\infty}\left(\frac{m}{m+1}\right)^{k'} = \frac{m+1}{N} \tag{38}$$

It is obvious that $\left|\frac{m}{m+1}\right| < 1$, and so we can treat it as a convergent geometric sum:

$$\begin{aligned}
&\left(\frac{1}{m+1}\right)^{k_1-m}\left(\frac{m}{1-\frac{m}{m+1}}\right) = \frac{m+1}{N} \\
&\Rightarrow \left(\frac{m}{m+1}\right)^{k_1-m}(m+1) = \frac{m+1}{N} \\
&\Rightarrow \left(\frac{m}{m+1}\right)^{k_1} = \frac{m^m}{N(m+1)^m}
\end{aligned} \tag{39}$$

Taking the logarithm,

$$\begin{aligned}
k_1 log\left(\frac{m}{m+1}\right) &= m log\left(\frac{m}{m+1}\right) - log(N) \\
k_1 &= m - \frac{log(N)}{log(m) - log(m+1)}
\end{aligned} \tag{40}$$

Therefore we have that $k_1 \propto log(N)$ for large N. Note that k_1 does not follow a power law relation with growth, implying again that this is not a scale-free network.

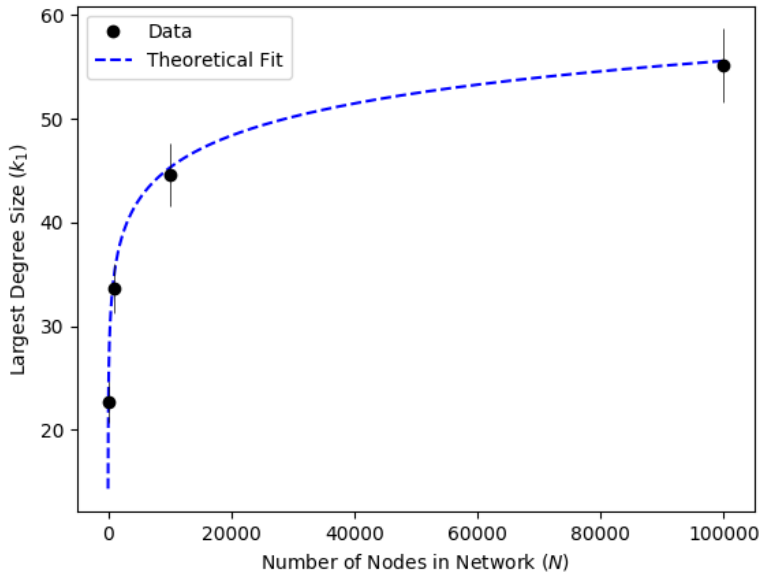


Figure 10: The mean largest degree k_1 computed from 100 trials, for $m = 4$, $N = 10^2, 10^3, 10^4, 10^5$. Error bars represent the standard deviation of each sample. The blue line is the theoretical fit desribed in (35). This data follows this fit to a high level of accuracy.

5 Random Walk Preferential Attachment

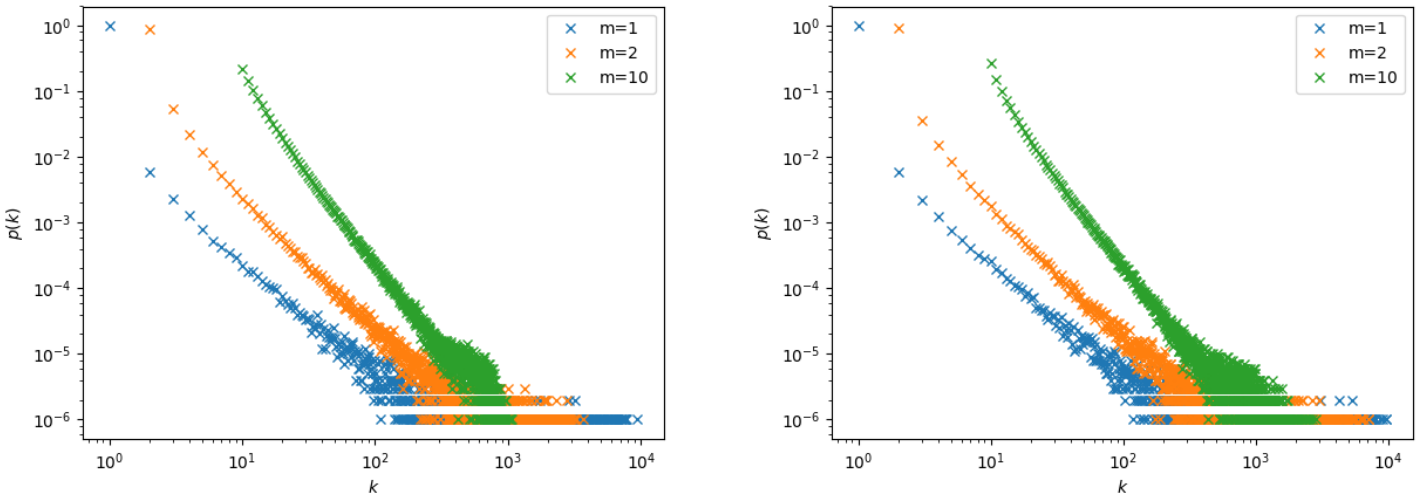
This case is similar to that of the random attachment, however every vertex we pick, we then perform a random walk of length L, and attach an edge to our finishing node. I will study the affects of this for different walk size L.

5.1 $L = 0$

This case is trivial. If L=0, then we do not perform a random walk, and we recover the random preferential attachment case from before.

5.2 $L = 1$

This case is more interesting. Figure 11 (a) shows the degree distributions for $m = 1, 2, 10$. One can see that fat tails are present and that the distributions do follow a powerlaw, implying it is a scale-free network. However, unlike the BA model, the powerlaw is dependent on m, the number of edges added per timestep. The algorithm used is very similar to that of the mediation-driven attachment networks(MDA)(Figure 11 (b)). The MDA model uses preferential attachment like the BA. However, for small m, instead of being desrcibed as using the *'rich-get-richer'* mechanism like the BA, it is often described as using the *'winners take it all'* mechanism, i.e. there are even fewer, more connected hubs.
These networks are created using an almost identical to our algoirthm. The only difference being that, one chooses a single random vertex and chooses from that node's neighbourhood m times, instead of choosing m random vertices and choosing once from each neighbourhood. These networks have been shown to exhibit a power law relation $P(k, m)\tilde{k}^{-\gamma(m)}$ where $\gamma = \frac{1}{\beta(m)} + 1$ is a function of m[2]. Looking at the numerical data, I propose that the model for L=1 follows a similar relation.



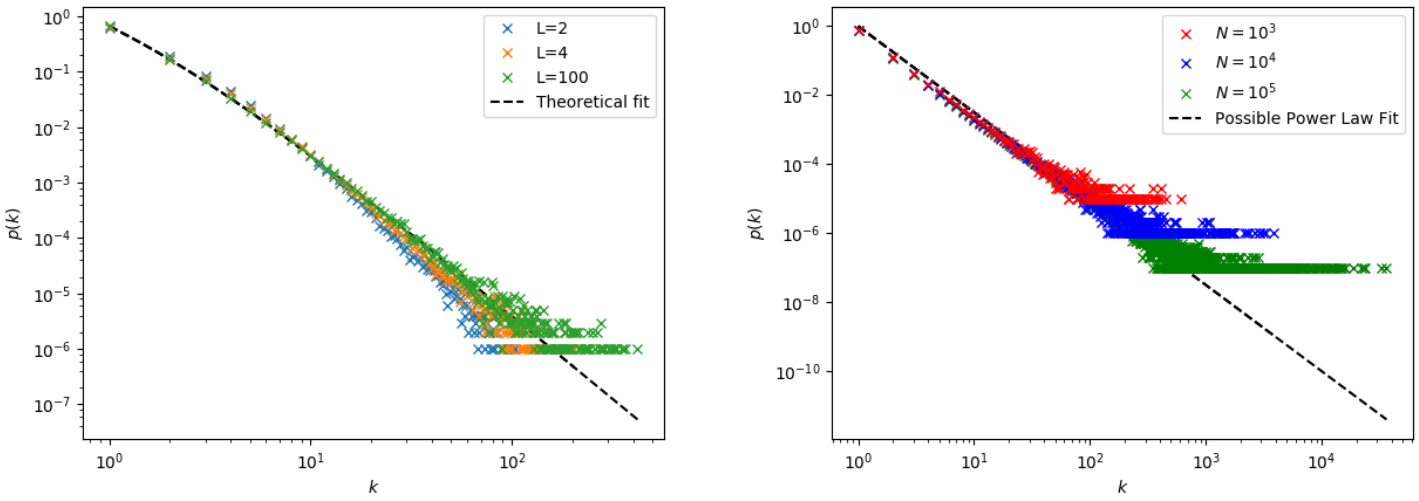
(a) (b)

Figure 11: (a) Shows the model for L=1, $m = 1, 2, 10$ and $N = 10,000$, over 100 trials. (b) Shows the MDA model for the same parameters. There is some differnece between the two models, especialy for m and k. However, with $m = 1, 2$, both models exhibit the 'Winners take it all' behaviour for small m. I.e. The $p(m) \approx 1$, so the majority of the nodes in our network have minumum degree, with a few small 'super hubs' with large degree.
Also note that the gradient of the log log plot in both cases increase with m. THis implies that with have a power law dependent on m as suggested.

5.3 $L > 1$

To explore how the graph changes with longer random walks, I chose 4 cases for L.

1. L=2, to understand what happens for L close to 1.
2. L=4, as the shortest path from a node tot he node with maximum degree is $< l > \sim \frac{\log(N)}{\log(\log(N))} \approx 4$, for N=10,000[3].
3. L=9, As the diameter in the BA model is $D \sim \log(N) \approx 9$ for N=10,000[3].
4. L=100 to reveal the behaviour distribution as $L >> 1$.



(a) (b)

Figure 12: a) shows the degree distribution for $N = 10^4$, $m = 1$, captured over 100 trials, $L = 2, 4, 100$. As L increases there is slight evidence of convergence towards the theoretical fit of BA model, although, all three cases obviously shows behaviour similar to preferential attachment.
(b) Shows the curious case one when L is the diameter of the graph. For this data was captured over 100 trials, from $N = 10^3, 10^4, 10^5$, and $L = 7, 9, 11$ (diameters for respecting network size). Unlike other cases, the degree distribution does not appear to follow the fit, however appears to follow an almost exact, different, power law. The example above(black line) is $p_\infty(k) \propto k^{-2.5}$. Another interesting feature, is the size of the fat tail. Lookng at $N = 10^4$, the fat tail for $L = 9$ is roughly 15 times larger than that of $L = 100$.

As figure 12a) shows, there is large change in the degree distribution for $L > 1$. An interesting feature is that we also recover the BA model. A possible reason for this could be that by performing large walks, we actually recover the noton of eigenvalue centrality. I.e. the probability of landing on a specific node is proportional to it's centrality in the network, and so central nodes are more likely to picked, recovering preferential attachment. This idea is used in analysis methods such as broadcasting and diffusion processes. Figure 12b) shows and describes the curios phenomenon of the third case, when $L \approx D$. In the future I wish to study this case in more detail, to understand why this feature occurs both mathematically and physically.

[2]Degree distribution, rank-size distribution, and leadership persistence in mediation-driven attachment networks, chapter 4, page 27. Kamrul Hassan et al. [3]Structural Properties of Scale-Free Networks, chapter 4.2.3, page 10. Reuven Cohen et al.