Network Project
A Growing Network Model

CID: 01854896

27th March 2023

Abstract: Derivations for the largest degree and the degree distribution, $p(k)$, for the Barabási–Albert (BA) model were found. These were simulated, and comparisons made both visually and tested using the Kolmogorov-Smirnov test statistic. Two other models were investigated: a random attachment model, and an existing vertices model, which uses a mixture of random and preferential attachment. The BA and existing vertices models were found to scale as $p(k) \propto k^{-3}$ and $p(k) \propto k^{-2.5}$ respectively, whereas random attachment exhibited no power-law scaling. All theoretical distributions for the degree distribution matched simulations, except for $m \geq 81$ for the existing vertices model. Finally, for all models, data collapses were produced to reveal the finite system size effects on the tail of the $p(k)$.

Word Count: 2434 words excluding font page, figure captions, table captions, acknowledgement and bibliography.

0 Introduction

We explore the Barabási–Albert (BA) model, demonstrating that preferential attachment leads to a fat-tailed degree distribution. Later, by comparing BA with the random attachment model, we aim to show that the fat-tailed distribution and scale-free behaviour cannot be recreated by random attachment alone. Finally, we explore the effects of an existing vertices model, which more accurately mimics networks in reality, such as the world wide web [1].

Definition

The BA model is a linear preferential attachment (PA) model, where the probability of an edge connecting to existing nodes, Π, scales linearly with the degree k of each node. This is given by

$$\Pi_{\text{pa}}\left(k,t\right) \propto k,$$

but $\sum_i \Pi(k_i,t) = 1$, leading to a normalisation

$$\Pi_{\text{pa}}\left(k,t\right) = \frac{k}{2E(t)}, \tag{1}$$

where $E(t)$ is the number of edges at time t, defined to increment by one per iteration. The normalisation is sensible, as each edge contributes +1 degree to a pair of nodes.

The number of nodes n with degree k at time t is given by the mean field master equation

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

with the boundary condition

$$\Pi(k,t)n(k,t) = 0 \text{ for } k < m, \tag{3}$$

where m is the number of edges added for each t.

The m edges were added simultaneously before updating Equation (1). Otherwise, this would render adding m edges as adding one edge m times for each t, making m redundant. However, this introduces the possibility of multi-edges if the same nodes are chosen, which is unaccounted for in Equation (2), since there are no $\Pi(k-2,t), \Pi(k-3,t)\ldots$ terms. On the other hand, if only simple graphs were allowed, then Equation (3) is incorrect, as the probability of a attaching to a node falls to 0 when it is already attached to at the same t. From numerical testing, the choice is irrelevant for $N \gg N_0$, but we allow multi-graphs for easier computation. Self-loops are forbidden, as the normalisation in Equation (1) excludes the node itself.

1 Phase 1: Pure Preferential Attachment Π_{pa}

1.1 Implementation

1.1.1 Numerical Implementation

A preferential attachment list was implemented. This list is first initialised with the existing nodes N_0, with the number of occurrences equal to k_i. Then, for every t,

1. m nodes are sampled with replacement from this list
2. Edges are attached from the new node to these sampled nodes
3. The preferential attachment list is extended by m occurrences of the new node and the list of sampled nodes

By sampling a node from this list, this directly corresponds to Equation (1), since the length of this list is the sum of the degrees for each node.

1.1.2 Initial Graph

To focus effects on preferential attachment only, the impact of the initial graph, $\mathcal{G}_0 = \mathcal{G}(N_0, E_0)$, was minimised. We wish to have the fewest number of nodes, and initialise the graph such that all nodes have equal probability of being chosen to prevent bias. To minimise multi-edges occurring, a *complete graph* was initialised, where each node is connected to every other node with one edge. Hence each node has degree $k = N_0 - 1$.

1.1.3 Type of Graph

Since multi-edges are allowed, the graph is highly likely to become a multi-graph after running for $N >> N_0$ and $m \geq 1$. The resulting graph is also undirected and unweighted. For $m = 1$, the resulting graph is a tree, since each node is connected by only one edge, creating a connected and acyclic graph. This was not simulated in the project.

1.1.4 Working Code

Several unit tests were written. Upon initialisation, we check that for $N_0 = 10$, we have $m_0 = 45$, and $k = N_0 - 1 = 9$ for all nodes. Furthermore, the BA model is initialised and driven until $N = 1000$. For each t, we assert the length of the preferential attachment list must be equal to both $2E(t)$ and $\sum_i k_i$.

1.1.5 Parameters

The parameters required were:

- N_0: the initial number of nodes. This depended on m, and is chosen to be $N_0 = m - 1$, the fewest nodes for a complete graph to satisfy Equation (3)
- N: the final number of nodes. This was at least $N = 100\,000$ in all sections to fulfill $N \gg N_0$, whilst keeping computational times short
- m: the number of edges per iteration. This was kept small when possible to reduce multi-edges created, unaccounted for in the theoretical derivations

1.2 Preferential Attachment Degree Distribution Theory

1.2.1 Theoretical Derivation

To begin, we substitute $\Pi = \Pi_{\text{pa}}$ into the Master Equation (1)

$$n(k, t+1) = n(k, t) + \frac{m(k-1)}{2E(t)} n(k-1, t) - \frac{mk}{2E(t)} n(k, t) + \delta_{km}.$$

By definition, the probability of a node with degree k is

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

Substituting in p for n yields,

$$N(t+1)p(k,t+1) = N(t)p(k,t) + \frac{m(k-1)}{2}\frac{N(t)}{E(t)}p(k,t) - \frac{mk}{2}\frac{N(t)}{E(t)}p(k,t) + \delta_{km}. \quad (4)$$

Consider the long time limit $t \to \infty$. At time t, we have

$$E(t) = E_0 + mt,$$
$$N(t) = N_0 + t,$$

where $E_0, N_0 < \infty$. Hence

$$\begin{aligned}\lim_{t\to\infty}\left(\frac{E(t)}{N(t)}\right) &= \lim_{t\to\infty}\left(\frac{E_0+mt}{N_0+t}\right)\\ &= \lim_{t\to\infty}\left(\frac{\frac{E_0}{t}+m}{\frac{N_0}{t}+1}\right)\\ &= m.\end{aligned}$$

This simplifies $N(t)/E(t)$ in Equation (4), assuming that for $t \to \infty$, p becomes stationary such that $p(k,t+1) = p(k,t) = p_\infty(k)$. This gives

$$N(t+1)p_\infty(k) = N(t)p_\infty + \frac{k-1}{2}p_\infty(k-1) - \frac{k}{2}p_\infty(k) + \delta_{km}.$$

Since only one node is added every iteration, $N(t+1) = N(t) + 1, \forall t \geq 0$, yielding

$$p_\infty(k) = \frac{1}{2}\left[(k-1)p_\infty(k-1) - kp_\infty(k)\right] + \delta_{km}. \quad (5)$$

For $k > m$

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

This has the solution

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

where $\Gamma(k)$ is the gamma function. We use the relation $\Gamma(k) = (k-1)!, \forall k \in \mathbb{Z}^+$ to find

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

To solve for A, we recognise that every new node comes with $k = m$, and Equation (3) implies that initial nodes with $k < m$ are never attached to. Both conditions combined means $n(k,t) = \text{const}$, but for $t \to \infty$, $N(t) \to \infty$, hence

$$p_\infty(k) = 0 \text{ for } k < m. \quad (6)$$

Substituting $k = m$ into Equation (5) and recognising $p(m-1) = 0$, we find

$$\frac{A}{m(m+1)(m+2)} = \frac{2}{m+2}$$
$$A = 2m(m+1).$$

Finally, we obtain

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \quad \text{for } k \geq m. \quad (7)$$

1.2.2 Theoretical Checks

Firstly, for $m < \infty$, Equation (7) shows

$$\lim_{k \to \infty} p(k) = k^{-3}. \tag{8}$$

This recovers the power-law relation $p(k) \propto k^{\gamma}$ for $k \gg 1$, with $\gamma = 3$ as found by the finite difference approximation in lectures.

From $p(k)$, the cumulative distribution function $p_<(k)$, representing the probability of randomly selecting a node i with degree $k_i < k$, is found by

$$\begin{aligned}
p_<(k) &= \sum_{k_i=m}^{k} p\left(k_i\right) \\
&= \sum_{k_i=m}^{k} \frac{2m(m+1)}{k_i(k_i+1)(k_i+2)} \\
&= m(m+1) \sum_{k_i=m}^{k} \left(\frac{1}{k_i} - \frac{2}{k_i+1} + \frac{1}{k_i+2} \right)
\end{aligned}$$

This is a telescoping series. The middle term cancels out contributions from the left and right terms in the series in the bulk of the sum.

$$\begin{aligned}
&= m(m+1) \left[\frac{1}{m} - \frac{1}{m+1} - \frac{1}{k+1} + \frac{1}{k+2} \right] \\
&= 1 - \frac{m^2+m}{(k+1)(k+2)}.
\end{aligned} \tag{9}$$

In the limit of $k \to \infty$

$$\lim_{k \to \infty} p_<(k) = 1,$$

confirming $p(k)$ is correctly normalised.

1.3 Preferential Attachment Degree Distribution Numerics

1.3.1 Fat-Tail

A fat-tailed distribution implies the decay in $p(k)$ is slower than an exponential. Since $k \in \mathbb{Z}^+$, $p(k)$ is also discrete. Thus, if linear binning method is used, there would be many bins in the tail with 0 counts, interspersed with occasional 1's over many orders of magnitude. This obscures the form of $p(k)$.

To circumvent this, logarithmic binning was used, where the increase in bin interval from the ith bin to the $i+1$th bin follows

$$\frac{b_{i+1}}{b_i} = \exp(\Delta) \quad \text{for } \Delta > 0.$$

The exponential increase in bin interval thus compensates for the diminishing statistics in the tail. This also results in horizontally equidistant data points in a log-log plot.

Furthermore, the statistical test used only $p_<(k)$, which is not required to be binned, removing the test's dependence on Δ.

1.3.2 Numerical Results

We simulate $m \in \{2, 4, 8, 16, 32, 64, 128\}$, where each m was repeated 50 times and $p(k)$ averaged. A plot of the degree distribution is shown in Figure 1. Best attempts were made to compute the error bars. Ideally, 50 runs is sufficient for each data point to be normally distributed under the Central Limit Theorem (CLT). However, bins in the tail have heavily skewed distributions, consisting zeros in most runs and only occasional non-zeros, and distributions with high skew converge to CLT slowly [2]. Thus, error bars on the tail should be taken qualitatively as their true spread is asymmetrical. Error bars are plotted as symmetric standard errors, equal to $\sigma/\sqrt{50}$ where σ is the standard deviation, as bins across runs are uncorrelated. This is useful as the sample mean is also normally distributed under CLT, broadly indicating how far each data point deviates if more runs were performed to estimate the true mean. Note that the logarithmic plot extends the lower error bar visually, breaking the symmetry.

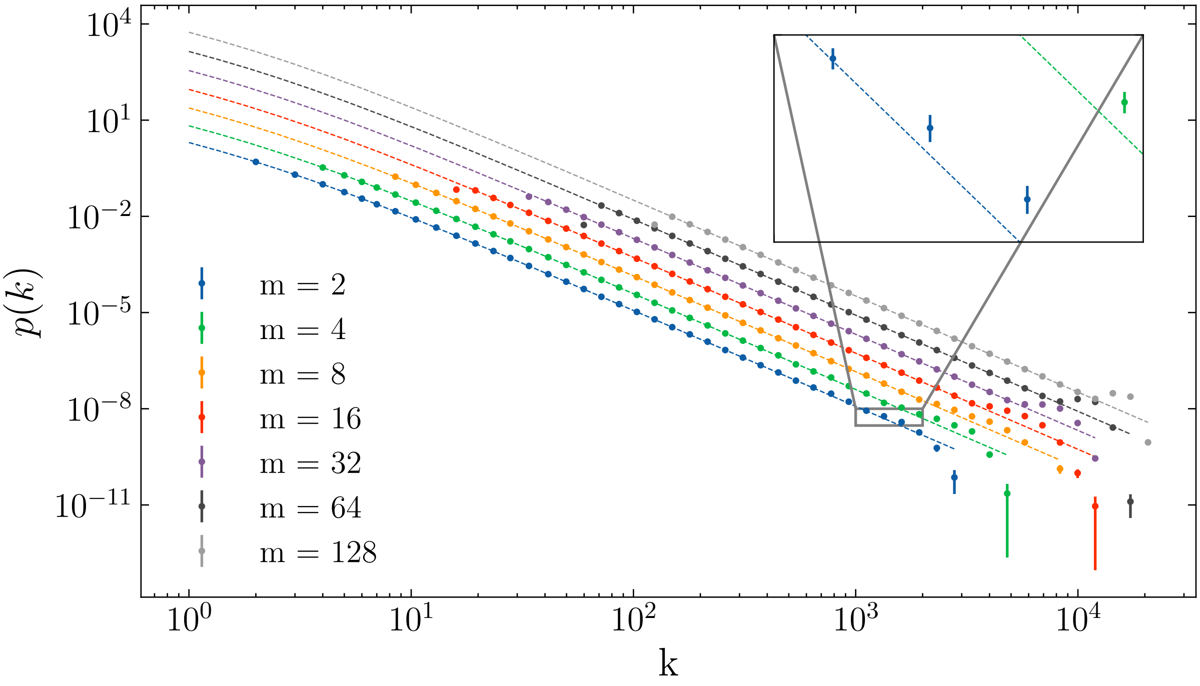

Figure 1: The degree distribution is plotted for $\Delta = 1.2$, up to $N = 1\,000\,000$, and data points are averaged over 50 runs. The standard error is plotted, though are too small to be visualised for most data points, as shown in the inset. Notice the first data point starts further out in the k-axis as m increases, as there are no nodes with $k < m$. The analytical shape, given by Equation (7), is marked with dashed lines. The first data point for each m is lower, which is an artifact from the log-binning. This is caused if the bin edges enclose $k = m$, causing fewer statistics for the first non-zero bin. This data point will hereafter be removed.

Assuming $p(k) \propto k^{-\gamma}$, to extract γ, a least squares regression was not performed on the log-log plot, as the estimator has significant bias for both $p(k)$ and $p_<(k)$ [3]. Instead, a Maximum Likelihood Estimator was used on the original set of node degrees. The estimator for discrete k has no closed-form solution, but can be approximated by

$$\hat{\gamma} \simeq 1 + n \left[\sum_{i=1}^{N} \ln \frac{k_i}{k_{\min} - \frac{1}{2}} \right]^{-1}, \tag{10}$$

where data points of $k < k_{\text{min}}$ are ignored, as Equation (7) suggests that the power-law approximation is worse for small k. k_{min} is found systematically by minimising the Kolmogorov-Smirnov distance, given by:

$$D = \sup_{x} \left| p_{<}^{th}(k) - p_{<}^{em}(k) \right| , \tag{11}$$

where sup is the supremum. This is the maximum difference between the empirical $p_{<}^{em}(k)$ and theoretical $p_{<}^{th}(k)$. Here, the theoretical $p_{<}^{th}(k)$ is for a power law. For $N \gg 1$, the standard error on the estimate is well approximated by

$$\sigma_{\hat{\gamma}} = \frac{\hat{\gamma} - 1}{\sqrt{S}} + \mathcal{O}\left(\frac{1}{S}\right) , \tag{12}$$

where S is the number of data points. Equations (10, 12) are conveniently implemented in the `powerlaw` library [4]. The values are shown in Table 1, which are all close to 3 within error. Furthermore, only one iteration was required for each m for this estimate to reach a high precision, as the unbinned data have many more data points.

Table 1: Power law exponent extracted using the `powerlaw` library. These should be close to 3, given by Equation(8). This is the case for all m within error.

m	k_{min}	$\hat{\gamma}$
2	37	3.03 ± 0.03
4	52	3.005 ± 0.023
8	101	3.016 ± 0.024
16	124	3.000 ± 0.015
32	116	3.003 ± 0.007
64	283	2.997 ± 0.009
128	375	2.996 ± 0.006

1.3.3 Statistics

A two sample Kolmogorov-Smirnov (KS) test was computed, which is a non-parametric, goodness of fit test, where parametric means the test is valid for any distribution. χ^2 was not used as it assumes Gaussian statistics in each bin, which is not true for the tail as argued above. Furthermore, it has a lower statistical power than the KS test [5].

This test computes the D as in Equation (11), but with different cumulative distribution functions. The first function is the empirical $p_{<}(k)$ from the simulation, and the second is a sample generated from the theoretical $p_{<}^{th}(k)$.

To construct the hypothesis test, we define the null hypothesis, H_0 to be that the sample follows the theoretical distribution, and the alternative hypothesis, H_1, as that it does not. To be the most precise, we should in theory sample from $p_{<}^{th}(k)$ over many iterations, such as using a Markov Chain Monte Carlo method, and perform KS tests on these samples. The fraction of accepted H_0 is then the p-value [3]. For simplicity in this project, $p_{<}^{th}(k)$ can be found by cumulatively summing $p^{th}(k)$. A threshold D_{crit} is then calculated by using

$$D_{\text{crit}} > \frac{1}{\sqrt{a}} \cdot \sqrt{- \ln\left(\frac{\alpha}{2}\right) \cdot \frac{1 + \frac{a}{b}}{2}},$$

where a, b are the lengths of the two samples, and α is the significance level, chosen to be $\alpha = 0.05$, giving $D_{\text{crit}} = 0.00192$ for all models in the project. Rather than comparing to a p-value, we can find the difference in D. This is because for $D > D_{\text{crit}}$, H_0 is rejected at a significance level α. $D_{\text{crit}} - D$ was calculated in Table 2, where positive values corresponds to accepting H_0. From this, we see that all m values are accepted, though this value decreases for increasing m. This is because for increasing m, more multi-edges are added, and the impact of $\mathcal{G}_0$ is greater, both unaccounted for theoretically.

Table 2: KS test values to 2 significant figures, where a positive $D_{\text{crit}} - D$ implies the data matches the theoretical distribution. All m passes the test, though the agreement becomes increasingly worse for larger m.

m	$D_{\text{crit}} - D$
2	0.0014
4	0.0016
8	0.0015
16	0.0015
32	0.0013
64	0.0010
128	0.00042

1.4 Preferential Attachment Largest Degree and Data Collapse

1.4.1 Largest Degree Theory

An equivalent definition for k_1 is that, within N nodes, only one node is expected to be found with $k > k_1$. Hence

$$1 = N \sum_{k=k_1}^{\infty} p(k)$$

recognise this as

$$\begin{aligned} &= N\left(1 - \sum_{k=m}^{k_1-1} p(k)\right) \\ &= N\left(1 - p_<(k_1 - 1)\right). \end{aligned}$$

Note that the $k_1 - 1$ argument in $p_<(k_1 - 1)$ arises as $k \in \mathbb{Z}^+$. Substituting in Equation (9) gives

$$1 = \frac{N(m^2 + m)}{k_1(k_1 + 1)},$$

resulting in a quadratic in k_1

$$0 = k_1^2 + k_1 - Nm(m + 1).$$

Only the positive square root is taken such that $k_1 > 0$, giving

$$k_1 = -\frac{1}{2} + \frac{\sqrt{1 + 4Nm(m+1)}}{2}. \tag{13}$$

The model only runs for $m \geq 1$ and $N \geq 1$, which implies $k_1 > 0$. Thus this solution is always physical. However, k_1 found empirically is always an integer, whereas this may yield non-integers for combinations of N and m.

1.4.2 Numerical Results for Largest Degree

We require $m \ll N$, such that the random network is run for iterations for effects of $\mathcal{G}_0$ to be small. Hence, $m = 2$ was chosen, avoiding $m = 1$ as m becomes a redundant parameter. This was run for $N \in \{100, 300, 1000, 3000, 10\,000, 30\,000, 100\,000\}$ to have sufficient data points and span 4 orders of magnitude.

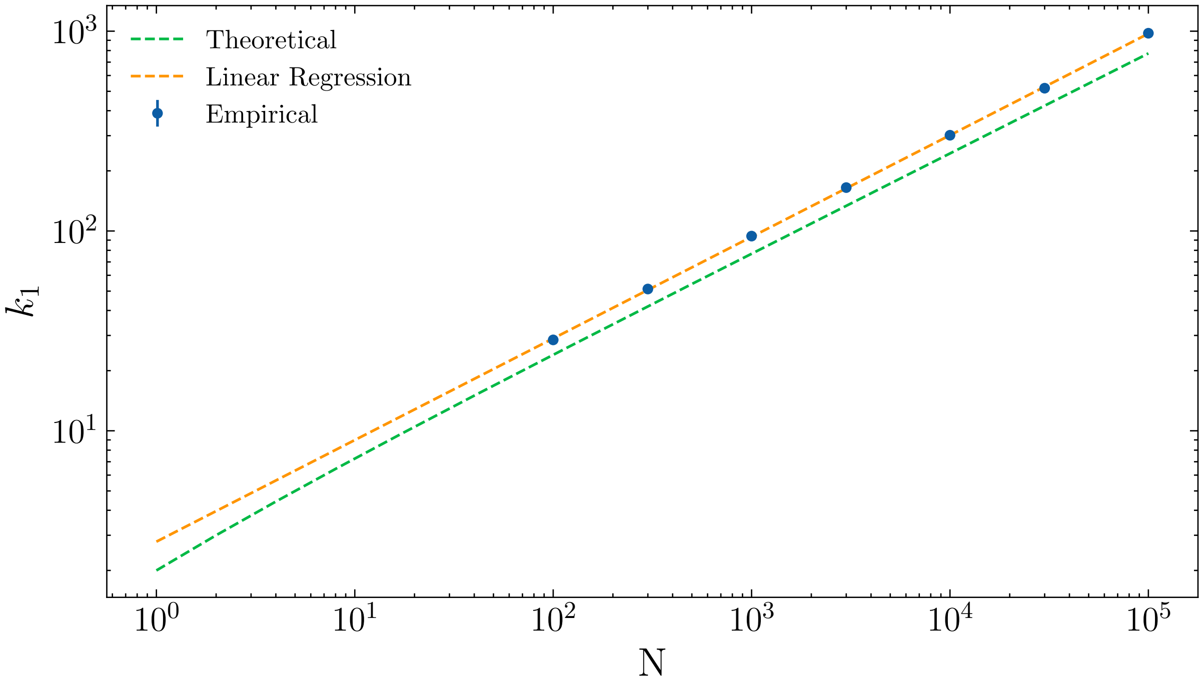

Figure 2: Error bars are the standard error, too small to be shown visually. The green dashed line, indicating Equation 13, is curved for small N, but rapidly tends to a straight line, since $k_1 \propto \sqrt{N}$ for $N \gg 1$. The orange dashed line is a linear regression performed on the data to find the fitted scaling exponent.

From Figure 2, there appears to be a systematic deviation, where all data points lie above the green line. Fitting a least squares linear regression with inverse variance weighting on Figure 2, the gradient yields 0.508 ± 0.002, where the error quoted is the standard deviation found by taking the square root of the relevant diagonal element in the covariance matrix. The slight deviation is due to the incorrect scaling in m in Equation (13) from numerical testing.

1.4.3 Data Collapse

To perform a data collapse, the k-axis was first divided by k_1 for each N. By using the k_1 found numerically, this is by definition the largest degree, and thus this division forces the largest k to be at the same horizontal position for each N. To collapse the vertical direction, we divide the log-binned $p(k)$ by the theoretical $p(k)$, where any deviations from theory would cause the vertical position to deviate from unity. The data collapse is shown in Figure 3. All points decay rapidly near $k/k_1 = 1$, as the finite system size forces $\sum_k^N p(k) = 1$, rather than $\sum_k^\infty p(k) = 1$. This also causes p to 'pile-up' just before decaying, exceeding $p^{th}(k)$ near the decay and is shown as a slight rise in the graph.

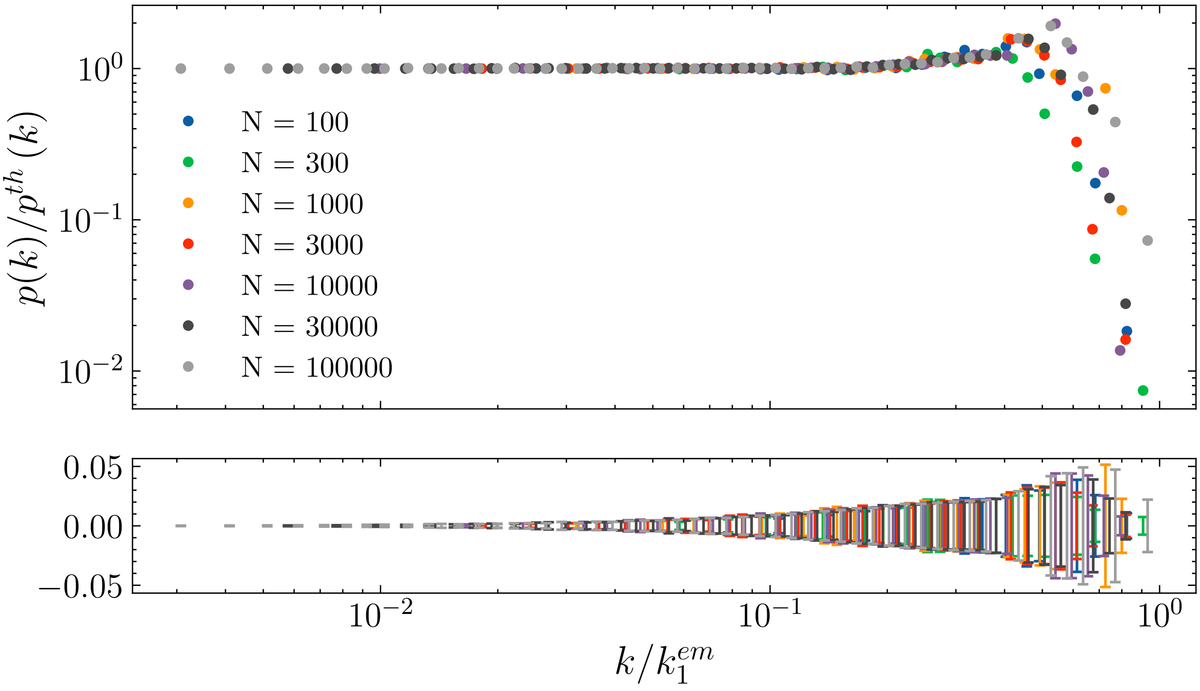

Figure 3: the k axis is divided by k_1 found empirically, denoted k_1^{em}, and $p(k)$ is divided by Equation (7), denoted $p^{th}(k)$, using $\Delta = 1.2$. Since data points are overlaid, error bars for subsequent data collapses are plotted below in linear scale for clearer visualisation. There are more data points for higher N, inferred from the many grey dots, which is sensible because k extend further out, filling out more non-zero bins in the logarithmic binning.

2 Phase 2: Pure Random Attachment Π_{rnd}

2.1 Random Attachment Theoretical Derivations

2.1.1 Degree Distribution Theory

For random attachment, we have

$$\Pi_{\text{rnd}}(k,t) = \frac{1}{N(t)}.$$

Substituting this into the Equation (2), and taking $t \to \infty$

$$p_\infty(k)\left[N(t+1) - N(t) + m\right] = m p_\infty(k-1) + \delta_{km},$$

Again, we apply $N(t+1) = N(t) + 1$, giving

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

For $k > m$, we have the recursive relation

$$p_\infty(k) = \frac{m}{1+m} p_\infty(k-1) = \left(\frac{m}{1+m}\right)^2 p_\infty(k-2) = \cdots = \left(\frac{m}{1+m}\right)^{k-m} p_\infty(m).$$

By inspection, this is in the form

$$p_\infty(k) = \frac{1}{1+m}\left(\frac{m}{1+m}\right)^{k-m} \qquad \text{for } k < m.$$

Furthermore, Equation (3) also applies here, giving $p_\infty(k) = 0$ for $k > m$. For $k = m$, we have

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

but $p_\infty(m-1) = 0$. This yields,

$$p_\infty(k) = \frac{1}{1+m} \qquad \text{for } k = m.$$

Putting it together, we find

$$p_\infty(k) = \frac{1}{1+m}\left(\frac{m}{1+m}\right)^{k-m} \qquad \text{for } k \geq m. \tag{14}$$

To verify, we compute $p_<(k)$,

$$\begin{aligned} p_<(k) &= \frac{1}{1+m}\sum_{k_i=m}^{k}\left(\frac{m}{1+m}\right)^{k_i-m} \\ &= \frac{1}{1+m}\sum_{n=0}^{k-m}\left(\frac{m}{1+m}\right)^{n}, \end{aligned}$$

where we relabelled $k_i \to n$. Since $r = |\frac{m}{1+m}| < 1$, we can apply both the finite and infinite geometric series formulae. The finite series is used here

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

Now use the infinite geometric sum to evaluate $k \to \infty$

$$\lim_{k\to\infty} p_<(k) = \frac{1}{1+m}\frac{1}{1-\frac{m}{1+m}} = 1,$$

indicating the normalisation is correct.

2.1.2 Largest Degree Theory

From

$$\begin{aligned} 1 &= N\left(1 - p_<(k_1 - 1)\right) \\ &= N\left(\frac{m}{m+1}\right)^{k_1-m}, \end{aligned}$$

and then taking the logarithms on both sides,

$$0 = \ln N + (k_1 - m) \ln \left(\frac{m}{m+1} \right),$$

we rearrange for k_1,

$$k_1 = m - \frac{\ln N}{\ln m - \ln(m+1)}. \tag{15}$$

2.2 Random Attachment Numerical Results

2.2.1 Degree Distribution Numerical Results

The $p(k)$ distribution is shown in Figure 4. The theoretical distribution is given by Equation (14), and indeed visually, there appears to be excellent agreement for all data points except the tail. From comparison to Figure 1, we see that $p(k)$ in this plot decay faster than preferential attachment, since data points do not reach high k values for the same N.

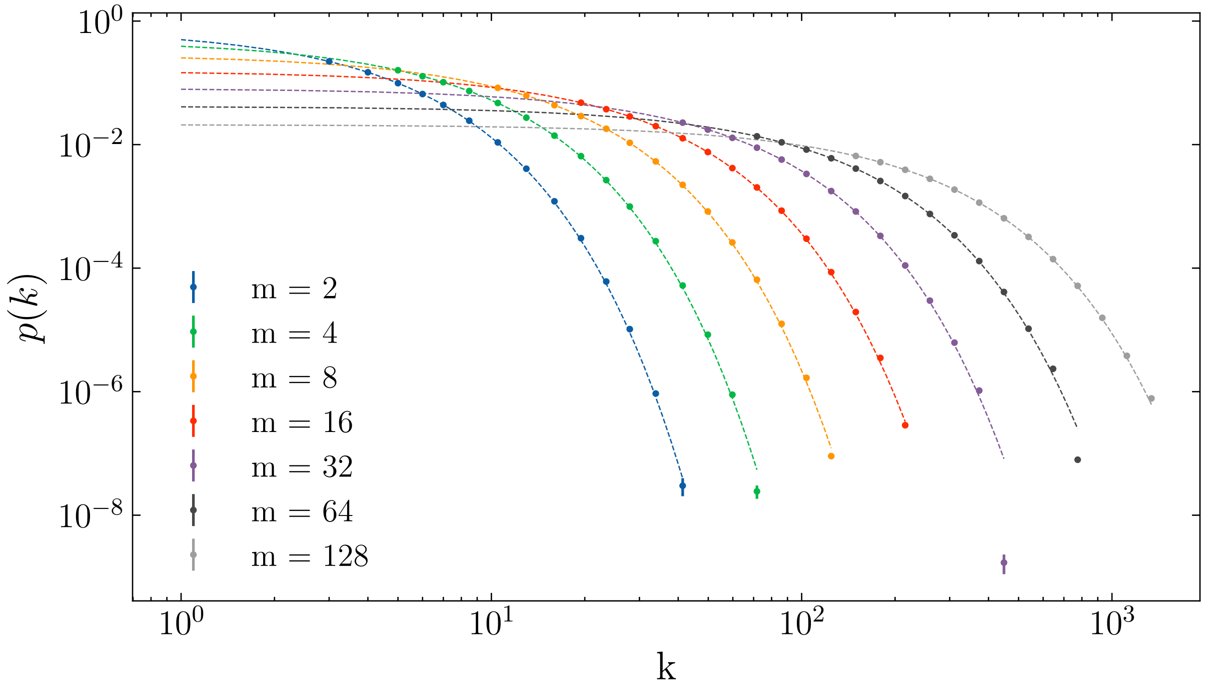

Figure 4: Each m was run for 50 times until $N = 1\,000\,000$, log-binned with $\Delta = 1.2$ and then averaged, where error bars are the standard error. The dashed lines are the theoretical prediction given by Equation (14). Notice that no region in this graph is a straight line. This suggests that random attachment does not yield a fat-tailed $p(k)$.

To test the agreement between theory and simulation, we perform the same KS test, shown in Table 3. H_0 is accepted for all m, where $D - D_{crit}$ values are greater, suggesting a better agreement compared to the BA model.

2.2.2 Largest Degree Numerical Results

Selecting $m = 2$ for the same reason as PA, Figure 5 shows k_1 averaged over 1000 instances of the model. The theoretical distribution fits the data better compared to PA. However

Table 3: KS values for random attachment to 2 significant figures. Positive values indicate the data fits the theoretical distribution at a significance level $\alpha = 0.05$

m	$D_{crit} - D$
2	0.0017
4	0.0016
8	0.0015
16	0.0015
32	0.0015
64	0.0016
128	0.0016

data points appear to be slightly under the theoretical distribution for low N. This can be fully explained due to the worse approximation in $N \gg N_0$ for low N. Furthermore, the graph does not appear linear in any region within the figure, since Equation (15) implies $k_1 \propto \ln(N)$ for $N \gg 1$, which is not a power-law.

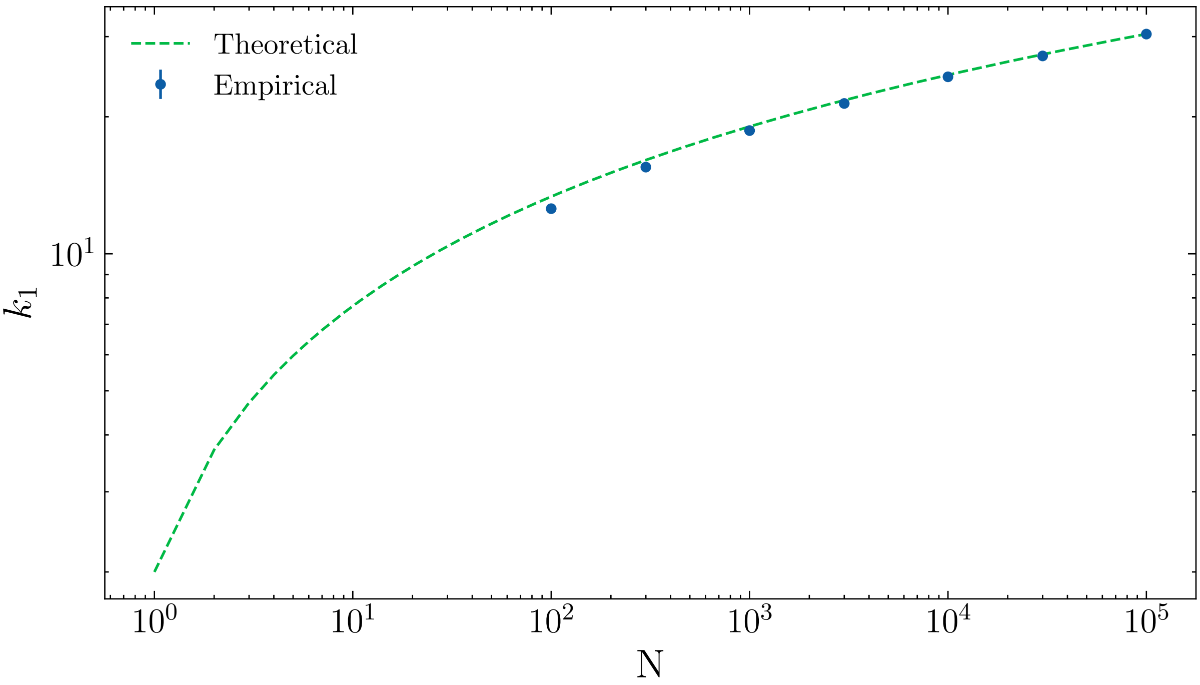

Figure 5: k_1 was found for $m = 2$, up to $N = 100\,000$, for 1000 runs and then averaged, such that the standard error is negligible. Data points do not agree with the theoretical line, given by Equation (15), for low N, due to $N \gg N_0$ approximation becoming worse.

A data collapse is shown in Figure 6. This collapse does not have a prominent bump before decaying away, which can be explained since the decay is quicker for random attachment. Hence, the finite-size effect is less prominent as $\sum_k p(k)$, or roughly the area under the dashed line in Figure 4, is closer to one upon reaching N.

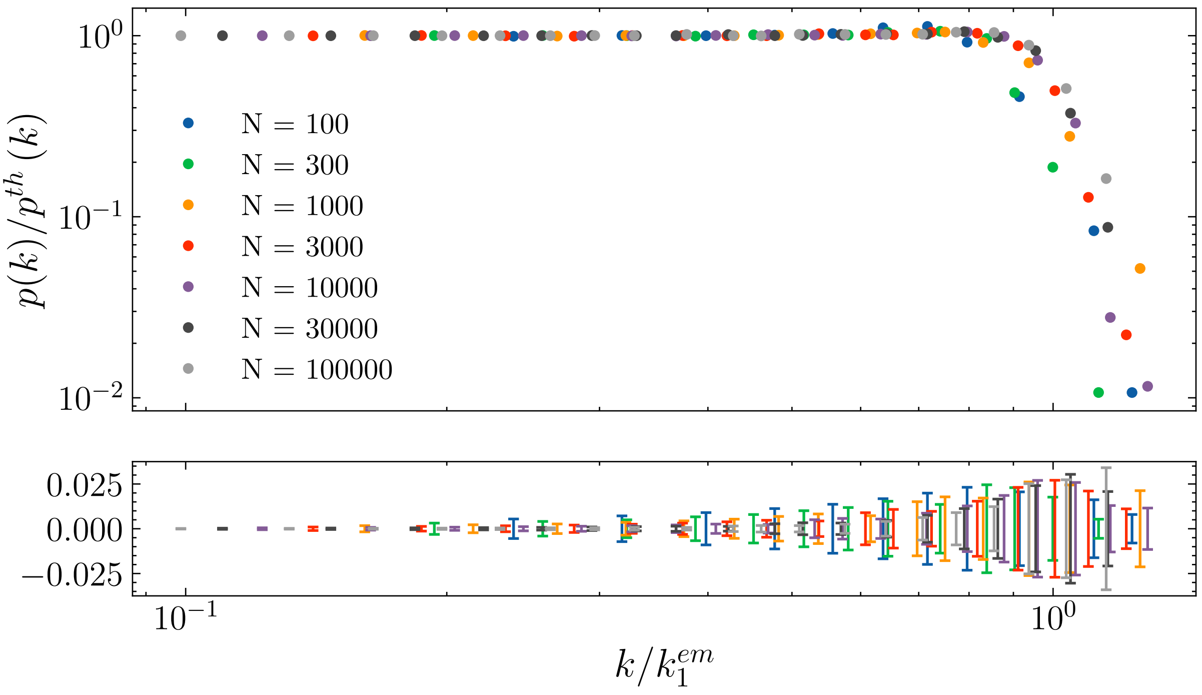

Figure 6: Data collapse produced for $\Delta = 1.2$. Each data point was averaged over 1000 runs and the standard error is plotted on the bottom in linear scale. There is no visible bump compared to BA.

3 Phase 3: Existing Vertices Model

3.1 Existing Vertices Model Theoretical Derivations

For m edges, the existing vertices model divides this into two attachment probabilities, where r edges follow Π_1 and $(m-r)$ follow Π_2, giving

$$m\Pi(k,t) = r\Pi_1 + (m-r)\Pi_2.$$

We consider $\Pi_1 = \Pi_{\text{rnd}}, \Pi_2 = \Pi_{\text{pa}}$ specifically. Since the preferential attachment is selecting two existing nodes in our network, each new edge added has twice the probability to connect to a node with degree $k-1$ or k, and so Π_{pa} is multiplied by 2. This results in

$$m\Pi(k,t) = \frac{r}{N(t)} + \frac{m-r}{E(t)}.$$

Each new node comes with r edges attached. Therefore the Kronecker delta reads δ_{kr}. Substituting this into the Master Equation (2),

$$\begin{aligned} n(k,t+1) = n(k,t) + \left(\frac{r}{N(t)} + \frac{(m-r)(k-1)}{E(t)}\right) n(k-1,t) \\ - \left(\frac{r}{N(t)} + \frac{k(m-r)}{E(t)}\right) n(k,t) + \delta_{kr}. \end{aligned}$$

Substituting $n(k,t)$ for $p(k,t)$, and using $N(t+1) = N(t)+1$,

$$\begin{aligned} (N(t)+1)\, p(k,t+1) = N(t)p(k,t) + \left(r + (m-r)(k-1)\frac{N(t)}{E(t)}\right) p(k-1,t) \\ - \left(r + k(m-r)\frac{N(t)}{E(t)}\right) p(k,t) + \delta_{kr}. \end{aligned}$$

For $t \to \infty$, we have $E(t)/N(t) \to m$. Again assume a stationary $p_\infty(k)$,

$$\begin{aligned}(N(t)+1)\,p_\infty(k) = N(t)p_\infty(k) &+ \left(r + \frac{(m-r)(k-1)}{m}\right) p_\infty(k-1) \\ &- \left(r + \frac{k(m-r)}{m}\right) p_\infty(k) + \delta_{kr}.\end{aligned} \tag{16}$$

For $k > r$,

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

yielding

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

We then substitute $k = r$ into Equation (16), recalling $p_\infty(r-1) = 0$ to determine A. Putting it together, we obtain

$$p_\infty(k) = \frac{m}{m + mr + (m-r)} \frac{\Gamma\left(1 + r + \frac{m+mr}{m-r}\right)}{\Gamma\left(r + \frac{mr}{m-r}\right)} \frac{\Gamma\left(k + \frac{mr}{m-r}\right)}{\Gamma\left(k + 1 + \frac{m+mr}{m-r}\right)} \quad \text{for } k \geq r.$$

To check our simulation, we seek the leading order term for $k \to \infty$. The asymptotic expansion of a ratio of gamma functions is known analytically in literature by using Stirling's approximation [6]:

$$\begin{aligned}\lim_{k\to\infty} p_\infty(k) &= \lim_{k\to\infty} A\frac{\Gamma\left(k + \frac{mr}{m-r}\right)}{\Gamma\left(k + 1 + \frac{m+mr}{m-r}\right)} \\ &= k^{\left(\frac{mr}{m-r} - 1 - \frac{m+mr}{m-r}\right)} \\ &= k^{-\left(1 + \frac{m}{m-r}\right)}.\end{aligned}$$

Notice for $r = m$, where we have pure random attachment, the denominator in the exponent becomes 0 and the leading order expression is invalid. This is consistent with our earlier findings where random attachment is not fat-tailed.

For the special case $r = m/3$,

$$p_\infty(k) = \frac{9}{9 + 5m} \frac{\Gamma\left(\frac{5}{2} + \frac{5m}{6}\right)}{\Gamma\left(\frac{5m}{6}\right)} \frac{\Gamma\left(k + \frac{m}{2}\right)}{\Gamma\left(k + \frac{5}{2} + \frac{m}{2}\right)} \quad \text{for } k \geq r, \tag{17}$$

$$\lim_{k\to\infty} p_\infty(k) = k^{-\frac{5}{2}} \quad \text{for } k \geq r. \tag{18}$$

3.2 Existing Vertices Model Numerical Results

The numerical results are plotted in Figure 7 for $m \in \{3, 9, 27, 81, 243\}$, ensuring $m \equiv 1 \pmod 3$. The dashed lines, indicating Equation (14), appears to match data points well for both the head and the middle of the distribution. Furthermore, Equation 18 implies that the degree distribution is well approximately by $p(k) \propto k^{-\gamma}$ for $k \gg 1$, where theoretically, $\gamma = 2.5$. An estimate, $\hat{\gamma}$, was also made for each m in simulation, with results shown in Table 4. These values match well for small m, but become smaller for increasing m.

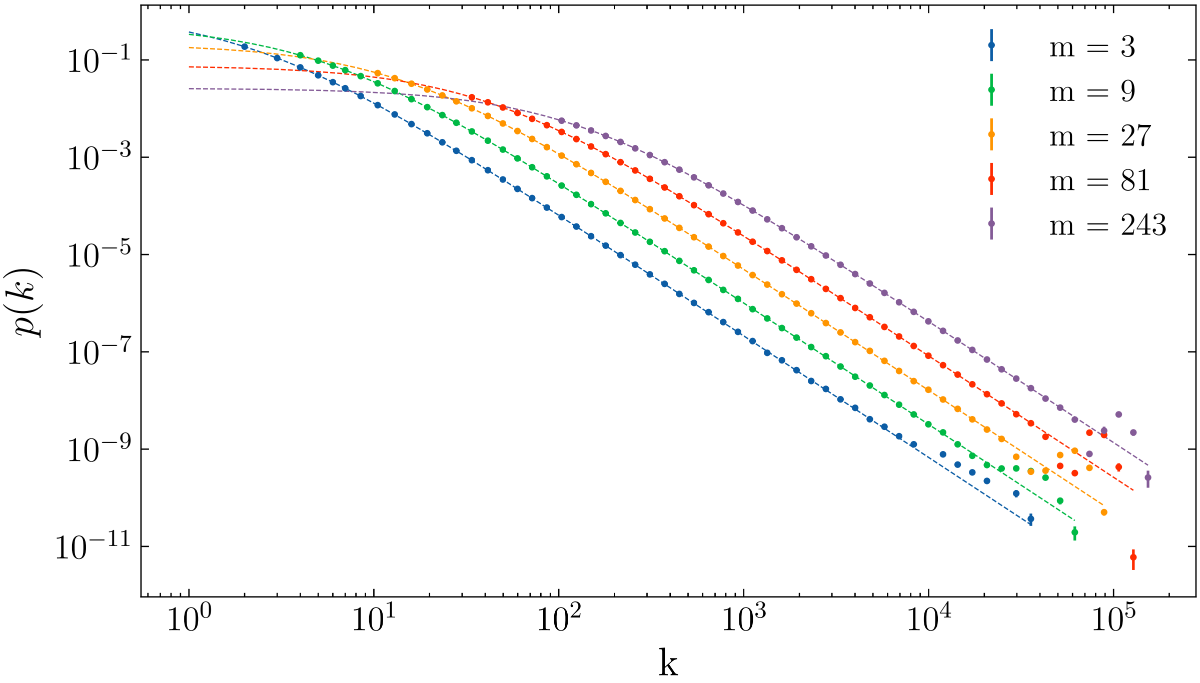

Figure 7: The degree distribution for the mixed attachment model, where $\Delta = 1.2$, $N = 1\,000\,000$, and each data point is averaged over 50 different instances of the model. Dashed lines are the analytical $p(k)$ given by Equation (17) for each m. Despite having a random attachment element, the graph is yet again linear in the log-log plot, implying a fat-tailed distribution. If the existing vertices model accurately reflects the world wide web, then we can conclude the internet must be fat-tailed.

Table 4: $\hat{\gamma}$ extracted from Figure 7 using `powerlaw`. These should be close to 2.5, found from Equation (5)

m	k_{min}	$\hat{\gamma}$
3	75	2.497 ± 0.019
9	216	2.483 ± 0.018
27	268	2.46 ± 0.01
81	1078	2.451 ± 0.012
243	1715	2.422 ± 0.007

This is because $k \gg m$ is required to approximate the asymptotic k behaviour, and this approximation becomes progressively worse for increasing m.

Beyond finding γ, we compute the KS test, given in Table 5. This test is actually rejected for $m \geq 81$, despite visually being a good fit. This is explained as γ is smaller compared to preferential attachment, leading to a slower decay in $p(k)$. Hence, the finite system size truncates $p(k)$ more abruptly, and the empirical and theoretical $p_<(k)$ has a greater mismatch.

For the data collapse, shown in Figure 8, we choose $m = 3$, being the smallest non-zero m that is divisible by 3. Compared to previous data collapses, we see a greater bump, again due to the truncation in $p(k)$.

Table 5: KS test run for each m to 2 significant figures. For $m \geq 81$, the KS test rejects the theoretical distribution, theorised due to increased contribution from finite system size effects.

m	$D_{crit} - D$
3	0.0016
9	0.0013
27	0.00081
81	-0.00026
243	-0.0024

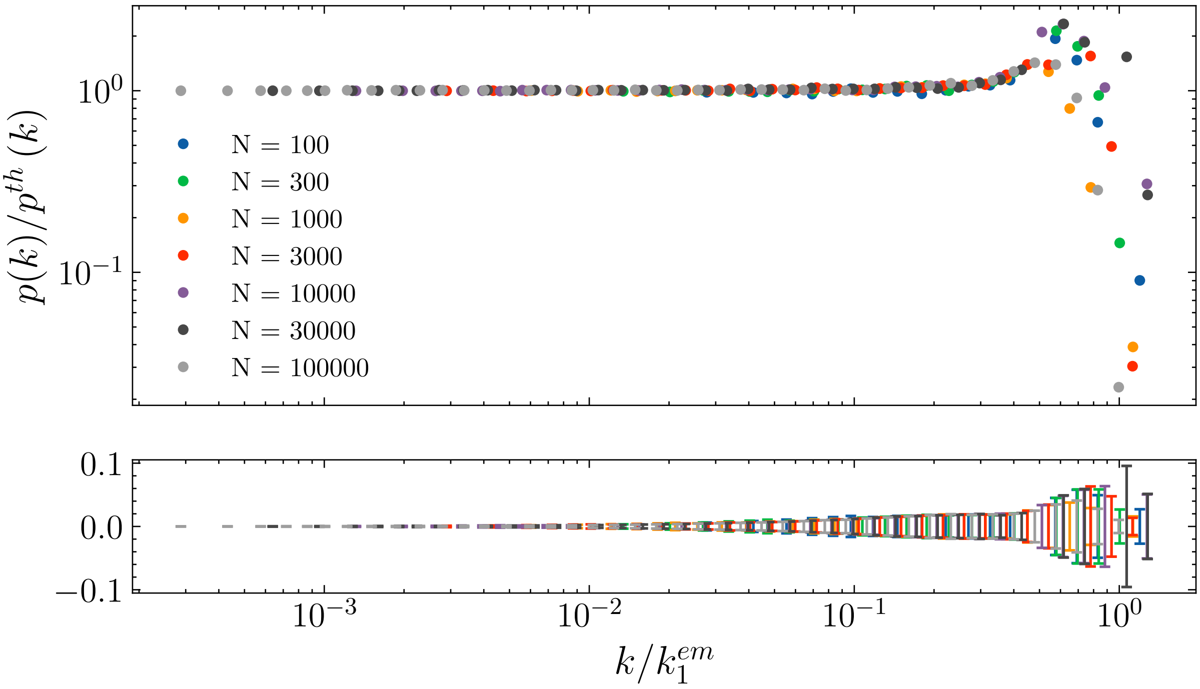

Figure 8: Data collapse for existing vertices for $\Delta = 1.2$, averaged over 1000 runs. Standard error is plotted in the linear scale. The bump is more noticeable than the collapse for random attachment.

4 Conclusions

Theoretical derivations and numerical simulations were compared, being largely consistent except when assumptions, such as the finite system size, becomes worse. For a perfect match between simulation and theoretical derivation, it is worth investigating the linearised chord diagram variant of the BA model, which gives a complete definition for the BA model. The mathematical derivations are beyond the scope [7].

References

[1] Albert-László Barabási, *Network Science by Albert-László Barabási.* [Online]. Available: `http://networksciencebook.com/` (visited on 03/09/2023).

[2] Mark E. Irwin, *Statistics 110*, Duke University, 2006. [Online]. Available: `http://www2.stat.duke.edu/~sayan/230/2017/Section53.pdf` (visited on 03/27/2023).

[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," en, *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 2009, ISSN: 0036-1445, 1095-7200. DOI: `10.1137/070710111`. [Online]. Available: `http://epubs.siam.org/doi/10.1137/070710111` (visited on 03/27/2023).

[4] J. Alstott, E. Bullmore, and D. Plenz, "Powerlaw: A Python package for analysis of heavy-tailed distributions," *PLoS ONE*, vol. 9, no. 1, e85777, Jan. 2014, arXiv:1305.0215 [physics], ISSN: 1932-6203. DOI: `10.1371/journal.pone.0085777`. [Online]. Available: `http://arxiv.org/abs/1305.0215` (visited on 03/08/2023).

[5] S. D. Horn, "Goodness-of-Fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale," *Biometrics*, vol. 33, no. 1, pp. 237–247, 1977, Publisher: [Wiley, International Biometric Society], ISSN: 0006-341X. DOI: `10.2307/2529319`. [Online]. Available: `https://www.jstor.org/stable/2529319` (visited on 03/27/2023).

[6] A. Erdélyi and F. G. Tricomi, "The asymptotic expansion of a ratio of gamma functions.," *Pacific Journal of Mathematics*, vol. 1, no. 1, pp. 133–142, Jan. 1951, Publisher: Pacific Journal of Mathematics, A Non-profit Corporation, ISSN: 0030-8730. [Online]. Available: `https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-1/issue-1/The-asymptotic-expansion-of-a-ratio-of-gamma-functions/pjm/1102613160.full` (visited on 03/26/2023).

[7] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, "The degree sequence of a scale-free random graph process," *Random Structures and Algorithms*, vol. 18, no. 3, pp. 279–290, May 2001, ISSN: 1042-9832. DOI: `10.1002/rsa.1009`. [Online]. Available: `http://www.scopus.com/inward/record.url?scp=0035625228&partnerID=8YFLogxK` (visited on 03/09/2023).