Networks Project
Preferential and Random Attachment Models

CID: 01850567

3rd April 2023

Abstract: This project analyses the growth of simple, undirected networks with preferential and random attachment styles. It also looks at a third type of attachment style, existing vertices, where new edges can be formed between existing nodes. The longtime degree distribution and expected largest degree for all models was derived and compared to the simulated data. From doing statistical tests on the data, it was found that networks with preferential and random attachment are modelled better by the theoretical degree distribution. There were finite-size effects exhibited in all networks and these effects were further analysed by implementing a data collapse.

Word count: 2496

1 Introduction

A Barabási-Albert model of network growth can be used to generate scale-free networks as new nodes form edges to existing ones according to preferential attachment [1]. The internet, citation networks, and social networks can be modelled as scale-free.

2 Preferential Attachment (PA)

In this model, a new node is added with m edges every time step. The probability of a new node connecting to an existing one is directly proportional to the degree of the existing node. There are no nodes in the network that have a degree less than m.

2.1 Creating the Programme

Initially, an adjacency list was used to store the network information. Other methods such as an edge list or adjacency matrix was considered but an adjacency list is the most efficient in storing data for large or sparse networks. An object-oriented programming approach was used, such that each network was an 'object' class.
The starter network for this attachment style consisted of a $m + 1$ nodes with each node connected to each other, as shown in Figure 1. Each node has m edges which avoids 'superhubs' appearing too early. When adding a new node, the node connects to m existing edges chosen at random, but weighted by the degree of the node.

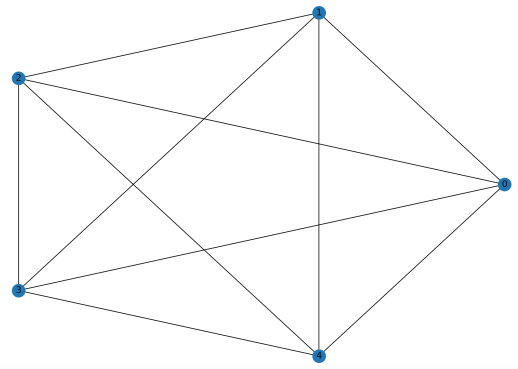


Figure 1: The starter network used for $m = 4$ consisting of $m + 1$ nodes all connected to each other such that each node has m edges.

When checking the code, the random seed was fixed such that the same network can be created. It was ensured that there were no double edges between two nodes as we want to produce a simple, undirected network. After each time step, the number of nodes and edges in the network were checked such that it matches expectations.

2.1.1 Revising the Programme

Although the outlined above programme works accurately, it was very slow as the code had a high level of complexity. A new method implemented.
Instead of using an adjacency list to store data, an `end_edges` list was used: for every edge, the list included the two nodes that the edge is connected to. The starter network in Figure 1 would have `end_edges = [0,0,0,0,1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4]`. When adding a new node, we can choose m nodes randomly (uniform random as nodes with more connections have a higher frequency in this list) from this list and append them to the list. The new node

is added to the list m times, as it has m connections. This method is much more efficient. There were mechanisms built in to prevent self edges and double edges. The downside is that the information on which nodes are connected to each other is lost. The graphs made with this method were compared to those with the previous method to ensure that they gave the same results. The degree distribution can be found by making a histogram of the list.

2.2 Theoretical Degree Distribution

2.2.1 Derivation of the Theoretical Degree Distribution

For preferential attachment, each new node is attached to an existing node with probability

$$\Pi(k,t) = \frac{k}{2E(t)} \tag{1}$$

where k is the degree on the node and E is the total number of edges. The master equation for adding a node to a graph is [2]

$$n(k,t+1) = n(k,t) + m\Pi(k-1,t)n(k-1,t) - m\Pi(k,t)n(k,t) + \delta_{k,m} \tag{2}$$

where $n(k,t)$ is the number of nodes with degree k at time t and $\delta(m,t)$ signifies a new node with degree $m = k$ added to the system. The probability of a node having degree k is given by

$$p(k,t) = \frac{n(k,t)}{N(t)} \tag{3}$$

and in the long time limit,

$$\lim_{t\to\infty} p(k,t) = p_\infty(k) \tag{4}$$

and

$$\lim_{t\to\infty} \frac{E(t)}{N(t)} = m \tag{5}$$

where $p_\infty(k)$ is the long-time or asymptotic degree distribution. Equation 5 is justified because we are adding m edges and one node every time step, so once the number of vertices added greatly exceeds the initial number, $N(t) \gg N(t=0)$ and $mt \gg E(t=0)$, then we can ignore the nodes in the initial graph. Equations 1 and 4 can be substituted into Equation 2 to find the master equation in terms of $p_\infty(k)$:

$$p_\infty(k) = m\Pi(k-1,t)p_\infty(k-1)N(t) - m\Pi(k,t)p_\infty(k)N(t) + \delta_{k,m} \tag{6}$$

where $N(t+1) = N(t) + 1$ was used.
To find an expression of $p_\infty(k)$ for a preferential attachment model, we can substitute the probability given by Equation 1 and Equation 5 into the above equation:

$$p_\infty(k) = \frac{1}{2}[(k-1)p_\infty(k-1) - kp_\infty(k)] + \delta_{k,m} \tag{7}$$

To solve Equation 7, we must rearrange it:

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2} \tag{8}$$

where the $\delta_{k,m}$ term was ignored because it is small.
To solve this equation, we must use the central properties of the Gamma function:

$$\Gamma(z+1) = \Gamma(z) \tag{9}$$

With this property, it can be shown that an equation of the form

$$\frac{f(z)}{f(z-1)} = \frac{z+a}{z+b} \tag{10}$$

has the solution

$$f(z) = A\frac{\Gamma(z+1+a)}{\Gamma(z+1+b)} \tag{11}$$

Equation 10 has a similar form to Equation 8. Hence, we can use Gamma functions and their properties to solve the equation:

$$p_\infty(k) = A\frac{\Gamma(k)}{\Gamma(k+3)} = \frac{A}{k(k+1)(k+2)} \tag{12}$$

where A is the normalisation constant. To normalise the solution, we must consider the boundary condition $k = m$. In our model, there are no nodes that have degree $< m$. From Equation 7, the probability of a node having degree m is

$$p_\infty(m) = \frac{2}{2+m} \tag{13}$$

The sum of the probabilities for each degree must be equal to unity. This, and the boundary condition for $k = m$ can be used to normalise the theoretical degree distribution.

$$\begin{aligned}
\sum_{k=m}^{\infty} p_\infty(k) &= 1 \\
&= \frac{2}{2+m} + \sum_{k=m+1}^{\infty} \frac{A}{k(k+1)(k+2)} \\
&= \frac{2}{2+m} + A\sum_{k=m+1}^{\infty} \frac{1}{2(k+2)} - \frac{1}{k+1} + \frac{1}{2k} \\
&= \frac{2}{m+2} + \frac{A}{2(m+1)(m+2)}
\end{aligned} \tag{14}$$

where the expression was split into partial fraction and the sum to infinity was simplified by observing which terms cancel with each other. Hence,

$$A = 2m(m+1) \tag{15}$$

Finally, the theoretical degree distribution in the long-time limit for the preferential attachment model is

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{16}$$

2.2.2 Simulated Degree Distribution

Using the model outlined in Section 2.1.1, ten different graphs for $m = 4, 8, 16, 32, 64, 128$ and 256 were simulated with $N = 10^6$ nodes added to an initial graph. Two different starter graphs were used: one were all nodes had m edges and one with a single superhub. Results from the simulations are shown in Figure 2. To remove the statistical noise in the tails of the plot, a log-binning method was employed, using an external module written by M. McGillivray (2020), with the constant $a = 1.1$ such that each bin is 10% larger than the previous. There are several distinct features exhibited in Figure 2 that result in deviations from the theoretical degree

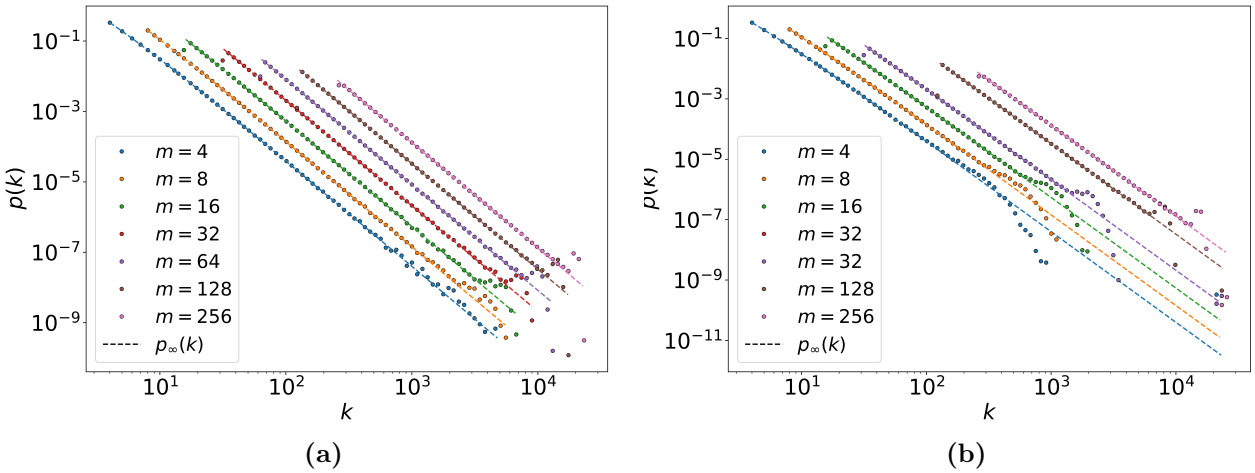


Figure 2: A simulated degree distribution with $N = 10^6$ nodes added to the initial graph for $m = 4, 8, 16, 32, 64, 128$ and 256. The plot was created with averaged data from 10 different graphs per m value and was smoothed using a log-binning method with $a = 1.1$. This number was chosen because it showed the distinct features of the plot the best by reducing noise and also retaining the important information of the data. The dashed lines are the theoretical degree distributions for each m value calculated from Equation 16 The first data point (first bin) in both **(a)** and **(b)** deviates from the theoretical degree distribution. The most likely cause of this is due to the log-binning process - $p_\infty(k)$ is 0 for $k < m$ so if the first bin is above and below the value of m, there will be an offset in the data.. **(a)** has a starter network with $m + 1$ edges per node and **(b)** has a superhub starter network. The bumps are due to the finite network size.

distribution.

The degree distribution follows a power law, so the network is scale-free. Figure 2a shows distinct 'bumps' near to the cutoff for each m. These are finite-sized effects. Equation 16 is the degree distribution for $N \to \infty$. It is impossible to simulate infinite nodes, so there will be a largest possible degree a node can have $\approx N$. The bumps are a result of nodes which, in the longtime limit, would have a degree larger than the cut-off degree, but are constrained by the finite size of the system. The clustering near the cut-off degree produces a pronounced 'bump'. The superhub contains 500 nodes each with m connections except one node which was connected to all other nodes in the system for this initial condition. For each m there is a node that has a much larger degree than the others in the network: this is the superhub node. The characteristic bump occurs at a much lower value of k than in Figure 2a - the effect of the superhub is that the smaller nodes grow much slower as new connections have a strong preference for the superhub.

2.2.3 Statistical Test

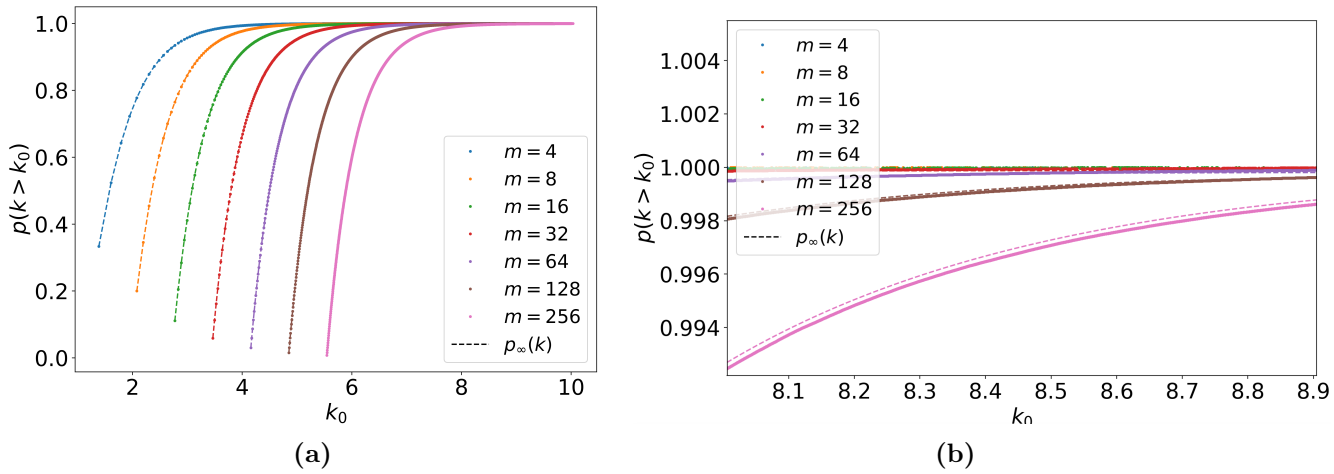
A statistical test was done on the data displayed in Figure 2 to determine the extent at it fits the theoretical distribution. It is important to choose the right test for the data and several were considered. The data is discrete and exhibits fat-tails (large variance) - a non-parametric, discrete test would be preferred.

A chi-squared test is a discrete test, but it is also a parametric test. The results for the test are affected by the choice of binning of the data. We can do objective binning such that each degree has > 5 data points in the bin, as per convention, but for bigger nodes there may be bins with a small number of data points spanning many nodes - there will not be enough values

m	Test Statistic	p-value
4	9.36e-5	0.999
8	1.39e-4	0.990
16	1.12e-4	0.999
32	1.83e-4	0.890
64	1.87e-4	0.877
128	2.96e-4	0.344
256	4.64e-4	0.0269

Table 1: Results from the KS test on the CDF in Figure **??**. The theoretical distribution is a good fit of the data for smaller values of m because it takes longer for large values of m to converge to $p_\infty(k)$.

in each bin without many groupings of bins. An R-squared test was also considered, but this works only for linear regression which is not relevant here.
Lastly, a Kolmogorov-Smirnov (KS) test was decided to be the best for the data. Although the KS test is best used for continuous data, there many data points that we can assume a continuous distribution. The KS test is a measure of the biggest difference between the theoretical cumulative distribution function (CDF) and the actual CDF of the data. It is not affected by any binning choices.



(a) **(b)**

Figure 3: The CDF of the raw data (without any log-binning method employed) shown in Figure 2. The dashed line shows the theoretical degree distribution. **(b)** is a zoomed in image of **(a)** to show the deviations from the theoretical CDF more clearly.

Visually, the CDF of the simulated data is a very close match to that of that of the theoretical CDF. The results from the KS test can be found in Table 1.
The KS test calculates the likelihood that the simulated data was sampled from the theoretical distribution as a p-value. The lower values of m tend to be a better fit to the theoretical distribution. This is to be expected because the theoretical distribution is for the long-time limit. It would take longer for larger values of m to converge to the theoretical distribution. Figure 3b shows the deviations from $p_\infty(k)$ at higher m values.

2.3 Finite Size Effect

To analyse the finite size effect, a single value of m was chosen and several graphs were created with $N = 10^i$ for $i = 2, 3, 4, 5, 6$ nodes added. $m = 128$ produces the most prominent bump

whilst also being feasible to run. Ten different graphs were produced for each value of N to smooth out the data. The simulated distribution is shown in Figure 4. A data collapse can be

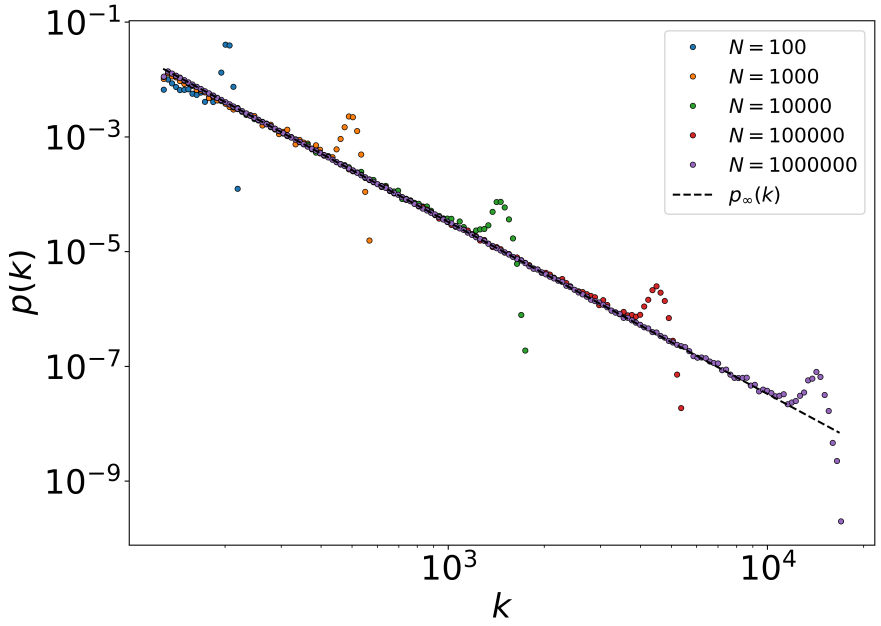


Figure 4: Degree distribution for $m = 128$ and $N = 10^i$ for $i = 2, 3, 4, 5, 6$. A log-binning method was employed to remove statistical noise with $a = 1.03$. For each value of N, ten different graphs were produced to smooth the data. The dashed line shows the theoretical degree distribution.

carried out to further explore the finite-size effect and show where the simulated data deviates from the theoretical expectation. First, we must derive a theoretical expression for the largest expected degree, k_1.

2.3.1 Derivation of the Largest Expected Degree

There is only one node which has the largest expected degree, and no nodes that have a degree that is greater:

$$\sum_{k=k_1}^{\infty} N p_\infty(k) = 1 \tag{17}$$

By substituting in Equation 16, an expression for k_1 can be produced.

$$\begin{aligned} \sum_{k=k_1}^{\infty} N p_\infty(k) = 1 &= N \sum_{k=k_1}^{\infty} \frac{2m(m+1)}{k(k+1)(k+2)} \\ \frac{1}{N} &= \sum_{k=k_1}^{\infty} \frac{1}{2(k+2)} - \frac{1}{k+1} + \frac{1}{2k} \\ &= \frac{2m(m+1)}{2k_1(k+1)} \\ k_1(k_1+1) &= Nm(m+1) \end{aligned} \tag{18}$$

This is a quadratic expression for k_1. We must only consider the positive solution because a negative degree is meaningless.

$$k_1 = \frac{-1 + \sqrt{1 + 4N(m^2+m)}}{2} \tag{19}$$

Hence, Equation 19 is the theoretical expression for the largest expected degree in a preferential attachment graph in terms of N and m.
Another expression for k_1 can be produced by considering the rate at which the degree of a node, k_i, increases:

$$\frac{dk_i}{dN} = m\Pi = \frac{k_i}{2N} \tag{20}$$

where Equation 1 was used. This can be solved by integrating between $t = 0$ where, according to the starter graph used, the node has degree m and there are $m + 1$ nodes, to $t = N$ where the node has largest degree k_1 and there are N nodes.

$$\begin{aligned} \int_m^{k_1} \frac{dk_i}{k_i} &= \frac{1}{2}\int_{m+1}^{N} \frac{dN}{N} \\ \ln\left(\frac{k_1}{m}\right) &= \frac{1}{2}\ln\left(\frac{N}{m+1}\right) \end{aligned} \tag{21}$$

$$k_1^\star = m\sqrt{\frac{N}{m+1}} \tag{22}$$

where the $\star$ symbol was used to differentiate between the two expressions for k_1. Both equations have the relationship $k_1 \propto N^{\frac{1}{2}}$ showing the scale-free property of the PA style.

2.3.2 Data Collapse

Figure 5 shows the different values of k_1 for $m = 4, 64$ and 128 and a data collapse of Figure 4 with these different k_1 values. The data was collapsed by scaling the y-axis by $p_\infty(k)$ and the x-axis by the different values of k_1. There is an offset between the theoretical and empirical values of k_1. It was found that for larger values of N or smaller values of m, this offset became smaller. The derivation of k_1 (Equation 19) makes use of the long-time degree distribution, $p_\infty(k)$ - the simulated k_1 converges to the theoretical value for larger N values.
The data collapse shows where the simulated data deviates from the theoretical data - there is a characteristic bump that occurs right before $k/k_1 = 1$ displaying the effect of the finite graph size. After the bump, the graph rapidly decays, due to the constraint of finite N.
The first point for every N value is lower than the others. As with Figure 2, the explanation for this is likely to do with the log-binning process.

3 Random Attachment (RA)

In a RA model, the probability of a new node connecting to an existing node does not depend on the node's degree. Instead, there is a uniform random probability of it connecting. Only a small alteration of the programme - changing the probability of attachment to a uniform probability - was required to implement the RA model.

3.1 Theoretical Degree Distribution

3.1.1 Derivation of the Theoretical Degree Distribution

The probability of connecting just one edge from a new node ($m = 1$) to the existing graph is

$$\Pi(k, t) = \frac{1}{N(t)} \tag{23}$$

This expression more complicated for connecting greater than 1 node in a simple graph because there cannot be more than 1 edge between nodes. But it is a good approximation for the limit

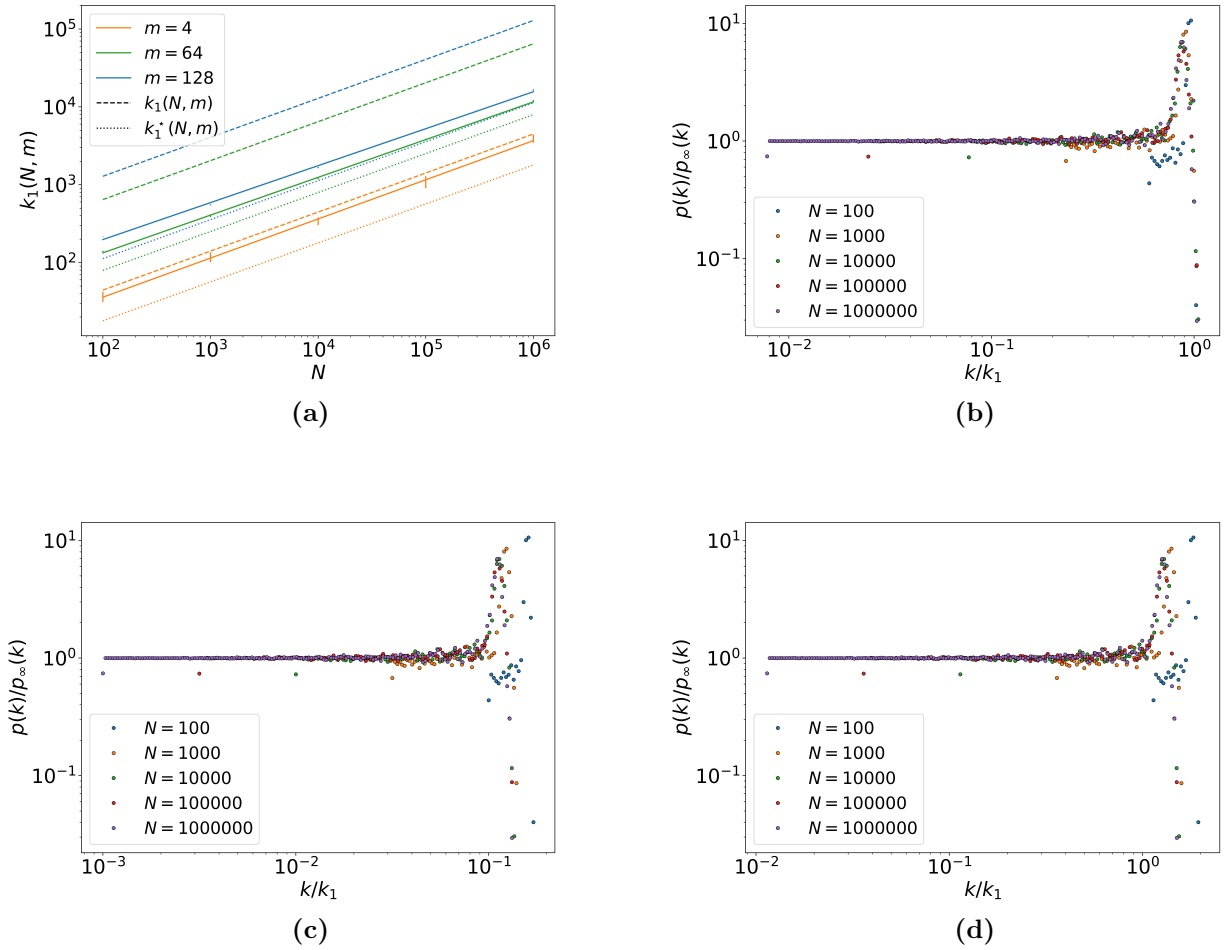


Figure 5: **(a)** is the k_1 plots for the simulated data the theoretical value using Equation 19 (dashed line) and the theoretical value using Equation 22. The solid line shows a linear fit on the empirical values of k_1. According to the linear fit, for $m = 128$ the empirical k_1 has relationship $k_1 \propto N^{0.4746 \pm 0.0001}$, the theoretical k_1 has relationship $k_1 \propto N^{0.500038}$ and $k_1^\star \propto N^{0.500000}$. The different k_1 values are closer for smaller values of m. A data collapse was done on Figure 4 by scaling the y-axis by $p_\infty(k)$ and scaling the x-axis by the empirical value of k_1 in **(b)**, the theoretical value of k_1 with Equation 19 in **(c)**, and $k_1^\star$ in **(d)**. Visually, **(b)** provides the best data collapse, followed by **(d)** then **(c)**.

as $N \to \infty$. By substituting Equation 23 into the master equation given in Equation 6, a recurrence relation for $p_\infty(k)$ for the RA model can be produced.

$$(m+1)p_\infty(k) = mp_\infty(k-1) + \delta_{k,m} \tag{24}$$

which can be written as

$$p_\infty(k) = \left[\frac{m}{m+1}\right]^{k-m} p_\infty(k-1) \tag{25}$$

for $k > m$ and $p(k < m) = 0$ and by ignoring the $\delta_{k,m}$ term. The probability of a node having degree m can be found by substituting $p_\infty(m-1) = 0$ into Equation 24

$$p_\infty(m) = \frac{1}{m+1} \tag{26}$$

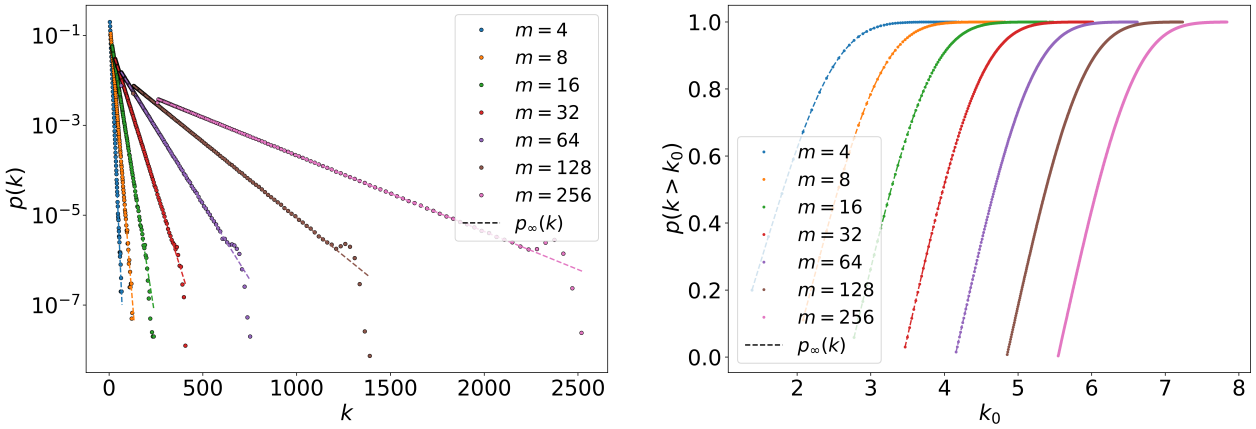
Hence, the degree distribution for the random attachment model in the longtime limit is [3]

$$p_\infty(k) = \frac{m^{k-1}}{(m+1)^{k-m+1}} \tag{27}$$

3.1.2 Simulated Degree Distribution

Figure 6a shows the degree distribution for $m = 4, 8, 16, 32, 64, 128, 256$.

The x-axis for Figure 6a is not logarithmic: the RA model does not produce a fat-tailed dis-



(a) **(b)**

Figure 6: A simulated degree distribution with $N = 10^6$ nodes added to the initial graph for $m = 4, 8, 16, 32, 64, 128, 256$. The plot was created with averaged data from 10 different graphs per m value and was smoothed using a log-binning method with $a = 1.02$. The dashed lines are the theoretical degree distributions for each m value calculated from Equation 27. In **(a)**, there is a distinctive bump, similar to the PA model, before the cut-off degree size. This shows the effect of a finite-sized system. The bump is only visible for larger system sizes (the slope of small m make these bumps difficult to see). The cut-off degree is much smaller than in PA because in PA, bigger nodes have a higher probability of attachment so have a capacity to be more connected than in RA. **(b)** is the CDF of **(a)**.

tribution unlike the preferential attachment. This can be inferred by the theoretical expression of $p_\infty(k)$ for the two models - the logarithm of $p_\infty(k)$ is linear for RA, but is a power law for PA.

m	Test Statistic	p-value
4	8.34e-5	1.000
8	2.04e-4	0.768
16	1.80e-4	0.902
32	1.01e-4	1.000
64	1.14e-4	1.000
128	1.12e-4	0.999
256	1.18e-4	0.886

Table 2: Results of the KS test on the random attachment model degree distribution. The simulated data seems to fit the theoretical distribution better than for the preferential attachment data. The data for $m = 8$ and 256 do not fit as well as the others, but it is still a good fit. This is likely a statistical anomaly.

3.1.3 Statistical Test

A KS test was performed on the CDF of the data, as seen in Figure 6b. The results from the test are in Table 2. The KS test results are good for all values of m.

3.2 Finite-Size Effect

Again, the value of $m = 128$ was analysed for different values of N.

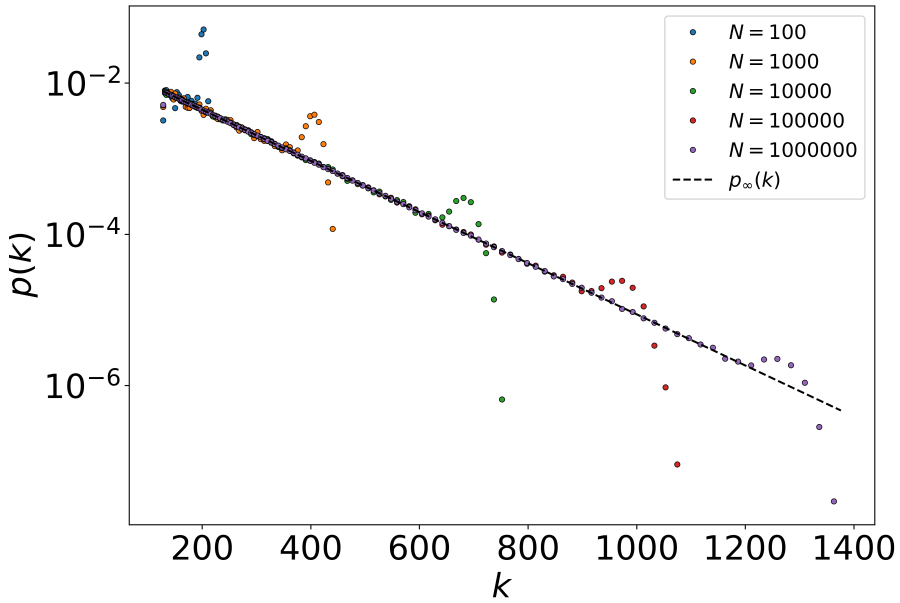


Figure 7: Degree distribution for $m = 128$ and $N = 10^i$ for $i = 2, 3, 4, 5, 6$. A log-binning method was employed with $a = 1.02$. For each value of N, ten different graphs were produced to smooth the data. The dashed line shows the theoretical degree distribution.

3.2.1 Derivation of the Largest Expected Degree

Substituting the theoretical degree distribution, Equation 27 into the definition of k_1, Equation 17, the following expression is produced:

$$\begin{aligned} \sum_{k=k_1}^{\infty} \frac{m^{k-m}}{(m+1)^{k-m+1}} &= \frac{1}{N} \\ \frac{m^m(m+1)^{1-m}}{N} &= \sum_{k=k_1}^{\infty} \frac{m^k}{(m+1)^k} \end{aligned} \tag{28}$$

Using the substitution $k = k_1 + i$,

$$\begin{aligned} \frac{m^m(m+1)^{1-m}}{N} &= \sum_{i=0}^{\infty} \left[\frac{m}{(m+1)}\right]^{k_1+i} \\ &= \frac{m}{m+1}^{k_1} \sum_{i=0}^{\infty} \left[\frac{m}{(m+1)}\right]^{i} \\ &= \left[\frac{m}{m+1}\right]^{k_1} (1+m) \\ \left[\frac{m}{m+1}\right]^{k_1} &= \frac{m^m}{N(m+1)^m} \end{aligned} \tag{29}$$

where the sum of a geometric series was used. By taking the logarithm of both sides, the expression for the k_1 is [3]

$$k_1 = m - \frac{\ln N}{\ln m - \ln m + 1} \tag{30}$$

By considering the rate at which a node's degree increases, another theoretical expression for k_1 can be produced:

$$k_1^\star = m\left(1 + \ln\left(\frac{N}{m+1}\right)\right) \tag{31}$$

This expression was found by substituting the probability of attachment for RA into Equation 20 then solving it.
For the PA, $\ln k_1 \propto \ln N$ whereas in RA, $k_1 \propto \ln N$. This difference stems from the form of $p_\infty(k)$ for the two models and shows that the RA model does not produce a scale-free network.

3.2.2 Data Collapse

Figure 8 shows the empirical value of k_1, and the theoretical values of k_1 from Equations 30 and 31, and the data collapse of Figure 7. The data collapse in Figure 8d is poor - the bumps do not align. This suggests that k_1 given by Equation 30 is a poor description of the empirical data and $k_1^\star$ is better. k_1 from Equation 30 uses $p_\infty(k)$ in the derivation whereas $k_1^\star$ does not. It takes time for the degree distribution to converge to $p_\infty(k)$, especially large m. But since $k_1^\star$ does not depend on $p_\infty(k)$, it is a better fit for smaller N.

4 Existing Vertices (EV)

Every time step, a new node with r is connected to the graph with probability Π_1, and $m - r$ edges are formed between existing nodes with probability Π_2. Again, no double or self-edges were allowed in the model. In this project, $r = m/3$ was chosen, Π_1 is equal to the RA probability and Π_2 is equal to the RA probability.

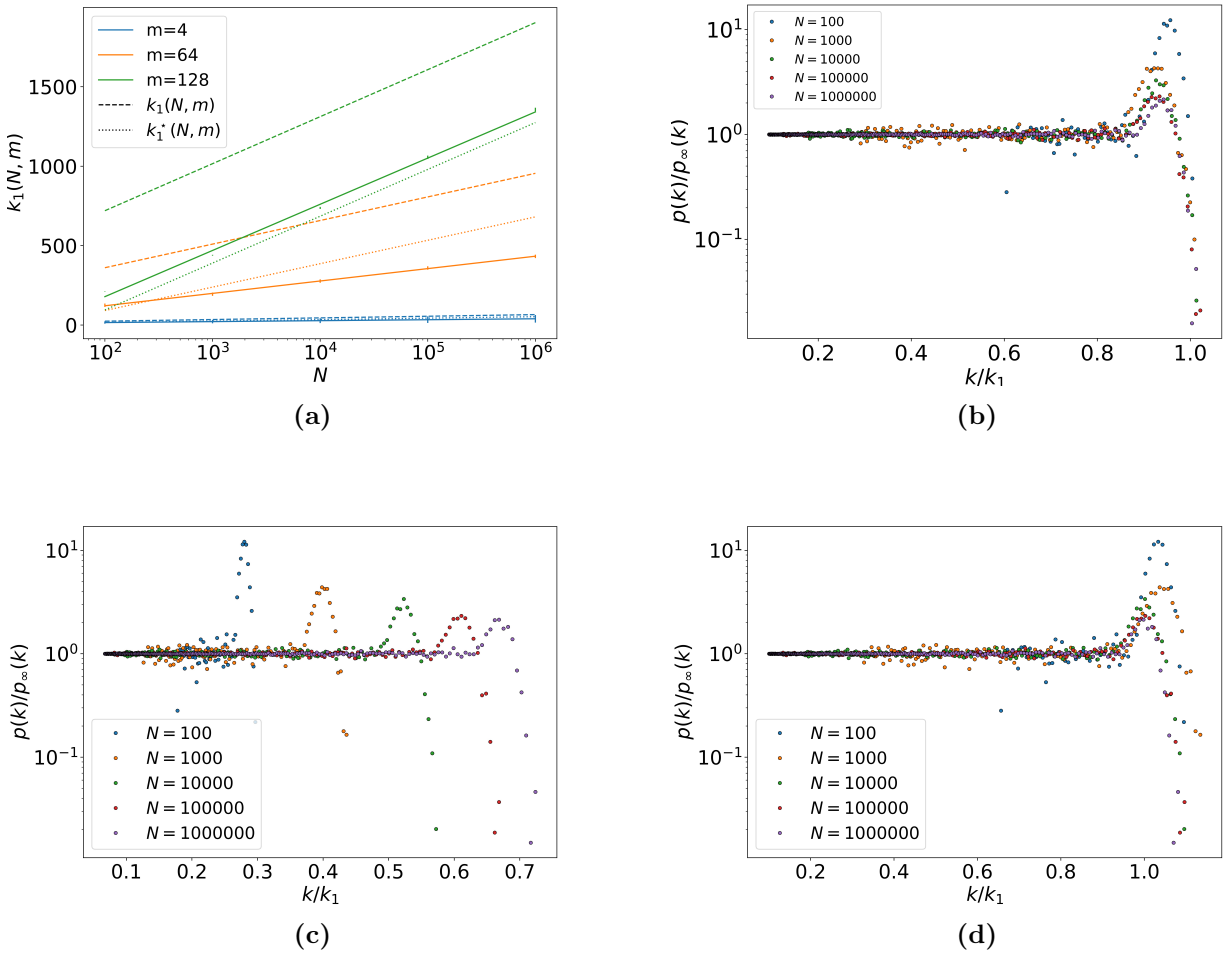


Figure 8: **(a)** is the k_1 plots for the simulated data the theoretical value using Equation 30 (dashed line) and the theoretical value using Equation 31. The solid line shows a linear fit on the empirical values of k_1. A data collapse was done on Figure 7 by scaling the y-axis by $p_\infty(k)$ and scaling the x-axis by the empirical value of k_1 in **(b)**, the theoretical value of k_1 with Equation 19 in **(c)**, and $k_1^\star$ in **(d)**. Visually, **(b)** provides the best data collapse, followed by **(d)**. **(c)** is a poor collapse of the data. In the collapse, most data points lie on $y = 1$ and there is a bump before $k/k_1 = 1$ showing the finite-size effect. There is an anomalous data point for $N = 100$ whose origins are unknown.

4.1 Theoretical Degree Distribution

4.1.1 Derivation of the Theoretical Degree Distribution

The master equation for $p_\infty(k)$, Equation 6, needed to be modified for this attachment style.

$$\begin{aligned} p_\infty(k) = r\Pi_1(k-1,t)N(t)p_\infty(k-1) - r\Pi_1(k,t)N(t)p_\infty(k) \\ +2(m-r)\Pi_2(k-1,t)p_\infty(k-1)N(t) - 2(m-r)\Pi_2(k,t)p_\infty(k)N(t) + \delta_{k,r} \end{aligned} \tag{32}$$

Substituting in Equation 23 and Equation 1 for Π_1 and Π_2 respectively, and inserting Equation 5,

$$p_\infty(k)\left[1 + r + \frac{m-r}{m}k\right] = p_\infty(k-1)\left[r + \frac{m-r}{m}(k-1)\right] + \delta_{k,r} \tag{33}$$

Ignoring the $\delta_{k,r}$, this can be rearranged to get

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{(k-1) + \frac{mr}{m-r}}{k + \frac{m+mr}{m-r}} \tag{34}$$

and this can be solved by using the properties of a Gamma function in Equation 10.

$$p_\infty(k) = A\frac{\Gamma\left(k + \frac{mr}{m-r}\right)}{\Gamma\left(k + 1 + \frac{m+mr}{m-r}\right)} \tag{35}$$

Normalisation constant, A, can be found by considering $p_\infty(k < r) = 0$.

$$p_\infty(r) = \frac{m}{m + 2mr - r^2} \tag{36}$$

and by using the Beta function defined by

$$B(z_1, z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,dt \tag{37}$$

which has the property

$$B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)} \tag{38}$$

Hence,

$$\begin{aligned} \sum_{k=r}^{\infty} = 1 &= \frac{m}{m + 2mr - r^2} + \frac{A}{\Gamma\left(\frac{m}{m-r} + 1\right)} \sum_{k=r+1}^{\infty} B\left(k + \frac{mr}{m-r}, \frac{m}{m-r} + 1\right) \\ &= \frac{m}{m + 2mr - r^2} + \frac{A}{\Gamma\left(\frac{m}{m-r} + 1\right)} \sum_{k=r+1}^{\infty} \int_0^1 t^{k - \frac{mr}{m-r} - 1}(1-t)^{\frac{m}{m-r}}\,dt \\ &= \frac{m}{m + 2mr - r^2} + \frac{A}{\Gamma\left(\frac{m}{m-r} + 1\right)} \int_0^1 t^{\frac{mr}{m-r} - 1}(1-t)^{\frac{m}{m-r}} \left[\sum_{k=0}^{\infty} t^k - \sum_{k=0}^{r+1} t^k\right] dt \\ &= \frac{m}{m + 2mr - r^2} + \frac{A}{\Gamma\left(\frac{m}{m-r} + 1\right)} \int_0^1 t^{\frac{mr}{m-r} + r}(1-t)^{\frac{m}{m-r} - 1} \\ &= \frac{m}{m + 2mr - r^2} + \frac{A}{\Gamma\left(\frac{m}{m-r} + 1\right)} B\left(\frac{mr}{m-r} + r + 1, \frac{m}{m-r}\right) \end{aligned} \tag{39}$$

Finally, the normalisation constant is

$$A = \frac{m + 2mr - r^2}{mB\left(\frac{mr}{m-r} + r + 1, \frac{m}{m-r}\right)} \tag{40}$$

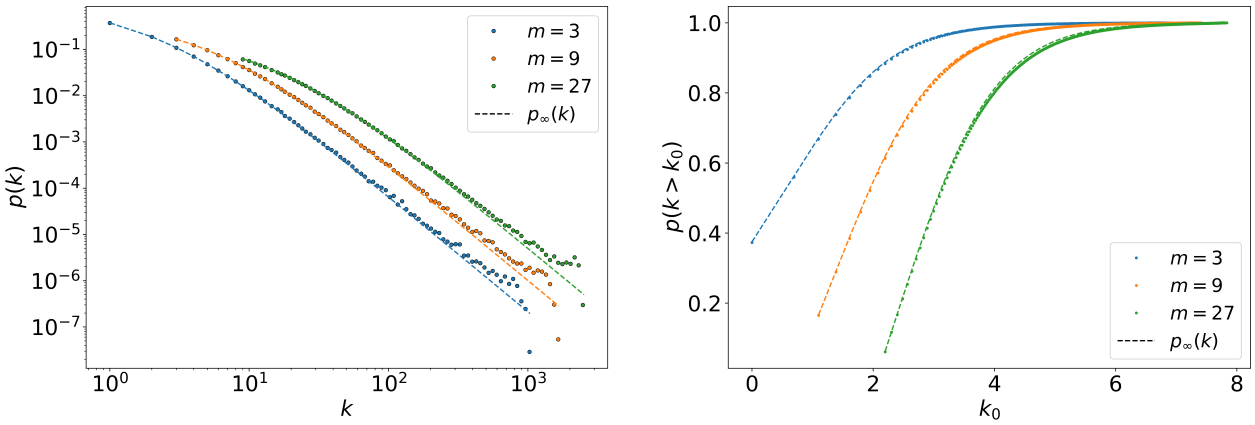
Substituting A into Equation 35 gives the theoretical degree distribution in the long-time limit.

m	Test Statistic	p-value
3	4.64e-3	9.19e-10
9	7.12e-3	2.03e-22
27	1.11e-2	2.30e-54

Table 3: KS test for the EV model for $m = 3, 9, 27$. The KS test fails for all values of m and is worse for higher m.

4.1.2 Simulated Degree Distribution

Figure 9 shows the simulated degree distribution for a fixed $N = 10^5$ and $m = 3, 9, 27$. The graph shows a power-law relationship between $p(k)$ and k suggesting that the EV model produces a scale-free network.



(a) **(b)**

Figure 9: A simulated degree distribution with $N = 10^5$ nodes added to the initial graph for $m = 3, 9, 27$. The plot was created with averaged data from 10 different graphs per m value and was smoothed using a log-binning method with $a = 1.07$. The dashed lines are the theoretical degree distributions for each m value calculated from Equation 35. There is a distinctive bump before the cutoff degree size. For higher values of k, the simulated data seems to deviate from the theoretical degree distribution more than the other models. **(b)** is the CDF of **(a)**. The data is shifted lower than the theoretical distribution.

4.1.3 Statistical Test

Figure 9b shows the CDF of Figure 9a. Visually, the fit is not as strong as the other two models. Table 3 shows the results from the KS test on the data - the test fails.

4.2 Finite Size Effect

The value $m = 3$ was analysed at $N = 10^i$ for $i = 2, 3, 4, 5$.

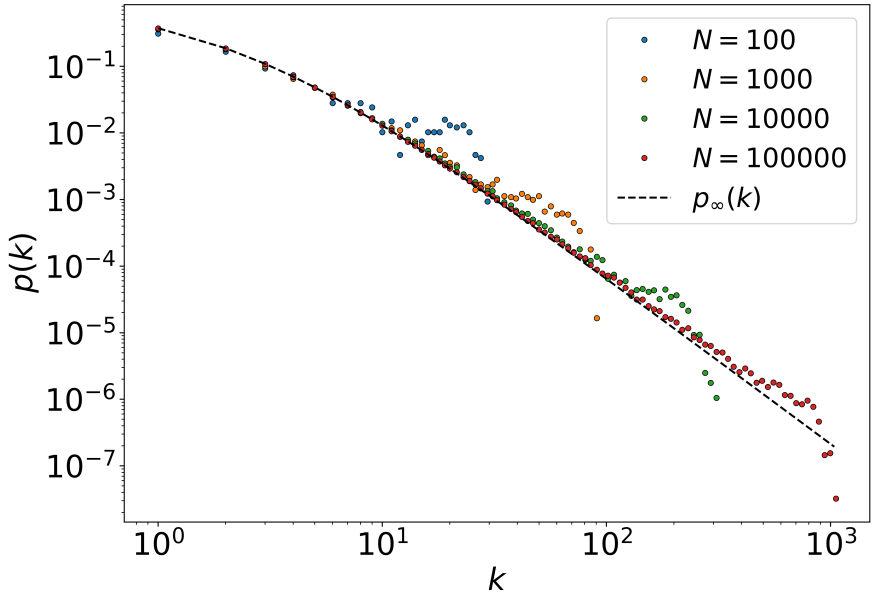


Figure 10: Degree distribution for $m = 13$ and $N = 10^i$ for $i = 2, 3, 4, 5$. A log-binning method was employed with $a = 1.06$. For each value of N, ten different graphs were produced to smooth the data. The dashed line shows the theoretical degree distribution.

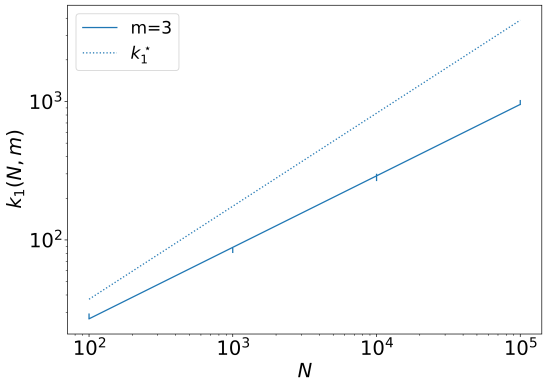
4.2.1 Derivation of the Largest Expected Degree

For the EV model, only $k_1^\star$ was considered. By substituting the probabilities of attachment into Equation 20 and solving the differential equation, the following expression for k_1 is produced:

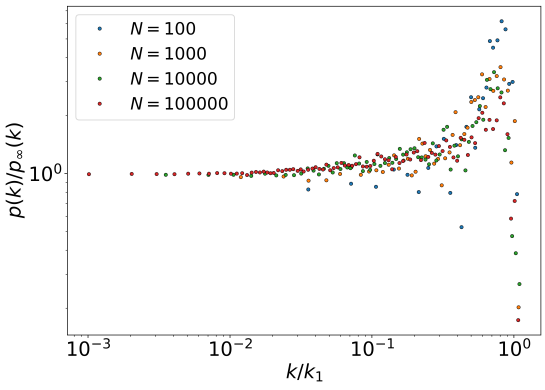
$$k_1^\star = \left(\frac{rm}{m-r} + m\right)\left(\frac{N}{m+1}\right)^{\frac{m-r}{m}} - \frac{rm}{m-r} \tag{41}$$

4.2.2 Data Collapse

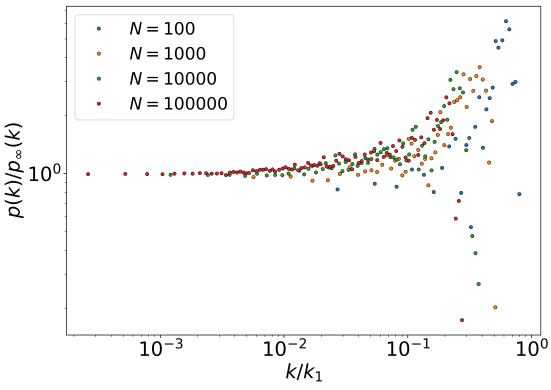
Figure 11 shows the k_1 values and the data collapse.



(a)



(b)



(c)

Figure 11: (a) is the k_1 plots for the simulated data and the theoretical value $k_1^\star$ using Equation 41 (dotted line). The solid line shows a linear fit on the empirical values of k_1. A data collapse was done on Figure 7 by scaling the y-axis by $p_\infty(k)$ and scaling the x-axis by the empirical value of k_1 in **(b)** and $k_1^\star$ in **(c)**. Visually, **(b)** provides a better data collapse.

5 Conclusion

The theoretical deviations of $p_\infty(k)$ are more accurate for PA and RA models. The data collapse using theoretical k_1 is the best for the PA model, but for all models the collapse using $k_1^\star$ is better. The degree distributions for all models exhibit a bump before the cut-off degree due to the finite size of the network.

References

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of Complex Networks". In: *Reviews of Modern Physics* 74.1 (2002), pp. 47–97. DOI: 10.1103/revmodphys.74.47.

[2] Tim Evans. *Part II: Networks*. 2021.

[3] Paul Secular. "Preferential and random attachment models of a complex network". In: *Imperial College London* (2015).