Network Project
A Growing Network Model

CID: 00941460

25th March 2019

Abstract:

The growth of networks is investigated in Python via three different attachment models: preferential (aka the Barabsi-Albert model), random attachment, and existing vertices model. Specifically, the aim is to compare numerical results with theory. Through the Kolmogorov-Smirnov statistical testing, all the numerical results for degree distribution were found to be consistent with theory with a much higher p-value than a typical 0.1 significance test. The numerical largest degree for preferential and random attachment also show good agreement with the theoretical scaling with N, whereas existing vertices model does not. Finite-size effects such as bumps and rapid decays are revealed through data collapse of the degree distribution.

Word Count: 2483 words excluding font page, figure captions, table captions, and bibliography.

1 Introduction

A graph (or network) is a collection of nodes (N) and edges (E). Graph growth requires adding new nodes and edges. The degree of a node represents the number of edges it has, denoted k. The growth of graphs is investigated using preferential (or BA model), random and existing vertices models. By simulating a variety of graphs under different initial conditions, this project aims to compare the numerical and theoretical results of the degree distribution $p(k)$ and largest degree k_1 of a graph in the asymptotic limit and explore the finite size effects.

2 Phase 1: Pure Preferential Attachment Π_{pa}

2.1 Implementation

2.1.1 Numerical Implementation

The BA model was implemented in Python using the following algorithm.

1. Initiate a graph $\mathcal{G}_0$ at time $t = 0$, with N_0 nodes and E_0 edges.
2. Increment time $t \to t + 1$
3. Add one new node.
4. Add m new edges, with one end attached to the new node, the other to an existing node chosen with probability Π.
5. Repeat from step 2 until node count reaches final number of N nodes.

Nodes are labelled with numbers 0,1,2,etc. Edges are labelled as source node and target node. For step 4, all m edges must be attached to unique nodes to ensure a simple graph is produced. The list of all the nodes is stored in `node_ls`. All the edges are stored as a list of tuples in `edge_ls`. Every time a new edge is attached to a node, that node is added to a list called `attached_node_ls`. The number of times a particular node appears in this list is proportional to its degree.

In this work, Π takes one of three forms. For preferential attachment, $\Pi = \Pi_{pa}(k) = k/(2E) \propto k$. This means a node is more likely to be chosen if it has more edges to begin with (BA model [1]). Thus, nodes are chosen from the `attached_node_ls` using `random.choice` method. The form of Π is changed as required for random attachment and existing vertices model. Random attachment requires choosing nodes uniformly. This is done using `random.choice` on `node_ls` since each node only appears once in this list. Existing vertices model used a combination of the above.

To calculate the degree distribution, `collections.Counter` is used to count frequency (i.e. the degree) of each node in `attached_node_ls`. The corresponding probability $p(k)$ is obtained using the logarithmic binning routine provided.

Several numerical runs (with unique seeds) are done for each set of initial conditions to allow calculation of a mean value for the quantity of interest. The standard error of the mean is then given by the standard deviation of all the runs over the square root of the number of runs.

2.1.2 Initial Graph

A complete simple graph with $N_0 = m + 1$ nodes is used as the initial graph at $t = 0$. The argument is as follows. The number of edges in a complete graph is $E_0 = N_0(N_0 - 1)/2$. Note the -1 because of no self-loops and the factor of $1/2$ because each edge has 2 nodes. On the other hand, the number of edges in a simple graph with uniform degree m is $E_0 = mN_0/2$. This is expected to happen in the large graph limit. Therefore, the initial graph should have uniform degree to minimise the effects of initialisation on the growth dynamics. Equating the two expressions and ignoring the trivial solution gives $N_0 = m+1$. A sparse graph with uniform degree of 2 is also tested, although the numerical results are similar. Subsequently, only complete graphs are used for initialisation.

For the existing vertices model, the initial graph is not complete. There are $m/2$ edges unconnected in the initial graph such that a complete graph is made when $m/2$ new edges are added between existing vertices for the first timestep at $t = 1$. This then allows the first new node to be attached to a complete graph.

2.1.3 Type of Graph

The simulations produce simple graphs as the model does not consider repeated edges (i.e. no multigraph) or directed edges. In addition, the edges are non-weighted and there are no self-loops.

2.1.4 Working Code

The `networkx` package is used as a testing tool to visualise very small graphs (e.g. 10 nodes) at each timestep. The following variables are printed in the Python console to compare with the visual graphs: initial number of nodes and edges, initial node list, edge list, and degree distribution, node and edges added at each timestep, and the new degree distribution at each timestep. This procedure ensures that the programme is working as expected.

2.1.5 Parameters

The parameters required for the simulation are: final number of nodes (N), number of edges added per timestep (m), number of seeds ($Nseeds$), and the log-binning scale (*logbin_scale*) which quantifies how much bigger each bin size is compared to the previous bin (e.g. 1.1 means each bin is 10% bigger than the previous one.)

For fixed N but varying m, $N = 10^4$, $m = [1, 3, 9, 27, 81, 243]$ with $Nseeds = [10^3, 10^3, 10^2, 10^2, 10^2, 10^2]$ respectively, and *logbin_scale*$= 1.1$.

For fixed m but varying N, $N = [10, 10^2, 10^3, 10^4, 10^5, 10^6]$ with $Nseeds = [10^4, 10^4, 10^4, 10^3, 10^2, 10^1]$ respectively, and *logbin_scale*$= 1.1$.

Specifically for existing vertices model, $m = [2, 4, 8, 16, 32]$ with $Nseeds = [10^2, 10^2, 10, 10, 10]$ respectively. m must be even to split the attachment half-half between new node and existing nodes.

These values were chosen on the basis of hardware and time constraints. In particular for preferential and random attachment, base 3 was chosen for m so that a big enough range was covered without the data being too cluttered as in the case for base 2. *logbin_scale*$= 1.1$ was chosen as it was found to be optimal for displaying the underlying distribution of the data.

2.2 Preferential Attachment Degree Distribution Theory

2.2.1 Theoretical Derivation

The growth of a graph is described generally using the master equation [2]

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

where $n(k,t)$ is the number of nodes with degree k at time t, m is the number of edges added per timestep, $\Pi(k,t)$ is the probability that a new node is connected to a node with degree k, and δ_{km} is the Kronecker-delta such that it equals one when $k = m$ and zero otherwise. Importantly, this equation only considers nodes with degree $(k-1)$ at the current timestep t which imply that only one edge can be added to a particular node per timestep, i.e. nodes with $k-1 \to k$. Thus for $m > 1$, the master equation fails to capture the contributions to $n(k,t+1)$ from multiple edges connecting to the same node of lower degree, say $(k-2), (k-3)$ and so on. However, in the large graph limit, the likelihood of multiple edges attaching to the same node diminishes as $N >> m$. Therefore, it is reasonable to ignore terms $n(k-2,t)n(k-3,t)$ in the master equation. Using the fact that one node is added per timestep, the total number of nodes in the graph evolves as $N(t+1) = N(t) + 1$. We can define degree probability distribution as

$$p(k,t) = \frac{n(k,t)}{N(t)}. \tag{2}$$

Dividing equation 1 through by $N(t+1)$ and replacing n using equation 2 transforms the master equation to

$$p(k,t+1) = N(t)[p(k,t) - p(k,t+1)] + mN(t)[\Pi(k-1,t)p(k-1,t) - \Pi(k,t)p(k,t)] + \delta_{km}. \tag{3}$$

Assuming $p(k,t)$ becomes asymptotically stationary in large time limit, the first term on the right-hand side vanishes. Thus, equation 3 can be expressed in terms of $p_\infty(k) \equiv lim_{t\to\infty} p(k,t)$ giving

$$p_\infty(k) = mN(t)[\Pi(k-1,t)p_\infty(k-1) - \Pi(k,t)p_\infty(k)] + \delta_{km}. \tag{4}$$

The exact form of $p_\infty(k)$ depends on $\Pi(k,t)$. For the attachment schemes considered, we make use of the following three results:

1. $\frac{E(t)}{N(t)} \to m$ as $t \to \infty$, for any initial graph configurations. This can be shown as follows. Consider an initial graph with N_0 nodes and E_0 edges. It evolves as

$$\begin{aligned} E(t) &= E_0 + mt \\ N(t) &= N_0 + t \end{aligned} \tag{5}$$

For finite graphs

$$\frac{E(t)}{N(t)} = \frac{E_0 + m(N - N_0)}{N} = m\Big(1 - \frac{N_0}{N}\Big) + \frac{E_0}{N}. \tag{6}$$

However, in the limit $t \to \infty$, $N \to \infty$. Thus, equation 6 reduces to

$$\lim_{t\to\infty} \frac{E(t)}{N(t)} = m. \tag{7}$$

This would be a reasonable assumption if $N \gg N_0$ and $N \gg E_0$.

2. If a graph grows one new node of degree m per timestep,

$$p_\infty(k|k<m)=0. \tag{8}$$

This is true because all new nodes already have degree m. These new edges will be attached to existing nodes so that in the asymptotic limit, all the existing nodes will have at least degree m.

3. The solution to the difference equation

$$\frac{f(z)}{f(z-1)}=\frac{z+a}{z+b} \tag{9}$$

is

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

where A is an arbitrary constant. This is done using the Gamma function, $\Gamma(z)$, which has the properties [3]

$$\Gamma(z+1)=z\Gamma(z), \qquad \Gamma(1)=1. \tag{11}$$

This is equivalent to $\Gamma(z+1)\equiv z!$.

In pure preferential attachment, $\Pi=\Pi_{pa}(k)=k/(2E)\propto k$, meaning nodes with higher degree are more likely to be chosen. The factor of 2 comes from each edge contributing to the degree of 2 nodes. Substituting Π into equation 4, and using equation 7 gives

$$\begin{aligned} p_\infty(k) &= mN(t)\frac{k-1}{2E(t)}p_\infty(k-1)-mN(t)\frac{k}{2E(t)}p_\infty(k)+\delta_{km} \\ &= \frac{k-1}{2}p_\infty(k-1)-\frac{k}{2}p_\infty(k)+\delta_{km}. \end{aligned} \tag{12}$$

For $k>m$, $\delta_{km}=0$. Equation 12 becomes

$$\frac{p_\infty(k)}{p_\infty(k-1)}=\frac{k-1}{k+2}, \tag{13}$$

which has the same form as equation 9. This is solved using equation 10 giving

$$\begin{aligned} p_\infty(k) &= A\frac{\Gamma(k)}{\Gamma(k+3)} \\ &= \frac{A}{k(k+1)(k+2)}. \end{aligned} \tag{14}$$

For $k=m$, $\delta_{km}=1$. Equation 12 becomes

$$p_\infty(m)=\frac{2}{m+2}, \tag{15}$$

where equation 8 is used. Equating equation 14 and 15 with $k=m$ gives

$$A=2m(m+1). \tag{16}$$

Thus, the asymptotic degree distribution for preferential attachment is

$$p_\infty(k)=\frac{2m(m+1)}{k(k+1)(k+2)} \qquad \text{for} \quad k\geq m. \tag{17}$$

In the large k limit, $p_\infty(k)\propto k^{-3}$.

2.2.2 Theoretical Checks

To check that A gives the correct normalisation for $p_\infty(k)$,

$$\begin{aligned}
\sum_{k=1}^{\infty} p_\infty(k) &= \sum_{k=m}^{\infty} \frac{A}{k(k+1)(k+2)} && \text{Since } p_\infty(k|k<m)=0) \\
&= \frac{A}{2}\sum_{k=1}^{\infty}\Big(\frac{1}{k} - \frac{2}{k+1} + \frac{1}{k+2}\Big) && \text{Partial fractions} \\
&= \frac{A}{2}\Big(\frac{1}{m} - \frac{1}{m+1}\Big) && \text{Cancellation of expanded sum} \\
&= \frac{A}{2m(m+1)} \\
&= 1,
\end{aligned}$$

as required for a probability distribution.

The average degree $\langle k\rangle = \sum_{k=m}^{\infty} k p_\infty(k) = 2m$, again utilising the cancellations of partial fractions when expanding the sum. This means on average each node added with degree m will also receive edges from m other nodes in the large graph limit.

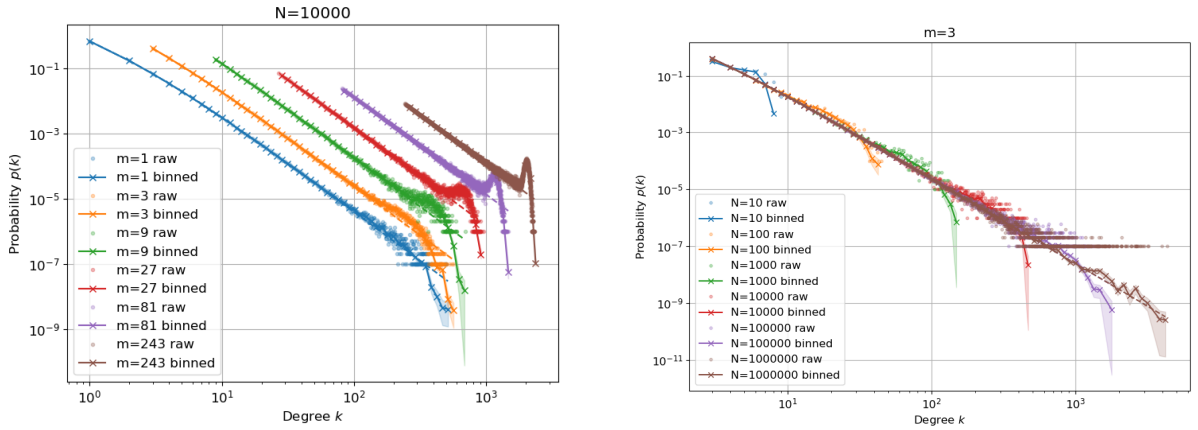
2.3 Preferential Attachment Degree Distribution Numerics

2.3.1 Fat-Tail

Fat-tailed distributions are expected as a result of power law distributions. This is clearly the case for preferential attachment as seen in equation 17. Logarithmic binning was used to deal with the statistical noise caused by the fat-tailed distribution in the raw data. By trial and error, *logbin_scale*= 1.1 was found to be optimal for displaying the underlying distribution of the data without being too smoothed out such that it loses important features.

2.3.2 Numerical Results

Figure 1a shows $p(k)$ against k on a log-log scale for various m values and for $N = 10^4$. As mentioned previously, it is expected to have a fat tail. This is seen clearly in the raw data (circles). The filled-in regions show the standard error of the mean for the log-binned data. The asymptotic degree distributions $p_\infty(k)$ are plotted as dashed lines. The numerical data follows the theoretical result closely up to a characteristic bump after which the avalanche size probability decays rapidly. Therefore, the degree at which the bump occurs can be thought of as the largest degree k_1 (or cut-off degree) since higher degrees are extremely unlikely (see section 2.4.1 for the theoretical value and section 2.4.2 for numerical results). These bumps look very similar to those observed in finite systems that display self-organised criticality [5]. They are characteristics of a finite system since the number of edges on a node is limited by the total number of nodes and how many edges are added at each timestep, thus a restriction on the maximum degree allowed. Therefore, there is a preference for these specific degrees producing the excess probability bump. The bump gets larger, narrower and occur at higher k for higher values of m as it approaches the $N-1$ edge limit of a complete graph. In addition, larger m systems have shorter power-law scaling regimes. This is because the smallest degree is shifted to a value close to m as reasoned in equation 7.



(a) (b)

Figure 1: a) The raw and log-binned degree distributions for preferential attachment, plotted with errors (filled-in colour), for varying m and fixed $N = 10^4$. The dashed lines indicate the theoretical power-law behaviour which agree with numerical data up to the characteristic bump at large k due to a finite graph in the simulation. For large m, the bump is much more prominent. b) The degree distribution for preferential attachment, plotted with errors (filled-in colour), for varying N and fixed $m = 3$. The bump moves to the right as N increases.

2.3.3 Statistics

The Kolmogorov-Smirnov (KS) test was used to statistically test the null hypothesis that the numerical and theoretical distributions are in fact the same distribution. The reason for using KS test is due to its superiority to linear regression when it comes to testing goodness of fit for data that follow an asymptotic power law with a fat-tailed distribution [6]. The key statistic of the KS test is the p-value, where a high p-value means we cannot reject the null hypothesis. A typical significance level to reject the null hypothesis is 0.1, which means there is a 10% chance that the distributions are the same.

To perform the KS test, a sample of 1000 values of log-binned k was drawn using `numpy.random.choice` with numerical probabilities $p(k)$ and similarly 1000 values of binned k with the theoretical probabilities $p_\infty(k)$. Then these two samples were compared using the `scipy.stats.ks_2samp` function to calculate a p-value. This procedure was repeated 10 times to get an average p-value for this configuration of m and N. As stated in section 2.3.2, the deviation of $p(k)$ from theory at the tail is due to finite size effects. The reason for using the binned values instead of the raw values is because the raw values contains many fluctuations due to the fat-tailed distribution. This would make the statistics very unreliable. To improve the statistics further, the last 4 binned data points (crosses in figure 1a) were neglected to minimise finite-size effects. Implementing this change improved the p-value score noticeably.

The p-values for each of the m and N combination is summarised in the table 1. These values are clearly much higher than the standard 0.1 significance level which means the null hypotheses that the numerical data is consistent with their respective theoretical distributions cannot be rejected.

Π_{pa}			Π_{rnd}			Π_{pr}		
m	N	p-value	m	N	p-value	m	N	p-value
1	10^4	0.983	1	10^4	0.842	2	10^4	0.620
3	10^4	0.766	3	10^4	0.708	4	10^4	0.706
9	10^4	0.803	9	10^4	0.762	8	10^4	0.598
27	10^4	0.697	27	10^4	0.644	16	10^4	0.663
81	10^4	0.670				32	10^4	0.849
243	10^4	0.779						

Table 1: The Kolmogorov-Smirnov (KS) test statistics: The p-values are shown for every combination of m and N used in the simulations, under all three attachment models: preferential (Π_{pa}), random (Π_{rnd}), existing vertices (Π_{pr}). All p-values are much higher than the typical significance level of 0.1, indicating a very high probability that the numerical and theoretical degree distributions are the same. Note the missing values for Π_{rnd} due to overflow errors in Python.

2.4 Preferential Attachment Largest Degree and Data Collapse

2.4.1 Largest Degree Theory

The largest degree k_1 (or cut-off degree) of a graph refers to the value of k after which only one node is expected to be found with that degree on average [4]:

$$\sum_{k=k_1}^{\infty} N p_\infty(k) = 1 \tag{18}$$

Using partial fractions for $p_\infty(k)$, equation 18 becomes a quadratic in k_1

$$\begin{aligned} \frac{1}{N} &= \sum_{k=k_1}^{\infty} p_\infty(k) \\ &= \frac{2m(m+1)}{2}\Big(\frac{1}{k_1} - \frac{1}{k_1+1}\Big) \\ &= \frac{m(m+1)}{k_1(k_1+1)} \\ k_1^2 + k_1 - mN(m+1) &= 0. \end{aligned}$$

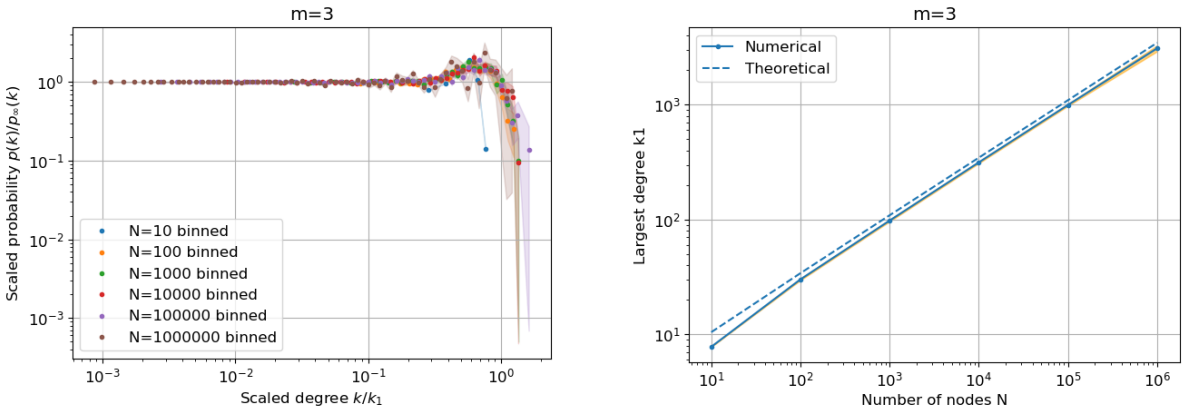
Applying the quadratic formula and taking the physical positive root gives the largest degree

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

In the large N limit, $k_1 \propto N^{1/2}$.

2.4.2 Numerical Results for Largest Degree

The largest degree (or cut-off degree) k_1 is found using the `max` function on the degree distribution `degreeLs`. The average value of k_1 is obtained by averaging over the largest degree of each distribution from each numerical run. The parameters used were $N = [10^n]$ for n in range 1 to 6, and $m = 3$. The number of numerical runs for each configuration is



(a) (b)

Figure 2: a) Data collapse of log-binned degree distribution for preferential attachment, plotted with errors (filled-in colour), for fixed $m = 3$ and varying N. The probability is scaled by the theoretical distribution collapsing to 1 for all N for $k < k_1$. Close to $k = k_1$, finite-size effects are revealed as a bump followed by rapid decay. b) A plot of k_1 against N, with error filled-in, showing good agreement between theory and numerical results. The slope of the dashed line is 0.5.

stated in section 2.1.5. The small m value was chosen on the basis of hardware and time constraints as it still allowed a broad range of N to be explored.

The uncertainties in k_1 comes from the standard deviation σ of all the different $k_1's$ from the `Nseeds` numerical runs. The standard error of the mean k_1 is then $\sigma/\sqrt{Nseeds}$.

Figure 1b clearly shows the cut-off k_1 increases with N. The plot of k_1 against N is shown in the figure 2b. The theoretical k_1 results from section 2.4.1 are shown as a dashed line. The `scipy.stats.linregress` function was used to do a linear regression on the numerical data giving a slope of 0.51 ± 0.01, which implies it agrees with the theoretical scaling of $N^{0.5}$. However, there is a clear discrepancy in the y-intercept suggesting that the cut-off k_1 depends on m as well in finite size networks.

In the above section, we found that k_1 is the cut-off degree unique to the network size N; the number of nodes. This means the degree values of different graphs can be normalised to its characteristic k_1 and the data should collapse vertically such that the cut-off is at $k/k_1 = 1$ if the theory value is correct. Similarly, the degree probability can also be normalised to its asymptotic theoretical value $p_\infty(k)$ such that $p(k)/p_\infty(k) = 1$.

Plotting $p(k)/p_\infty(k)$ against k/k_1 indeed shows a data collapse as expected (see figure 2a). $p(k)/p_\infty(k)$ is close to 1 before the bump indicating a good agreement with theory. The finite-size effect (i.e. the bump) is seen at the expected cutoff $k = k_1$, followed by a rapid decline in the ratio $p(k)/p_\infty(k)$ which indicates theory no longer holds.

2.5 Phase 2: Pure Random Attachment Π_{rnd}

2.6 Random Attachment Theoretical Derivations

2.6.1 Degree Distribution Theory

In pure random attachment, $\Pi = \Pi_{rnd}(k) = 1/(N(t)) \propto 1$ is constant, meaning all nodes are equally likely to be chosen. Substituting Π into equation 4 gives

$$p_\infty(k) = mN(t)\frac{1}{N(t)}p_\infty(k-1) - mN(t)\frac{1}{N(t)}p_\infty(k) + \delta_{km}. \tag{20}$$

For $k > m$, $\delta_{km} = 0$, equation 20 becomes

$$p_\infty(k) = \frac{m}{m+1}p_\infty(k-1). \tag{21}$$

Through induction (i.e. substituting $k = m+1$ and then $k = m+c$ where c is an arbitrary positive integer), equation 21 becomes

$$p_\infty(k) = \Big(\frac{m}{m+1}\Big)^{k-m} p_\infty(m). \tag{22}$$

For $k = m$, $\delta_{km} = 1$, equation 20 becomes

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

where equation 8 is used. Substituting this into equation 23 gives the asymptotic degree distribution for random attachment

$$p_\infty(k) = \frac{m^{k-m}}{(m+1)^{k-m+1}} \qquad \text{for} \quad k \geq m. \tag{24}$$

Checking $p_\infty(k)$ has the correct normalisation,

$$\begin{aligned}
\sum_{k=1}^{\infty} p_\infty(k) &= \sum_{k=m}^{\infty} \frac{m^{k-m}}{(m+1)^k - m + 1} && \text{Since } p_\infty(k|k<m) = 0) \\
&= \frac{m^{-m}}{(m+1)^{1-m}}\Big[\sum_{k=0}^{\infty}\frac{m^k}{(m+1)^k} - \sum_{k=0}^{m-1}\frac{m^k}{(m+1)^k}\Big] && \text{Split the sum} \\
&= \frac{m^{-m}}{(m+1)^{1-m}}\Big[\frac{1}{1-m/(m+1)} - \frac{1-(m/m+1)^m}{1-m/(m+1)}\Big] && \text{Geometric series} \\
&= 1,
\end{aligned}$$

as required for a probability distribution.

2.6.2 Largest Degree Theory

As in section 2.4.1, the largest degree is obtained by solving equation 18. Using $p_\infty(k)$ from equation 24, equation 18 becomes

$$\begin{aligned}
\frac{1}{N} &= \sum_{k=k_1}^{\infty} \frac{m^{k-m}}{(m+1)^{k-m+1}} \\
\frac{m^m}{N(1+m)^{m-1}} &= \sum_{j=0}^{\infty} \Big(\frac{m}{m+1}\Big)^{k_1+j} && \text{Change summation index } j = k - k_1 \\
&= \Big(\frac{m}{m+1}\Big)^{k_1} \sum_{j=0}^{\infty} \Big(\frac{m}{m+1}\Big)^{j} \\
&= \Big(\frac{m}{m+1}\Big)^{k_1} \frac{1}{1-m/(m+1)} && \text{Geometric series} \\
\Big(\frac{m}{m+1}\Big)^{k_1} &= \frac{m^m}{N(1+m)^m} \\
k_1 \ln \frac{m}{m+1} &= m \ln \frac{m}{m+1} - \ln N && \text{Taking ln of both sides.}
\end{aligned}$$

Rearranging for k_1 gives

$$k_1 = \frac{lnN}{ln(m+1) - \ln m} + m. \tag{25}$$

This shows that $k_1 \propto \ln N$ for large N in random attachment (not a power law as in preferential attachment).

2.7 Random Attachment Numerical Results

2.7.1 Numerical Results for Largest Degree

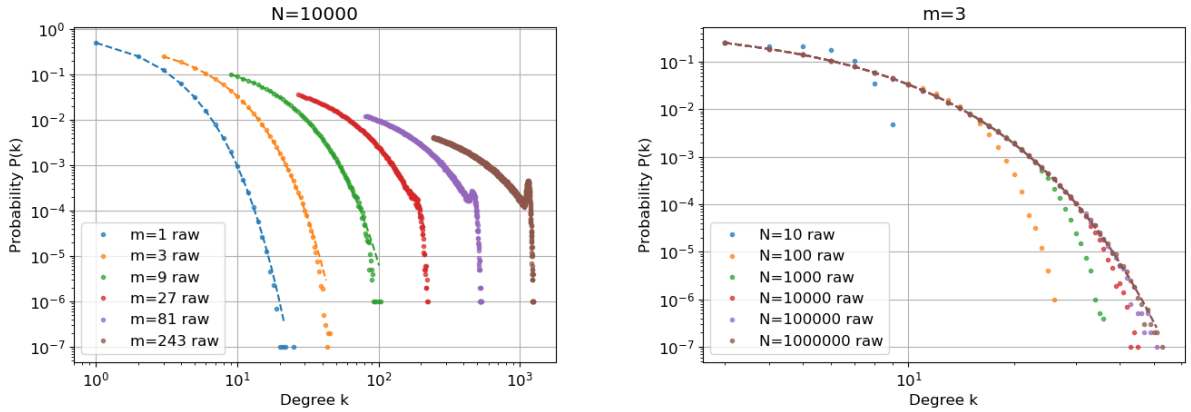
Figure 3a below shows $p(k)$ against k on a log-log scale for various m values and for $N = 10^4$. The numerical data did not show signs of fat-tailed distribution which is consistent with the theory since $p_\infty(k)$ is not a power law. Thus, the data was plotted raw without any log-binning. The bumps on the tail are only visible from $m = 27$ which imply that finite size effects are less significant in random attachment than in preferential attachment. For fix m and varying N as shown in figure 3b, the cutoff k_1 increases with N which is consistent with theory.

Visually, the numerical data appears to follow the theory very well below the bump. This is further supported statistically by a very high average p-value from the KS test (see table 1).

Data collapse was performed using the same method as before. This is shown below in figure 4a. The cut-off of $N > 10$ occurs at $k = k_1$ as expected, whereas $N = 10$ did not. Below the cut-off, $p(k)/p_\infty(k)$=1 for all N except $N = 10$, which has a visible bump well before $k = k_1$ due to finite-size effects being more significant.

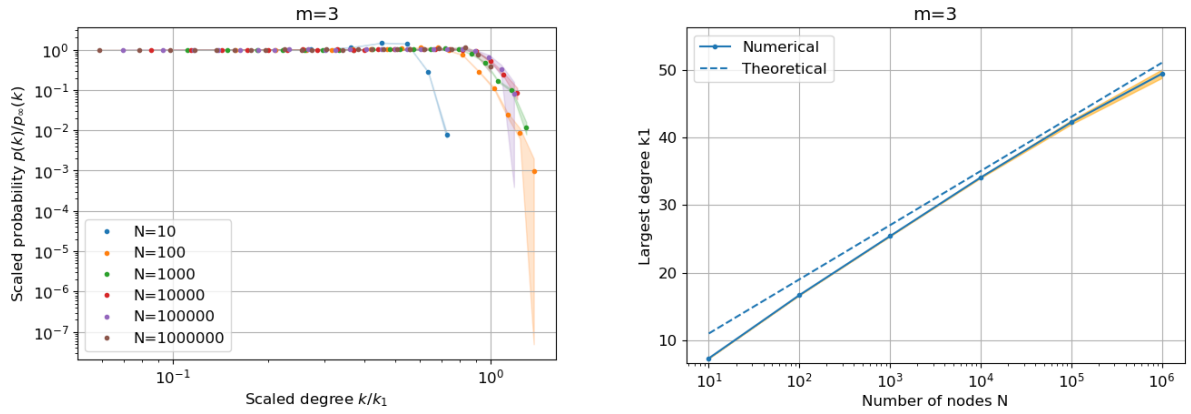
2.7.2 Largest Degree Numerical Results

The plot of k_1 against N is shown in figure 4b with log scale on x-axis and linear scale on y-axis. A straight line is observed as expected from equation 25 which showed $k_1 \propto \ln N$



(a) (b)

Figure 3: a) The raw degree distribution for random attachment for varying m and fixed $N = 10^4$. The dashed lines indicate the theoretical power-law behaviour which agree with numerical data up to the characteristic bump at large k due to a finite graph in the simulation. For large m, the bump is much more prominent. b) The degree distribution for preferential attachment, plotted with errors (filled-in colour), for varying N and fixed $m = 3$. The cut-off degree moves to the right and becomes less visible as N increases



(a) (b)

Figure 4: a) Data collapsed degree distribution for random attachment, plotted with errors (filled-in colour), for fixed $m = 3$ and varying N. The probability is scaled by the theoretical distribution collapsing to 1 for all N for $k < k_1$. Close to $k = k_1$, finite-size effects are revealed as a bump followed by rapid decay. b) A plot of k_1 against N, with error filled-in, showing good agreement between theory and numerical results when considering large N. The slope of the dashed line is 3.48.

for large N. The `scipy.stats.linregress` function was used to do a linear regression on the numerical data giving a slope of 3.67 ± 0.08, which is very close to the theoretical slope of $(\ln m + 1 - \ln m)^{-1} = 3.48$ for $m = 3$, although they do not agree within error. However, when only considering data points with $N > 100$ to minimise the finite-size effects, the numerical slope becomes 3.48 ± 0.10 which agrees with theory slope exactly.

3 Phase 3: Existing Vertices Model

3.1 Existing Vertices Model Theoretical Derivations

In existing vertices model, $\Pi = \Pi_{rnd}(k) = 1/N$ for $r = m/2$ edges where both ends are attached to existing nodes, whilst $\Pi = \Pi_{pa}(k) = k/(2E)$ for the remaining half of edges with one end to the new node and the other to an existing node. Following the same procedure as in preferential attachment and substituting this definition of Πinto equation 4 gives

$$\begin{aligned} p_\infty(k) = &\frac{m}{2}\Pi_{pa}(k-1)N(t)p_\infty(k-1) - \frac{m}{2}\Pi_{pa}(k)N(t)p_\infty(k) \\ &+ 2\times\frac{m}{2}\Pi_{rnd}(k-1)N(t)p_\infty(k-1) - 2\times\frac{m}{2}\Pi_{rnd}(k)N(t)p_\infty(k) \\ &+ \delta_{k,m/2}. \end{aligned} \tag{26}$$

where the factor of 2 in the second line is due to contribution to the degree of 2 existing nodes. For $k > m$, $\delta_{k,m/2} = 0$. Furthermore, using equation 7, equation 26 can be simplified to

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k+4m-1}{k+4m+4}, \tag{27}$$

which takes the form of a difference equation (eqn. 9) with $a = 4m-1$ and $b = 4m+4$. Using the Gamma function property (eqn. 10),

$$\begin{aligned} p_\infty(k) &= A\frac{\Gamma(k+4m)}{\Gamma(k+4m+5)} \\ &= A\frac{(k+4m-1)!}{(k+4m+4)!} \\ &= \frac{A}{(k+4m+4)(k+4m+3)(k+4m+2)(k+4m+1)(k+4m)}. \end{aligned} \tag{28}$$

For $k = m/2 = r$, $\delta_{k,m/2} = 1$. Making use of equations 7 and 8, equation 26 becomes

$$\begin{aligned} p_\infty(r) &= -r\Pi_{pa}(r)N(t)P_\infty(r) - m\Pi_{rnd}(r)N(t)P_\infty(r) + 1 \\ &= \frac{8}{8+9m}. \end{aligned} \tag{29}$$

Substituting $k = m/2 = r$ into equation 28 and equate with equation 29 gives the constant A to be

$$A = 4(9r+3)(9r+2)(9r+1)(9r). \tag{30}$$

Thus, the asymptotic degree distribution for preferential attachment is

$$p_\infty(k) = \frac{4(9r+3)(9r+2)(9r+1)(9r)}{(k+4m+4)(k+4m+3)(k+4m+2)(k+4m+1)(k+4m)} \quad \text{for} \quad k \geq \frac{m}{2}. \tag{31}$$

In the large k limit, $p_\infty(k) \propto k^{-5}$.

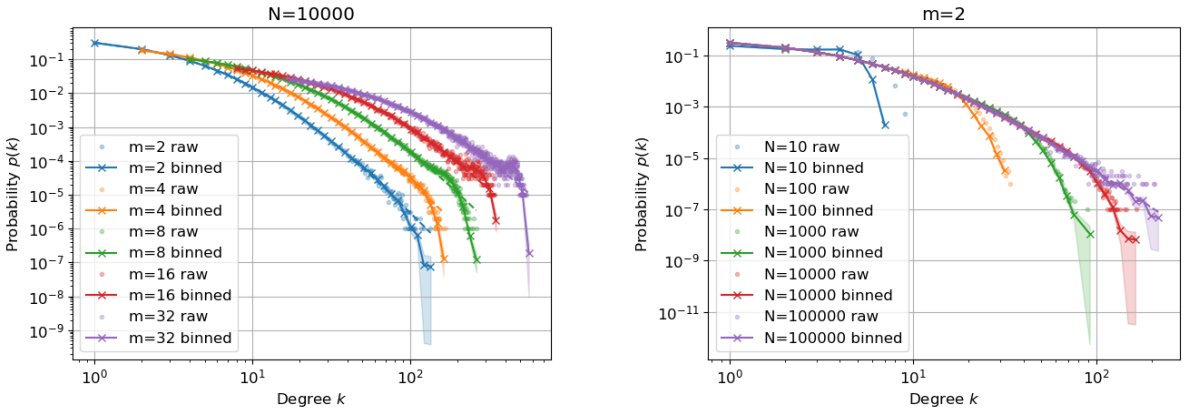
The normalisation of $p_\infty(k)$ is satisfied when writing $p_\infty(k)$ as partial fractions and cancelling terms in the sum

$$\begin{aligned}\sum_{k=1}^{\infty} p_\infty(k) &= A \sum_{k=m/2=r}^{\infty} \Big(\frac{1}{24(k+4m+4)} - \frac{1}{6(k+4m+3)} + \frac{1}{4(k+4m+2)} - \frac{1}{6(k+4m+1)} \\ &\qquad + \frac{1}{24(k+4m)} \Big) \\ &= A\Big(- \frac{1}{24(9r+3)} + \frac{1}{8(9r+2)} - \frac{1}{8(9r+1)} + \frac{1}{24(9r)} \Big) \\ &= 1 \end{aligned} \tag{32}$$

as required for a probability distribution. The theoretical expression for the largest degree was also found using Wolfram Alpha (see code for detail).

3.2 Existing Vertices Model Numerical Results

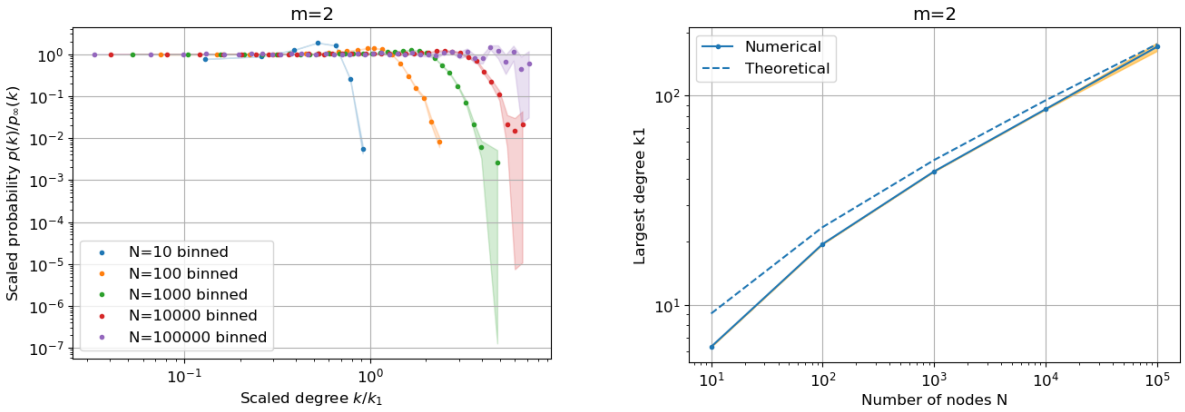
3.2.1 Numerical Results for Largest Degree



(a) (b)

Figure 5: a) The binned degree distribution for existing vertices model, plotted with errors (filled-in colour), for varying m and fixed $N = 10^4$. The dashed lines indicate the theoretical power-law behaviour which agree with numerical data up to the characteristic bump at large k due to a finite graph in the simulation. For large m, the bump is much more prominent. b) The degree distribution for preferential attachment, plotted with errors (filled-in colour), for varying N and fixed $m = 3$. The bump moves to the right and becomes less visible as N increases.

Figure 5a below shows $p(k)$ against k on a log-log scale for various m values and for $N = 10^4$. The numerical data does show signs of fat-tailed distribution which is consistent with the theory since $p_\infty(k)$ tends to a power law k^{-5} for large k. However, it is not a straight line like in preferential attachment. This shows that the model contains a mixture of preferential and random attachment. The bumps on the tail are visible from $m = 4$ which imply that finite size effects are quite significant in this model as for the case of preferential attachment. For fix m and varying N as shown in figure 5b, the cutoff k_1 increases with N which is consistent with theory (see code).



(a) (b)

Figure 6: a) Data collapsed degree distribution for existing vertices model, plotted with errors (filled-in colour), for fixed $m = 3$ and varying N. The probability is scaled by the theoretical distribution collapsing to 1 for all N for $k < k_1$. The bumps and rapid decay are quite spread around $k = k_1$, indicating an incorrect k_1 for this model. b) A plot of k_1 against N, with error filled-in, showing inconsistency between theory and numerical results.

Visually, the numerical data appears to follow the theory very well below the bump. This is further supported statistically by a very high average p-value from the KS test (see table 1).

Data collapse was performed using the same method as before. This is shown below in figure 6a. Below the cut-off for each respective line, $p(k)/p_\infty(k)$=1 for all N, which indicates that $p_\infty(k)$ is valid for this model. Crucially, the bumps all occur at different positions indicating that the theory value of k_1 is not valid. The bump is definitely more visible for smaller values of N which is consistent with finite-size effects.

4 Conclusions

Overall, the numerical degree distributions for the preferential, random and existing vertices attachment models are consistent with their corresponding asymptotic theoretical distributions for varying m and N. This is supported by visual confirmation and statistically by KS test. The numerical largest degree show good agreement with the theoretical variation with N, at least for the slope. However, there is an offset that is likely dependent on m. This is not explored in the present work. Finally, finite-size effects can be revealed more clearly through data collapse of the degree distribution.

References

[1] Barabsi, A.-L., and Albert R. (1999), 'Emergence of scaling in random networks', *Science* 286, 509-512.

[2] Evans, T. (2019), *Networks Lecture Course Notes*, Physics department, Imperial College London.

[3] Evans, T. (2019). *Networks Problem Sheet 2*: Random Networks, Physics department, Imperial College London.

[4] Barabsi, A.-L. (2016), *Network science*, Cambridge university press.

[5] K.Christensen and N.Maloney, *Complexity and Criticality*, Imperial College Press, London, 2005.

[6] Clauset, A. et al., (2009), Power-law distributions in empirical data, *SIAM review* 51(4), 661-703.