Network Project
A Growing Network Model

CID: 01701209

26th March 2022

Abstract: In this project, we implemented a computer simulation to model the Barabasi-Albert network using Python. We tested and compared three varied models, including the preferential attachment model, the random attachment model and the existing vertices model. Theoretical results for the degree distribution and its largest degree were documented, to which our data were compared. The reduced chi-square statistics were employed to quantify the goodness of the fit and to conduct further hypothesis testing. In addition, the scaling factor of the largest degree on the system size was determined to a 2% and a 7% error respectively.

Word count: 2493 (discounting the front page)

0 Introduction

The Barabasi-Albert model (BA model) undertakes a preferential attachment process, meaning more connected vertices in the network are more likely to experience further connections with new vertices. This behaviour gives rise to a fat-tailed distribution which can be explained by a "richer get richer" principle - a legitimate mechanism that is seen in our everyday life from wealth, social influence to modern politics.

Definition

We define the BA model as follows:

1. Create an initial graph G_0 at time $t = t_0$.
2. Increment time by one unit $t \to t + 1$.
3. Add a new vertex.
4. Add m new vertices between the new vertex and existing vertices with a probability Π. In the pure preferential attachment case, $\Pi = \Pi_{pa}$ which is proportional to the total degree k of existing vertices.
5. Repeat steps 2-4 until a specified time $t = t_{max}$ is reached.

Note unless otherwise specified, we do not allow: multiple attachments between a pair of vertices, self-loops, the total degree k of existing vertices to be smaller than m.

1 Phase 1: Pure Preferential Attachment Π_{pa}

1.1 Implementation

1.1.1 Numerical Implementation

An initial graph is first created using NetworkX. To enable preferential attachment, we also initialise a "preferential attachment" list (PA list) which stores all vertices in existing edges. For example, if there are three existing edges (0, 1), (0, 2) and (1, 2) to start, the PA list will be [0, 1, 0, 2, 1, 2].

A loop is used to iterate the network by one unit in time $t \to t + 1$ until a total number of vertices $t = t_{max}$ is reached. Inside each iteration, we add a new vertex to the system and form m edges between this new node and existing vertices. To do this, an existing vertex from the PA list is randomly chosen. Note, since vertices with higher degrees have more occurrences inside the list, the attachment probability Π_{pa} to an existing vertex will be proportional to its degree k. While a new edge is formed, both the chosen and new vertices are appended to the PA list. Extra caution was taken to avoid self-loops and duplicate attachments. For instance, if the chosen existing vertex is the same as the new vertex or an edge has already formed before, a new random vertex will be chosen again.

After all iterations are completed, elements inside the PA list are counted by computing a histogram of the dataset, where the bin centres correspond to our vertices and the values of the histogram correspond to the degrees of each vertex.

1.1.2 Initial Graph

A complete graph with $m+1$ vertices was chosen as the initial graph. An important property is that every vertex has a maximum number of edges across the entire network. This removes any bias towards a specific vertex when new edges are formed initially. Having initial vertices of $N(t_0) = m+1$ ensures that every vertex contains m degrees which is a requirement under the definition of the BA model: $k \geq m$ at all times. Furthermore, limiting t_0 to a small number also minimises the influence of the initial conditions.

1.1.3 Type of Graph

A simple graph is produced. It is undirected, unweighted and contains no graph loops or multiple edges. We made this choice because it is straightforward to implement computationally. Although most networks are directed in reality, it is important to note that an undirected graph still follows the original mathematical formulation of Barabasi and Albert [1] and we do not expect any difference in the final degree distribution. For $t \to \infty$, the graph also grows into a sparse network according to Eq (1) and Eq (2), both satisfying the required conditions: total edges $E(t) \sim O(N^1)$ and average degree $\langle k \rangle \sim O(1)$ [1].

$$E(t) = mN(t), \qquad (1) \qquad\qquad \langle k \rangle = 2m. \qquad (2)$$

1.1.4 Working Code

As an initial check, we drew the graph for small m ($m = 2$) using NetworkX over a finite time interval and inspected if any condition was violated. In addition, we also implemented a numerical test by comparing the measured $\langle k \rangle$ for a range of m against their theoretical values predicted by Eq (2). In the long-time limit $N = 10^5$:

Table 1: Measurements of $\langle k \rangle$ for $m \in \{2, 4, 8, 16, 32, 64, 128\}$

m	$\langle k \rangle$
2	3.99 ± 0.01
4	7.99 ± 0.01
8	15.99 ± 0.02
16	31.98 ± 0.02
32	63.97 ± 0.04
64	127.92 ± 0.12
128	255.79 ± 0.23

Note uncertainties were estimated from the standard deviation of respective degree distribution.

1.1.5 Parameters

Our programme requires three parameters, including the number of edges added per unit time m, the final number of vertices N and the number of numerical runs. To study the degree distribution of preferential attachment, we chose $m \in \{2, 4, 8, 16, 32, 64, 128\}$ as they are all

powers of two, which will result in linearly spaced plots on a logarithmic scale. N was chosen to be 10^5, a large enough number to approximate the general requirement of a long-time limit $t \to \infty$ and hence to minimise the finite-size effect. Across all phases, we averaged our data over 50 numerical runs to reduce the effect of statistical fluctuations.

1.2 Preferential Attachment Degree Distribution Theory

1.2.1 Theoretical Derivation

The master equation describes how the degree distribution of a network changes as it evolves in time:

$$n(k,t+1) = n(k,t) + m\Pi(k-1,t)n(k-1,t) - m\Pi(k,t)n(k,t) + \delta_{k,m}, \tag{3}$$

where $n(k,t)$ is the number of vertices at time t with degree k, $\Pi(k,t)$ is the probability that a vertex with degree k is connected to and $\delta_{k,m}$ represents the new vertex when $k = m$. Defining $p(k,t)$, the probability of finding a vertex with degree k in the system, we have $n(k,t) = p(k,t)N(t)$. Rearranging Eq (3) gives:

$$N(t+1)p(k,t+1) = N(t)p(k,t) + m\Pi(k-1,t)N(t)p(k-1,t) - m\Pi(k,t)N(t)p(k,t) + \delta_{k,m}. \tag{4}$$

Since $N(t=1) = N(t)+1$, the $N(t)p(k,t)$ term cancels. For preferential attachment:

$$\Pi(k,t) = \Pi_{pa} = \frac{k}{2E(t)} \approx \frac{k}{2mN(t)}. \tag{5}$$

Substituting Eq (5) into (4), we obtain the difference equation:

$$p(k,t+1) = \frac{1}{2}[\,(k-1)p(k-1,t) - kp(k,t)] + \delta_{k,m}. \tag{6}$$

In the long-time limit, where $t \to \infty$, we assume $p(k,t+1) = p(k,t) = p_\infty(k)$, so Eq (6) becomes:

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

For $k > m$, $\delta_{k,m} = 0$. Rearranging Eq (7) we have:

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

Crucially, Eq (7) can be restated in terms of a gamma function Γ:

$$p_\infty(k) = A\,\frac{\Gamma(k+1-1)}{\Gamma(k+1+2)} = A\,\frac{\Gamma(k)}{\Gamma(k+3)}, \tag{9}$$

which has the following property [1]:

$$p_\infty(k) = A\,\frac{(k-1)!}{(k+2)!} = \frac{A}{k(k+1)(k+2)}. \tag{10}$$

For $k = m$, $\delta_{k,m} = 1$ and the $p_\infty(k-1)$ term vanishes, thus Eq (7) simply becomes:

$$p_\infty(k) = \frac{2}{2+m}. \tag{11}$$

Finally, by setting $k = m$ in Eq (10), we can equate the expression with Eq (11), leading to $A = 2m(m+1)$. We conclude that the degree distribution in the long-time limit is given by:

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \quad \text{for } k \geq m. \tag{12}$$

1.2.2 Theoretical Checks

An important property to check is that the sum of all probabilities $p(k)$ must sum to unity. Since only $k \geq m$ are allowed, starting from Eq (12):

$$\sum_{k=m}^{\infty} p_\infty(k) = \sum_{k=m}^{\infty} \frac{2m(m+1)}{k(k+1)(k+2)}. \tag{13}$$

The sum can be simplified by rearranging Eq (13) into partial fractions [2]:

$$\sum_{k=m}^{\infty} p_\infty(k) = m(m+1) \sum_{k=m}^{\infty} \frac{1}{k} - \frac{2}{k+1} + \frac{1}{k+2}, \tag{14}$$

$$\sum_{k=m}^{\infty} p_\infty(k) = m(m+1) \sum_{k=m}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1} \right) - \sum_{k=m}^{\infty} \left(\frac{1}{k+1} - \frac{1}{k+2} \right). \tag{15}$$

Only the first term in the two sums remains as the rest cancel out:

$$\sum_{k=m}^{\infty} p_\infty(k) = m(m+1) \left(\frac{1}{m} - \frac{1}{m+1} \right) = 1, \tag{16}$$

probabilities sum to unity as required. Note as $t \to \infty$, $k \to \infty$, reducing Eq (12) to its continuous case where $p_\infty(k) = k^{-3}$.

1.3 Preferential Attachment Degree Distribution Numerics

1.3.1 Fat-Tail

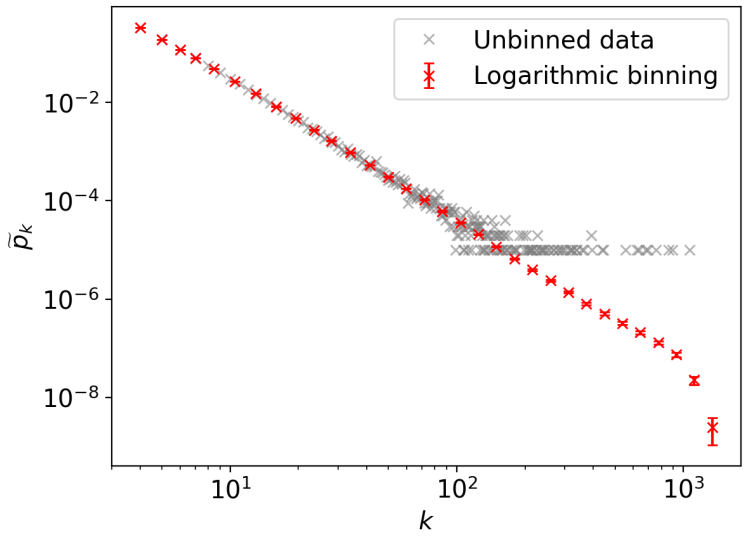


Figure 1: Degree distribution of the pure preferential attachment model for $m = 4$ and $N = 10^5$. Grey: unbinned data. Red: data after performing logarithmic binning with $s = 1.25$, which was chosen as a compromise between smoothness and keeping enough data. To account for the stochastic nature of a fat-tailed distribution, we ran 50 numerical runs and log binned the data separately. An average μ_i and a standard deviation σ_i were then computed for each bin i across all runs, which was displaced as the new probability $\widetilde{p_k}$ and an associated error bar from the standard error of the mean $(\sigma_i/\sqrt{50})$.

To reveal the scale-free behaviour of our networks, we organised our data in exponentially increasing bins [3], see Fig 1.

$$\frac{b_i}{b_{i+1}} = \exp\{\Delta\} = s, \tag{17}$$

where b_i and b_{i+1} are the bin widths of bin i and bin $i+1$.

1.3.2 Numerical Results

Degree distributions for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and fixed $N = 10^5$ were investigated (see Fig 2).

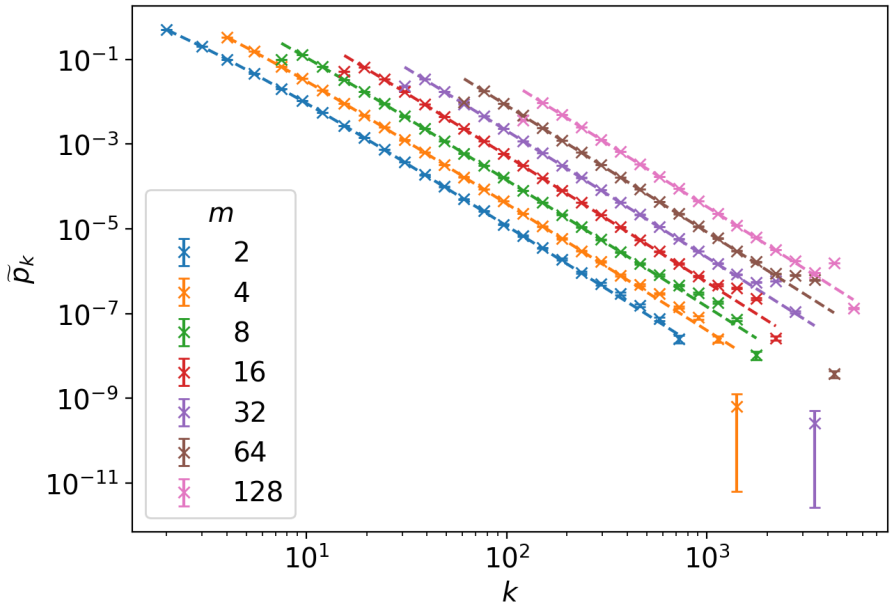


Figure 2: Degree distributions for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and fixed $N = 10^5$, dotted lines represent the theoretical fit according to Eq (12). Note a significant number of data located at higher k deviate away from the fit.

From Fig 2, it is evident that the majority of the data for small m such as $m = 2, 4, 8$ lies close to their theoretical fits given by Eq (12). For larger m, we observed more anomalies at larger k values, which increasingly deviate from their fits. Such behaviour can be attributed to the rise of a cutoff k due to a finite-size system, meaning the assumption made when deriving Eq (12) breaks down. As a result, the rest of the "missing" probabilities aggregate onto the last few bins, forming a bump.

If certain numerical runs have empty p_k at large k, zeros were padded. This resulted in occasional large uncertainties as the appearance of these values is inconsistent. The first point for $m = 8$ and beyond also appeared to deviate below the fit but can be solved by reducing the log bin scale.

1.3.3 Statistics

The reduced chi-square statistics χ^2_{red} was chosen over the Kolmogorov–Smirnov test (KS test) and the chi-square test χ^2 to examine the goodness of fit of our data. First, χ^2_{red} is the only test that considers a number of samples, in our case, the 50 numerical runs. Thus, its test statistics are more representative and reliable. Second, our data are discrete after log binning. Third, since each numerical run is independently run, it follows that p_k of the same bin (before

averaging) are independent and identically distributed. By the central limit theorem, we expect the data at a specific bin to obey the normal distribution.

We define the test statistics as follows:

$$\chi^2_{red} = \frac{1}{n-1}\sum_{i=0}^{n-1}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2, \tag{18}$$

where x_i, μ_i and σ_i are the theoretical frequency, mean frequency and standard deviation of the sampled frequency at bin i, and n is the number of data points after log binning. Since log binning produces probability values, we converted them into frequencies by multiplying with N. Ideally, a $\chi^2_{red} \sim 1$ indicates that the data and the theoretical distribution match in the agreement with the error variance [4]. To more quantitatively structure our argument, a one-sided hypothesis test at 0.05 significance level α was performed, where we made a null hypothesis H_0 that the set of μ_i is well described by the theoretical distribution given by Eq (12).

A corresponding test statistic $\chi^2_{red,\alpha}$ was computed using a chi-square distribution table [5]. If $\chi^2_{red} > \chi^2_{red,\alpha}$, we can be 95% confident in rejecting the null hypothesis. The initial results and conclusions were summarised in Tab 2 (a). Computing a χ^2_{red} on all data means that we would include the finite-size effect, thus it is trivial to see that for $m \geq 4$, the null hypothesis was rejected. A fairer comparison to the theoretical distribution would be to remove the anomalies. Through trial and error, we found that by removing the first and last four data, the resulting χ^2_{red} are very close to 1 for all m, see Tab 2 (b). This indicates a good model fit and hence not to reject the null hypothesis (apart from $m = 4$).

Table 2: Measurements of χ^2_{red} for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ for the preferential attachment model.

(a) Keeping all data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	0.41	1.52	No
4	2.30	1.51	Yes
8	3397.42	1.52	Yes
16	6248.24	1.54	Yes
32	10215.76	1.56	Yes
64	22108.15	1.59	Yes
128	52612.13	1.62	Yes

(b) Removing first and last four data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	0.49	1.59	No
4	2.38	1.51	Yes
8	1.04	1.59	No
16	0.83	1.62	No
32	1.23	1.64	No
64	1.21	1.69	No
128	1.58	1.64	No

1.4 Preferential Attachment Largest Degree and Data Collapse

1.4.1 Largest Degree Theory

It is by definition that only one vertex can have the largest degree, and in theory, this vertex must exist in degree k_1 and beyond. In the long-time limit, we have:

$$\sum_{k=k_1}^{\infty} n_\infty(k) = \sum_{k=k_1}^{\infty} N(t)p_\infty(k) = 1. \tag{19}$$

Substituting Eq (15) into Eq (19):

$$\sum_{k=m}^{\infty} p_\infty(k) = m(m+1) \sum_{k=k_1}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1} \right) - \sum_{k=k_1}^{\infty} \left(\frac{1}{k+1} - \frac{1}{k+2} \right). \tag{20}$$

Only the first term in the two sums remains as the rest cancel out:

$$\sum_{k=m}^{\infty} p_\infty(k) = m(m+1) \left(\frac{1}{k_1} - \frac{1}{k_1+1} \right) = \frac{1}{N(t)}. \tag{21}$$

Multiplying out terms in Eq (21), we obtain a quadratic equation:

$$k_1^2 + k_1 - N(t)m(m+1) = 0, \tag{22}$$

which means:

$$k_1 = \frac{1}{2} \left[-1 + \sqrt{1 + 4N(t)m(m+1)} \right]. \tag{23}$$

Note only the positive solution was chosen because it is not physically to have negative k_1.

1.4.2 Numerical Results for Largest Degree

Degree distributions for $N \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and fixed $m = 2$ were investigated (see Fig 3). A small m was chosen for three reasons. First, visually (see Fig 2) it appeared to be the best fit to the theoretical distribution, without any anomalous data and significant error bar. Second, it is the only m value that was not rejected in both Tab 2 (a) and (b). Third, a small m means the network is less susceptible to the initial graph.

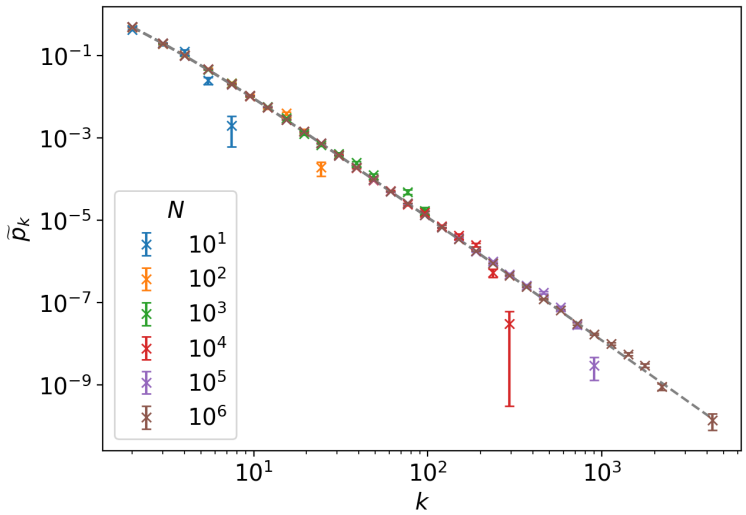
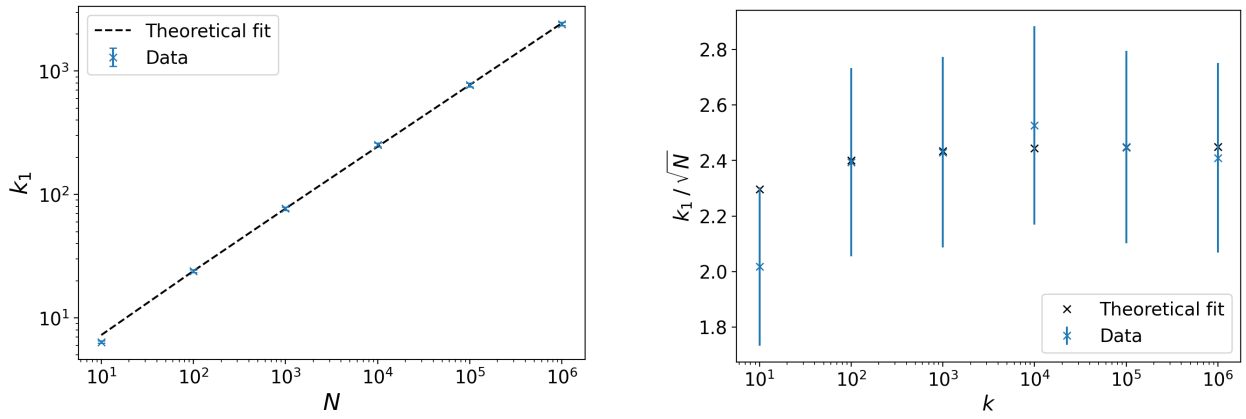


Figure 3: Degree distributions of the pure preferential attachment model for $N \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and fixed $m = 2$. The grey dotted line represents the theoretical fit according to Eq (23).

For $N \gg m$, Eq (23) becomes $k_1 \propto \sqrt{N}$. A linear line was fitted in Fig 4 (a). We found the gradient a of the log-log plot to be 0.51 ± 0.01, which is only differs from the expected value (0.5) by 2%, indicating a good overall fit. To more clearly reveal the difference between our data and the theoretical result, we divided k_1 by the N dependence, see Fig 4 (b). The data across all N are largely consistent (same order of magnitude). Apart from $N = 10^1$, the rest of the data are in accord with the error bars. The deviation at the small N value can be attributed to the breakdown of the $N \gg m$ assumption.



(a) (b)

Figure 4: (a) k_1 against N for $N \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and fixed $m = 2$. The black dotted line represents the theoretical distribution according to Eq (12). A linear line was fitted on the data, gradient a was measured to be 0.51 ± 0.01. This resulted in an overall 2% error, indicating a good fit. (b) $k_1/\sqrt{N}$ against N. $N = 10^1$ is only marginally agreeing with the theoretical value.

1.4.3 Data Collapse

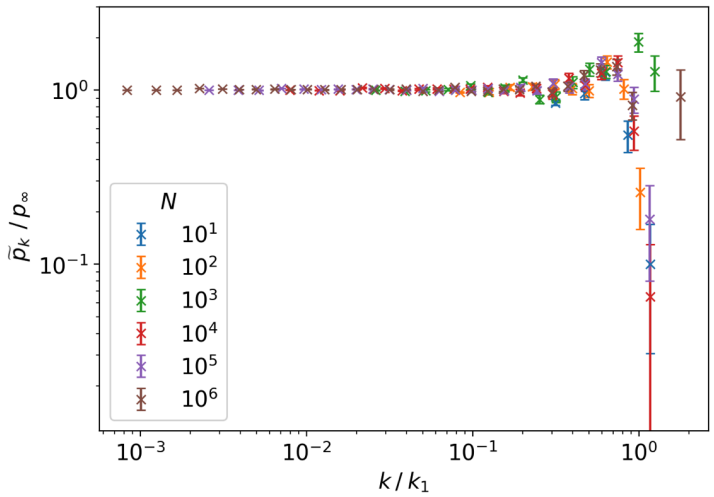


Figure 5: Preferential attachment data collapse of the degree distributions for $N \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and fixed $m = 2$. Error bars were scaled by the theoretical distribution according to Eq (12).

Recall from Fig 3, we observed that a cutoff degree k_1 exists for finite-size systems. This is evident from the sharp fall-off of p_k at large k which is typically associated with large error bars. This suggests a dependency of k_1 across all N. Similarly, the sampled probability is constrained by the theoretical distribution p_∞. Thus, to reveal the uniform behaviour of our networks, we shall re-scale $\widetilde{p_k}$ to $\widetilde{p_k}/p_\infty$ and k to k/k_1, leading onto the following finite-size scaling ansatz:

$$\widetilde{p_k} = p_\infty \, F\left(\frac{k}{k_1}\right), \tag{24}$$

where $F(k/k_1)$ is the scaling function. By doing the above, our degree distributions collapsed onto a single plot (see Fig 5). Crucially, we have found a "blueprint" that describes the commonality across all networks: a cutoff degree always results in a cluster of "missing" probabilities.

2 Phase 2: Pure Random Attachment Π_{rnd}

2.1 Random Attachment Theoretical Derivations

2.1.1 Degree Distribution Theory

For random attachment, it is equally likely to pick any existing vertex:

$$\Pi(k,t) = \Pi_{ra} = \frac{1}{N(t)}. \tag{25}$$

Substituting Eq (25) into the master equation Eq (4), we proceed with another difference equation:

$$p(k,t+1) = m[\,p(k-1,t) - p(k,t)\,] + \delta_{k,m}. \tag{26}$$

Again, in the long-time limit, we assume $p(k,t+1) = p(k,t) = p_\infty(k)$, so Eq (26) becomes:

$$p_\infty(k) = m[\,p_\infty(k-1) - p_\infty(k)\,] + \delta_{k,m}. \tag{27}$$

For $k > m$, $\delta_{k,m} = 0$. Rearranging Eq (27) we have:

$$p_\infty(k) = \frac{m}{1+m}\,[\,p_\infty(k-1) + \delta_{k,m}\,]\,. \tag{28}$$

For $k > m$, $\delta_{k,m} = 0$, by induction:

$$p_\infty(k) = \left(\frac{m}{1+m}\right)^{k-m} p_\infty(m). \tag{29}$$

For $k = m$, $\delta_{k,m} = 1$ and the $p_\infty(k-1)$ term vanishes. Equation (28) becomes:

$$p_\infty(k) = \frac{1}{1+m}. \tag{30}$$

To conclude, the degree distribution for random attachment in the long-time limit is as follows:

$$p_\infty(k) = \frac{m^{k-m}}{(1+m)^{k-m+1}} \quad \text{for } k \geq m. \tag{31}$$

To validate Eq (31), we can check if probabilities sum to unity:

$$\sum_{k=m}^{\infty} p_\infty(k) = \sum_{k=m}^{\infty} \frac{m^{k-m}}{(1+m)^{k-m+1}} = \frac{m^{-m}}{(1+m)^{-m+1}} \sum_{k=m}^{\infty} \left(\frac{m}{1+m}\right)^k. \tag{32}$$

By relabelling the index in the sum of Eq (32) as $k \to k' + m$:

$$\sum_{k=m}^{\infty} p_\infty(k) = \frac{m^{-m}}{(1+m)^{-m+1}} \sum_{k'=0}^{\infty} \left(\frac{m}{1+m}\right)^m \left(\frac{m}{1+m}\right)^{k'}, \tag{33}$$

we have a form that can be solved using the infinite geometric series $\sum_{n=0}^{\infty} ar^n = a/(1-r)$. Thus,

$$\sum_{k=m}^{\infty} p_\infty(k) = \frac{m^{-m}}{(1+m)^{-m+1}} \left(\frac{m}{1+m}\right)^m (1+m) = 1, \tag{34}$$

which is normalised as expected.

2.1.2 Largest Degree Theory

Similar to Sec 1.4.1, we sum the probabilities from degree k_1 and expect only one vertex. Substituting Eq (31) into Eq (19) and rearranging:

$$\sum_{k=k_1}^{\infty} p_\infty(k) = \frac{m^m}{(1+m)^{m+1}} \sum_{k=k_1}^{\infty} \left(\frac{m}{1+m}\right)^k. \tag{35}$$

Relabelling the index in the sum of Eq (35) as $k \to k' + k_1$:

$$\sum_{k=k_1}^{\infty} p_\infty(k) = \sum_{k'=0}^{\infty} \left(\frac{m}{1+m}\right)^{k_1} \left(\frac{m}{1+m}\right)^{k'}. \tag{36}$$

which has allowed us to apply the infinite geometric series, thus Eq (36) becomes:

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

$$\left(\frac{m}{1+m}\right)^{-m+k_1} = \frac{1}{N(t)}. \tag{38}$$

We can log both sides, rearranging gives:

$$k_1 = \frac{\log\left|\left(\frac{m}{m+1}\right)^m \frac{1}{N(t)}\right|}{\log\left|\frac{m}{m+1}\right|}. \tag{39}$$

2.2 Random Attachment Numerical Results

2.2.1 Degree Distribution Numerical Results

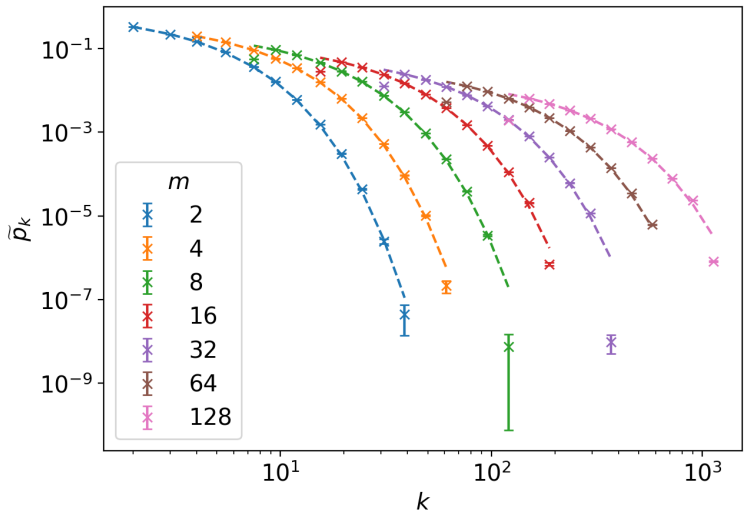


Figure 6: Degree distributions for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and fixed $N = 10^5$. A similar process took place in log binning the data and finding their means.

Figure 6 shows that data for $m = 2, 4$ are in good agreement with the theoretical distribution given by Eq (31). Similar to our investigation in Sec 1.4.2, data at high k for larger m increasingly deviate from the fit and larger error bars, as a result of the finite-size effect. Different from the preferential attachment case, the random attachment networks do not follow

a power-law relationship. A reduced chi-square statistic was computed for each m (see Tab 3). Initial findings showed that χ^2_{red} for $m = 2, 4$ are both ≈ 1, χ^2_{red} increases for larger m, indicating a worse fit. By removing the first and last three data, χ^2_{red} has become much closer to unity and lead us to not reject the null hypothesis for $m = 8, 16, 32, 64$ too.

Table 3: Measurements of χ^2_{red} for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ for the random attachment model.

(a) Keeping all data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	1.46	1.79	No
4	1.24	1.79	No
8	2407.07	1.64	Yes
16	5833.01	1.79	Yes
32	6778.34	1.79	Yes
64	14738.93	1.83	Yes
128	19928.68	1.83	Yes

(b) Removing first and last three data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	2.19	2.01	Yes
4	1.71	2.01	No
8	1.34	1.94	No
16	0.91	2.01	No
32	1.27	2.01	No
64	1.45	2.10	No
128	2.19	2.10	Yes

2.2.2 Largest Degree Numerical Results

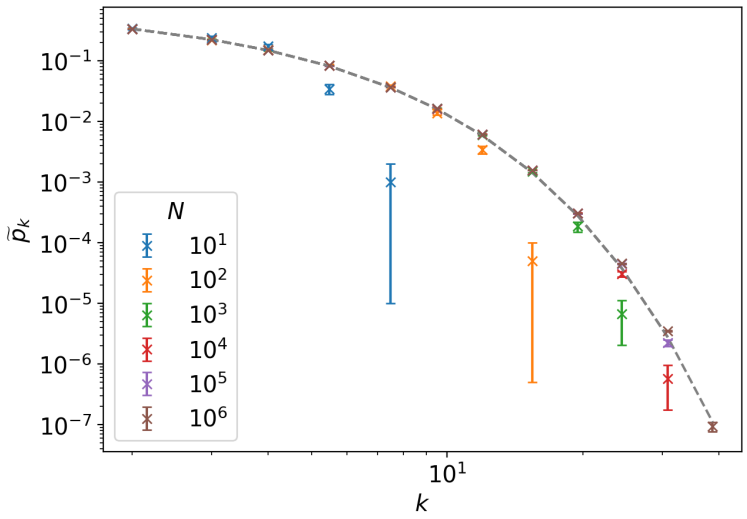
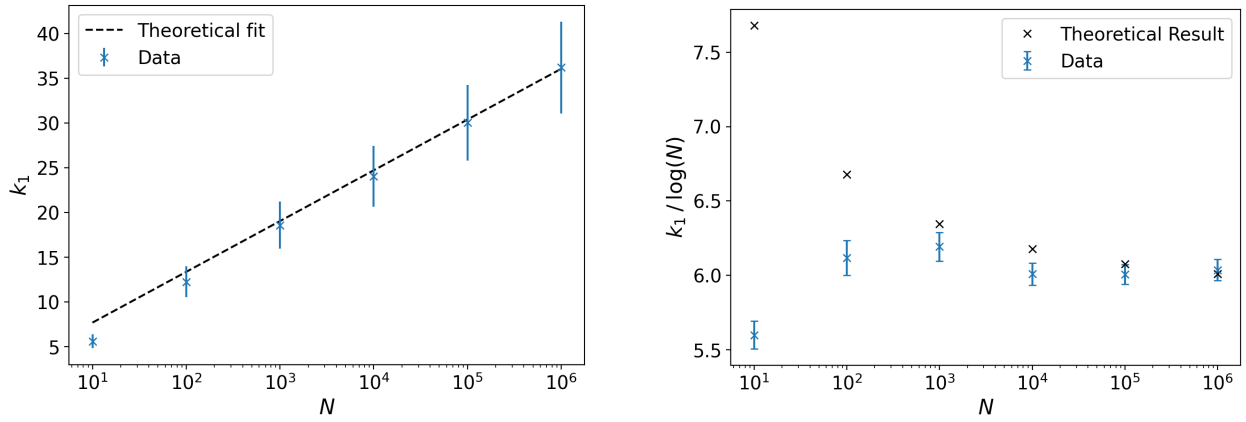


Figure 7: Degree distributions for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and fixed $N = 10^5$. The grey dotted line represents the theoretical fit according to Eq (39). Note for the largest $N = 10^6$, data are in strong agreement with the fit. Large errors are common for $N = 10^1, 10^2$, which can be attributed to the breakdown of the long time assumption.

For $N \gg m$, Eq (39) becomes $k_1 \propto \log(N^1)$. A linear line was fitted in Fig 8 (a). From the gradient, we measured the exponent of the N dependence to be 1.07 ± 0.01 which is 7% away from the true value. This was observed to be a larger relative difference than the preferential

attachment case. Dividing k_1 by the N dependence, we revealed that data for $N = 10^1, 10^2$ are more separated from the theoretical result, with $N = 10^1$ being the worst. This is because the limited size of the system means that the initial conditions would cause a significant effect.



(a) (b)

Figure 8: (a) k_1 against N for $N \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and fixed $m = 2$. The black dotted line represents the theoretical distribution according to Eq (31). Almost all data are described by the error bars. (b) $k_1/\log N$ against N. $N = 10^1 = 10^2$ significantly deviate from the theoretical result, which is attributed to the finite-size effect.

3 Phase 3: Existing Vertices Model

3.1 Existing Vertices Model Theoretical Derivations

For this model, $m/2$ edges first form between the new vertex and existing vertices by random attachment, the other $m/2$ edges subsequently form between pairs of existing vertices. A new master equation was defined:

$$\begin{aligned} n(k,t+1) = n(k,t) + \frac{m}{2}\left[\,\Pi_{ra}(k-1,t)n(k-1,t) - \Pi_{ra}(k,t)n(k,t)\,\right] \\ +m\left[\,\Pi_{pa}(k-1,t)n(k-1,t) - \Pi_{ra}(k,t)n(k,t)\,\right] + \delta_{k,m/2}. \end{aligned} \tag{40}$$

Note the delta function has changed because we would now include the new node after it has formed $m/2$ edges. Substituting Eq (5) and (25) (definition of Π_{ra} and Π_{pa} into Eq (40):

$$p(k,t+1) = \frac{m}{2}\left[\,p(k-1,t) - p(k,t)\,\right] + \frac{1}{2}\left[\,(k-1)p(k-1,t) - kp(k,t)\,\right] + \delta_{k,m/2}. \tag{41}$$

Assuming a long-time limit, Eq (40) becomes:

$$p_\infty(k) = \frac{m}{2}\left[\,p_\infty(k-1) - p_\infty(k)\,\right] + \frac{1}{2}\left[\,(k-1)p_\infty(k-1) - kp_\infty(k)\,\right] + \delta_{k,m/2}. \tag{42}$$

For $k > m/2$, $\delta_{k,m/2} = 0$. Rearranging Eq (42) we have:

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k+m-1}{k+m+2}. \tag{43}$$

Equation (43) can be restated in terms of the gamma function Γ:

$$p_\infty(k) = A\,\frac{\Gamma(k+m+1-1)}{\Gamma(k+m+1+2)} = A\,\frac{\Gamma(k+m)}{\Gamma(k+m+3)}, \tag{44}$$

which has the following property [1]:

$$p_\infty(k) = A\frac{(k+m-1)!}{(k+m+2)!} = \frac{A}{(k+m)(k+m+1)(k+m+2)}. \tag{45}$$

For $k = m/2$, $\delta_{k,m/2} = 1$ and the $p_\infty(k-1)$ term vanishes, thus Eq (42) trivially becomes:

$$p_\infty(k) = \frac{4}{4+3m}. \tag{46}$$

Finally, by setting $k = m/2$ in Eq (45), we can equate the expression with Eq (46), leading to $A = \frac{1}{2}(3m)(3m+2)$. We conclude that the degree distribution in the long-time limit is given by:

$$p_\infty(k) = \frac{3m(3m+2)}{2(k+m)(k+m+1)(k+m+2)} \quad \text{for } k \geq m/2. \tag{47}$$

3.2 Existing Vertices Model Numerical Results

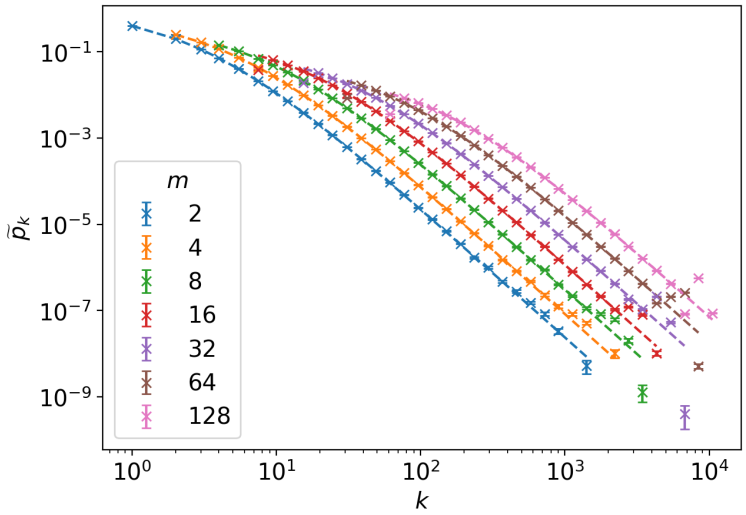


Figure 9: Degree distributions for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and fixed $N = 10^5$. A new scale of $s = 1.3$ was used to smooth out the data. A similar bump was observed in the data at larger values of k, which was caused by the finite-size system. Note we have allowed multigraphs in this model for ease of computation. The overall effect on the distribution is negligible because the number of repeated edges added becomes increasingly limited as $t \to \infty$.

From Fig 9, most data are well aligned to the theoretical distribution given by Eq (47). The agreement is less well for larger m such as $m = 8$ and beyond. Generally, error bars are relatively small compared to the previous cases. Interestingly, smaller m enter a near power-law relationship sooner ($p_\infty(k) \to k^-3$ for $k \gg m$), evident from the flatter lines on the log-log plot. A reduced chi-square statistic was calculated for each m to quantify the goodness of our fit (see Tab 4). Removing the first and last two data was also observed to produce better test statistics for larger m. Not rejecting any null hypothesis, ie. in Tab 4 (b), is not sufficient to conclude on the goodness of fit. Judging from the χ^2_{red} for all m across Tab 4 (a) and (b), no value ≈ 1, thus I would conclude that our data are not well fitted.

Table 4: Measurements of χ^2_{red} for $m \in \{2, 4, 8, 16, 32, 64, 128\}$ for the existing vertices model.

(a) Keeping all data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	0.28	1.67	No
4	0.46	1.51	No
8	0.76	1.51	No
16	1.52	1.52	No
32	3836.28	1.67	Yes
64	365.08	1.52	Yes
128	13678.00	1.51	Yes

(b) Removing first and last two data.

m	χ^2_{red}	$\chi^2_{red,\alpha}$	Reject H_0?
2	0.30	1.51	No
4	0.35	1.52	No
8	0.20	1.52	No
16	0.55	1.51	No
32	0.83	1.51	No
64	0.12	1.83	No
128	0.74	1.79	No

4 Conclusions

In conclusion, numerical solutions rarely match up exactly with theoretical derivations. In Phase 3, we presented a more realistic model combining preferential and random attachments and have shown that despite breaking certain theoretical constraints, the model still exhibits a fat-tailed distribution.

References

[1] Evans, T., 2021. C&N: Networks PS2: Random Graphs. p.3.

[2] Secular, P., 2015. Preferential and random attachment models of a complex network. p.6.

[3] McGillivray, M., 2020. logbin_2020.py. Imperial College London.

[4] En.wikipedia.org. n.d. Reduced chi-squared statistic - Wikipedia. [online] [Accessed 28 March 2022].

[5] People.richland.edu. n.d. Table: Chi-Square Probabilities. [online] [Accessed 28 March 2022].