Network Project
A Growing Network Model

CID: 01331868
Shaun Fendi Gan
March 26, 2020

Abstract: The Barabasi-Albert Model was investigated with three attachment models: preferential attachment, random attachment, and random walk with preferential attachment. Theoretical derivations with experimental tests were carried out for the degree distribution and largest degree for random and preferential attachment. A two-sample Kolmogorov–Smirnov test was used to test goodness of fits, with log-binning used to remedy fat-tails. Additionally, a data collapse of varying N was done for preferential attachment. For random walk, different q values were tested to show the progression of the degree distribution from random to preferential attachment. It was found for $N = 1000$, $m = 2$ with no repeats, the distribution transitioned to random at $q = 0.30 \pm 0.02$ and transitioned to preferential at $q = 0.76 \pm 0.02$.

Word Count: 2499

1 Introduction

Networks provide a method to understand complex systems as they demonstrate the interactions between the system's components. Networks manifest in cells, social mediums and in communication technology and as Barabási describes: "we will never understand complex systems unless we develop a deep understanding of the networks behind them". In this project, the Barabasi-Albert Model(BA Model), random attachment model, and random walk with preferential attachment model are investigated.

1.1 Definition

The Barabasi-Albert Model implements growth over time and preferential attachment to reflect real world networks. This can generate scale-free networks. Preferential attachment means new nodes are more likely to connect to existing nodes with a large number of links or degrees. In this model, it is defined that:

- Multi-loops are not allowed
- Self-loops are not allowed
- A node can't have a negative number of degrees
- A node can't have zero links
- A node must have $k \geq m$ degrees

where k is the number of degrees and m is the number of links added for every new node added.

2 Phase 1: Pure Preferential Attachment Π_{pa}

2.1 Implementation

2.1.1 Numerical Implementation

Numerically, the probability of a new node attaching to an existing node Π_{pa} was tracked using a preferential list. This was implemented such that when a pair of nodes connect, both node numbers would be appended to this list. A node with a larger degree occurs more often in this list, increasing the probability of picking that node. A random choice was made to pick nodes from the preferential list. This method was used as it was efficient.

Self-loops and multi-loops were disallowed by storing a list of disallowed vertices. From a complete graph starting with m_0 nodes at $t = 0$ where all nodes are connected together,

1. Initialise a preferential list with the starting nodes, containing $m_0 = m$ occurrences of each starting node.
2. At time t, add a node of number: $t + m_0$
3. Initialise a list of disallowed vertices containing the node itself - preventing self-loops.
4. For m links to be added
   (a) choose a node from the preferential list
   (b) while the chosen node is in list of disallowed vertices, randomly pick another node from the preferential list
   (c) connect nodes
   (d) append connected nodes to preferential list and to list of disallowed vertices
5. Repeat step 2-4 until specified end time.
6. Return a table of each node number and its degree count

2.1.2 Initial Graph

Graphs were initialised using complete graphs with $m_0 = m$ starting nodes all connected to each other. This was chosen as it prevented from generating artificial hubs that would bias the system. Complete graphs of $m_0 > m$ were avoided to limit the effects of initial conditions.

2.1.3 Type of Graph

A simple undirected graph was used. This gave a clearer overall picture of the network structure at large system sizes compared to weighted graphs. The BA model is also a critical point network, between the scale-free and random regime [1].

2.1.4 Working Code

Code was checked by testing if the graph's properties followed theoretical predictions

- Number of nodes: $N_{\text{nodes}} = t + m_0$
- Number of links: $N_{\text{links}} = \frac{1}{2}(m_0 - 1)m_0 + mt$
- Number of elements in preferential-list: $N_{\text{pref}} = 2 \times N_{\text{links}}$
- Limit of average degree size $\lim_{t\to\infty} < k >= 2m$

2.1.5 Parameters

- t, time or number of nodes
- m, number of nodes added at each timestep
- m_0, initial number of nodes. By default, $m_0 = m$ unless otherwise specified.
- $draw$ = True, creates NetworkX plot for visualisation

2.2 Preferential Attachment Degree Distribution Theory

2.2.1 Theoretical Derivation

Beginning from the master equation for networks of any attachment model,

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

where $N(k, t)$ is the number of nodes with degree k at time t, Π is the probability of picking a node to attach to and $\delta_{k,m}$ is a dirac delta function that counts the new node being added at $k = m$. This equation denotes that the number of nodes with k at $t + 1$ is equal to the number at the current time, plus nodes that change to and away from k.

For preferential attachment, the probability Π of picking a node is proportional to the degree of the picked node normalised across the system

$$\Pi_{\text{pa}}(k) = \frac{k}{\sum_j k_j} \approx \frac{k}{2mt}, \tag{2}$$

for j nodes, using small m_0 initial conditions over large time. Additionally, the probability of $p_k(t)$ picking a node with degree k is as follows

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

where $N(t)$ is the total number of nodes at time t.

Substituting Eq.(2) and Eq.(3) with the master equation Eq.(1), the following cases are obtained

$$(N+1)p_k(t+1) = \begin{cases} Np_k(t) + \frac{k-1}{2}p_{k-1}(t) - \frac{k}{2}p_k(t), & \text{if } k > m \\ Np_m(t) + 1 - \frac{m}{2}p_m(t), & \text{if } k = m \\ 0, & \text{if } k < m \end{cases} \tag{4}$$

defining all situations. For $k < m$, this is not allowed and thus is zero.

By considering the limit as $t \to \infty$, it is assumed that $p_k(\infty) = p_k$. This considers the probability of k at all times are equal, and thus $p_k(t+1) = p_k(t) = p_k$. This causes cancellations in the situations of Eq.(4) such that

$$p_k = \begin{cases} \frac{k-1}{k+2} p_{k-1}, & \text{if } k > m \\ \frac{2}{m+2}, & \text{if } k = m \\ 0, & \text{if } k < m \end{cases} \tag{5}$$

where p_k now describes the probability of k at all times $t > 0$. By observing the subsequent terms

$$p_{m+1} = \frac{m}{m+3} p_m = \frac{2m}{(m+2)(m+3)}, \tag{6}$$

$$p_{m+2} = \frac{m+1}{m+4} p_{m+1} = \frac{2m(m+1)}{(m+2)(m+3)(m+4)}, \tag{7}$$

$$p_{m+3} = \frac{m+2}{m+5} p_{m+2} = \frac{2m(m+1)\cancel{(m+2)}}{\cancel{(m+2)}(m+3)(m+4)(m+5)}, \tag{8}$$

where Eq.(7) is relabelled as $k \to m+2$, and Eq.(8) is relabelled as $k \to m+3$. This produces a generalised recursive relation as

$$p_k(\Pi_{\text{pa}}) = \frac{2m(m+1)}{k(k+1)(k+2)}, \text{where } k \geq m. \tag{9}$$

giving the discrete degree distribution for preferential attachment.

2.2.2 Theoretical Checks

Starting from the Eq.(9), a sum for all possible discrete k values can be taken to show that the probability is normalised. The sum is performed by method of partial fractions

$$\sum_{k=m}^{\infty} p_k = \sum_{k=m}^{\infty} \frac{2m(m+1)}{k(k+1)(k+2)}, \tag{10}$$

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

$$\sum_{k=m}^{\infty} p_k = m(m+1) \left[\sum_{k=m}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1} \right) - \sum_{k=m}^{\infty} \left(\frac{1}{k+1} - \frac{1}{k+2} \right) \right], \tag{11}$$

where in this specific sum, terms in between the first and last term cancel. This gives the following as the reciprocal of infinity tends to zero

$$\sum_{k=m}^{\infty} p_k = m(m+1) \left[\frac{1}{m} - \frac{1}{m+1} \right], \tag{12}$$

$$\sum_{k=m}^{\infty} p_k = \cancel{m(m+1)} \left[\frac{m+1-1}{\cancel{m(m+1)}} \right] = 1. \tag{13}$$

$$\tag{14}$$

showing probability as normalised.

Additionally as $t \to \infty$, $k \to \infty$, causing the discrete distribution to tend to the continuous at large k, i.e. $p_k \to k^{-3}$.

2.3 Preferential Attachment Degree Distribution Numerics

2.3.1 Fat-Tail

Fat-tailed characteristics are degree distributions that follow a power law, which causes scale-free networks. Mathematically, a fat-tail has a probability p_k that decays slower than an exponential [2]. This leads to a small but significant frequency and probability at large k, representing hubs. This region has insufficient data points to extract true behaviour, causing data points of the same probabilities to span orders of magnitude in k. Thus, log-binning can be used to remedy this.

Log-binning creates bins that scale exponentially with k as

$$\frac{b_{i+1}}{b_i} = \exp\{s\} \tag{15}$$

where b_i is the lower bound of a bin, b_{i+1} is the upper bound of a bin, and s is the scale. Log-binning produces a new probability from the frequency in each bin divided by its bin width. This probability $\tilde{p_k}$ is not equivalent to p_k, but represents the same information. A scale of $s = 1.2$ was chosen for preferential attachment, compromising between smoothness and exhibiting characteristics. This procedure is shown in Fig. 2.

2.3.2 Numerical Results

Fig. 1 shows the degree distribution p_k with k for $N = 100000$ and $m = [2,4,8,16,32,64,128]$. The BA model was run 100 times for repeats, where the degree counts of each run were collated into one list that was log-binned to give $\tilde{p_k}$.

Generally observed values of k follows the predicted distribution. The first point for $m \geq 16$ appears to deviate away from the expected distribution, but could be fixed by changing log-bin scale.

Deviations at small p_k are attributed to finite size system effects, where nodes that could reach larger degrees in infinite systems are cut off early and thus condensing those probabilities into the bump.

Errors were found by taking the standard deviations of the different probabilities at each k, across 100 repeats. This was divided by the square root of the number of repeats to obtain standard error on the mean. If a specific repeat did not have a p_k value at a specific k, $p_k = 0$ was appended. This increases uncertainty at larger k as their occurrences become unlikely.

2.3.3 Statistics

A two-sample Kolmogorov–Smirnov (K-S) test was employed to test how data fits theoretical predictions. The test finds the absolute maximum difference between the expected and observed data at all points on a cumulative distribution

$$D = \sup_k \left| F_{1,n}(k) - F_{2,l}(k) \right|, \tag{16}$$

where D is the K-S statistic, sup is supremum, n, l sample sizes (unlog-binned) and F_1, F_2 are the two respective cumulative distributions. The null hypothesis is that the two distributions are derived from the same function, and it can be rejected with confidence α if

$$D > \frac{1}{\sqrt{n}} \cdot \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1 + \frac{n}{l}}{2}}. \tag{17}$$

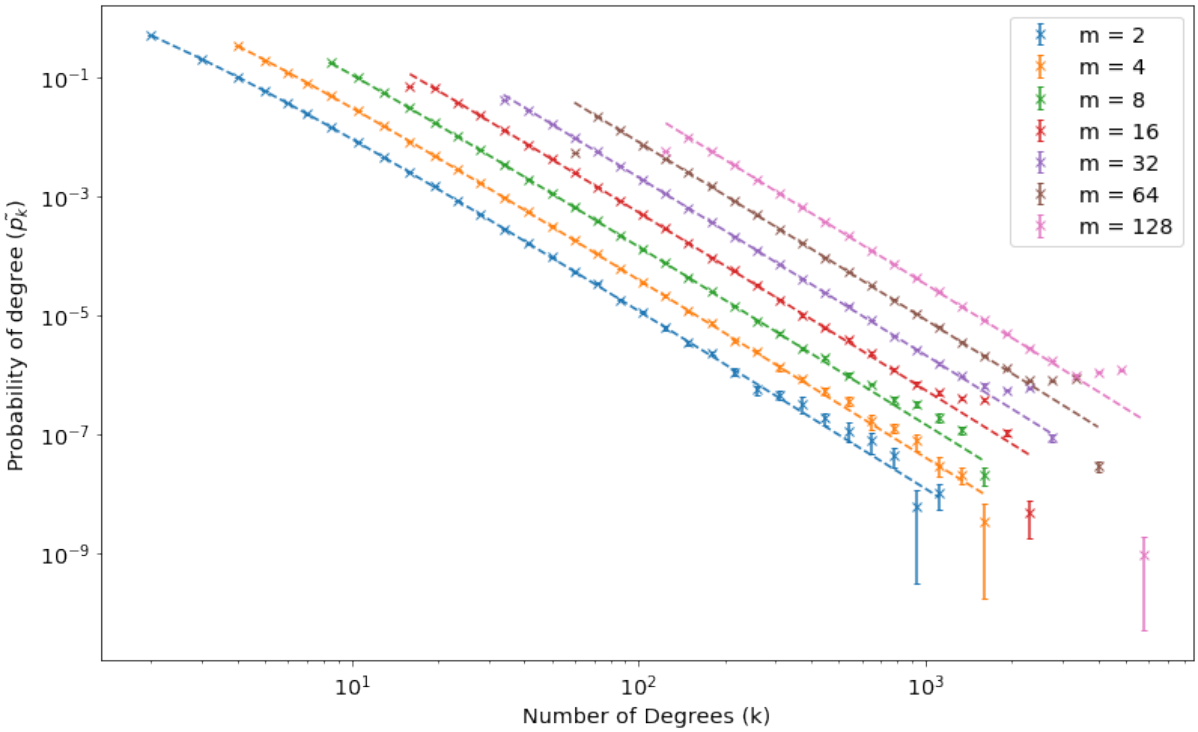

Figure 1: Plot shows probability of degree against number of degrees k, for the BA Model with preferential attachement. Crosses are observed points, whereas the dashed lines represent the theoretical equation Eq. (9). The distance of points on the bump to the theoretical is larger than the error, indicating it as not random error.

where α is the p-value and the RHS can be thought of as a threshold [3].

By converting probability degree distributions to cumulative distributions, different D values were found for different m and are seen in Table 1. A p-value of 0.05 was used as a standard measure to determine statistical significance. Results indicate that for $m = 2, 4$, the null hypothesis cannot be rejected, implying they follow the preferential attachment distribution with a probability higher than 5%.

For $m = [8, 16, 32, 64, 128]$, the null hypothesis can be rejected as D is larger than the threshold, implying different distributions. This is not as expected and is believed to be due to the more significant finite size bump at the end of the distribution for larger m.

This was chosen over chi-square test as the K-S test does not assume that sample follows a normal distribution, whereas chi-square does [4]. Additionally, the chi-square test averages across the separations of all points, indicating a good fit even with anomalies.

2.4 Preferential Attachment Largest Degree and Data Collapse

2.4.1 Largest Degree Theory

Similarly to Eq.(11) when checking for normalisation, the probability of the largest node can be found by considering the sum from k_1 instead of m to ∞. This encapsulates all possible probabilities for degrees $k \geq k_1$ where k_1 is the number of degrees on the largest node. Additionally, since only one node has the max degrees, this probability is equal to

m	D	Thresholds ($\alpha = 0.05$)	Δ
2	0.000344	0.001921	-0.001576
4	0.001257	0.001921	-0.000663
8	0.004002	0.001921	0.002081
16	0.043532	0.001920	0.041611
32	0.008820	0.001920	0.006900
64	0.031682	0.001920	0.029762
128	0.011195	0.001919	0.009275

Table 1: K-S statistic values for different m and $N = 100,000$ were found for preferential attachment, with their thresholds representing the RHS of Eq. (17). The p-value α is 0.05. Δ represents the difference of D and the threshold, a negative Δ indicates that the null hypothesis is accepted. It is seen that for values of $m \geq 8$ in the table, the null hypothesis is rejected indicating different distributions, this is believed to be due to the finite size bump.

picking one node

$$\sum_{k=k_1}^{\infty} p_{k_1} = m(m+1)\left[\sum_{k=k_1}^{\infty}\left(\frac{1}{k} - \frac{1}{k+1}\right) - \sum_{k=k_1}^{\infty}\left(\frac{1}{k+1} - \frac{1}{k+2}\right)\right] = \frac{1}{N}, \tag{18}$$

where the terms in between the first and last k values cancel with the limit of $\frac{1}{\infty} \to 0$, this gives

$$\begin{aligned} m(m+1)\left[\frac{1}{k_1} - \frac{1}{k_1+1}\right] &= \frac{1}{N}, \\ k_1^2 + k_1 - Nm(m+1) &= 0, \end{aligned} \tag{19}$$

where this form is solved as a quadratic equation. Thus k_1 is governed by

$$k_1 = -\frac{1}{2} + \frac{\sqrt{1 + 4Nm(m+1)}}{2}. \tag{20}$$

where only the positive solution was taken. It should be noted that this form is expected to work for $N \gg 1$, as evidently $k_1 > m$ and thus $k_1 \neq -\frac{1}{2}$.

2.4.2 Numerical Results for Largest Degree

Times of $N = [10^1, 10^2, 10^3, 10^4, 10^5, 10^6]$ and $m = 2$ were chosen. This is because $m \leq 0$ does not grow the network, $m = 1$ is a special network (i.e. tree), and choosing large m would require the network to grow over large times to reduce the effects of the initial conditions.

Fig. 2 shows the distribution of the maximum degree k_1 against N. One hundred repeats were taken, where 100 k_1 values were averaged. Errors in k_1 were found by taking a standard error on the mean. It is expected as $k_1 \propto N^{0.5}$ from Eq.(20) for $N \gg m$. Analyzing the gradient, the relationship is found as

$$\log k_1 = (0.505 \pm 0.002) \log N \tag{21}$$

which is 1% away from the expected value.

Dividing k_1 by the leading scaling order reveals the separation between predicted and observed k_1. The separation is roughly constant but is larger than the errors even though the scale is small, indicating a systematic error. This could be due to the assumption that the probability of $p_k(\infty) = p_k$ not holding true at small k.

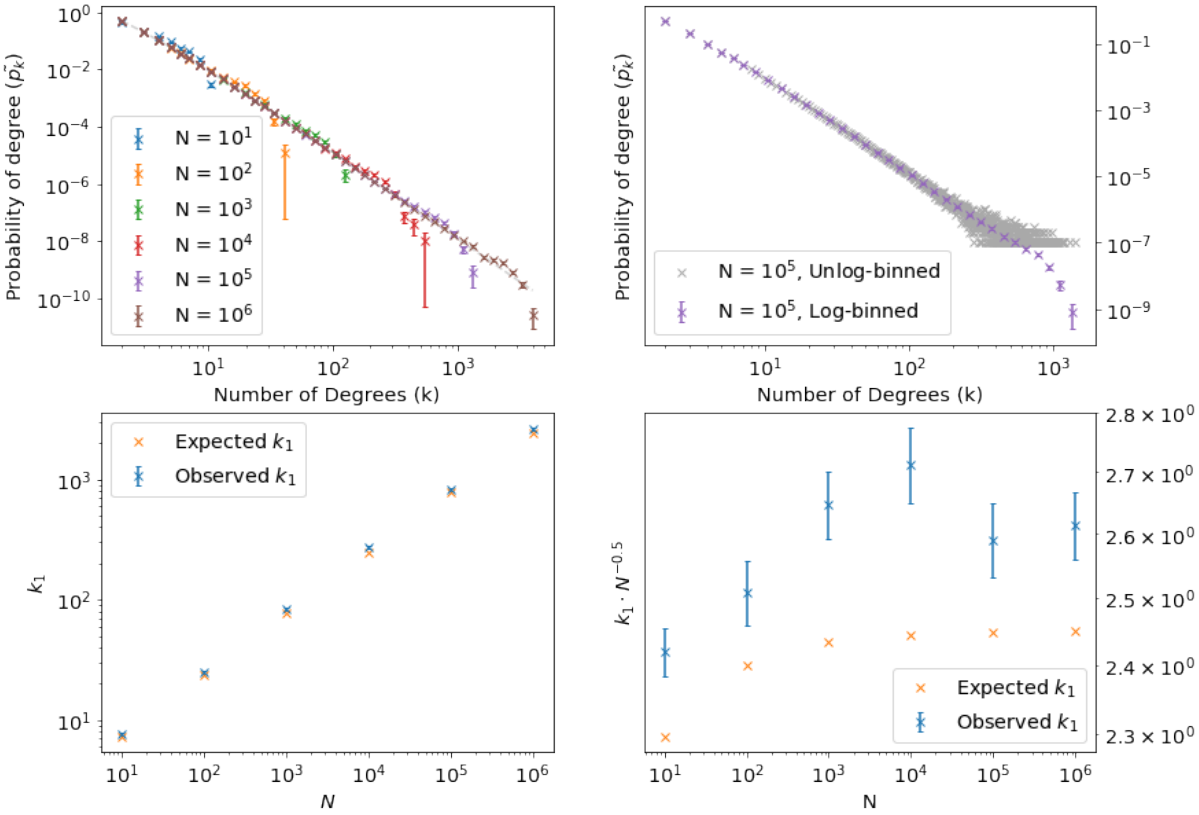

Figure 2: **Top Left**: Degree distribution of $m = 2$ for different N of nodes added. The light grey dashed line represents the theoretical solution seen in Eq. (9). **Top Right**: Shows the before and after of log-binning a sample with a fat-tail, displaying its benefit to extract statistical information. **Bottom Left**: shows the distribution of k_1 against N. **Bottom Right**: shows k_1 divided by the leading order in N to observe the differences between each points.

2.4.3 Data Collapse

Dividing k by the maximum degree k_1 will align all distributions to occur at the same k/k_1. Multiplying p_k by its dependence on k as seen in Eq. (9) will remove its dependence such that the height of each distribution is just a constant dependent on m.

As seen in Fig. 3, the data collapse remains a constant, before reaching a bump and decaying away towards $-\infty$. This is due the finite size effect. $N = 10^1$ was excluded from the data collapse as the system was not sufficiently large, causing initial conditions to affect the system, thus deviating from the data collapse. Errors were scaled in the same way as data points.

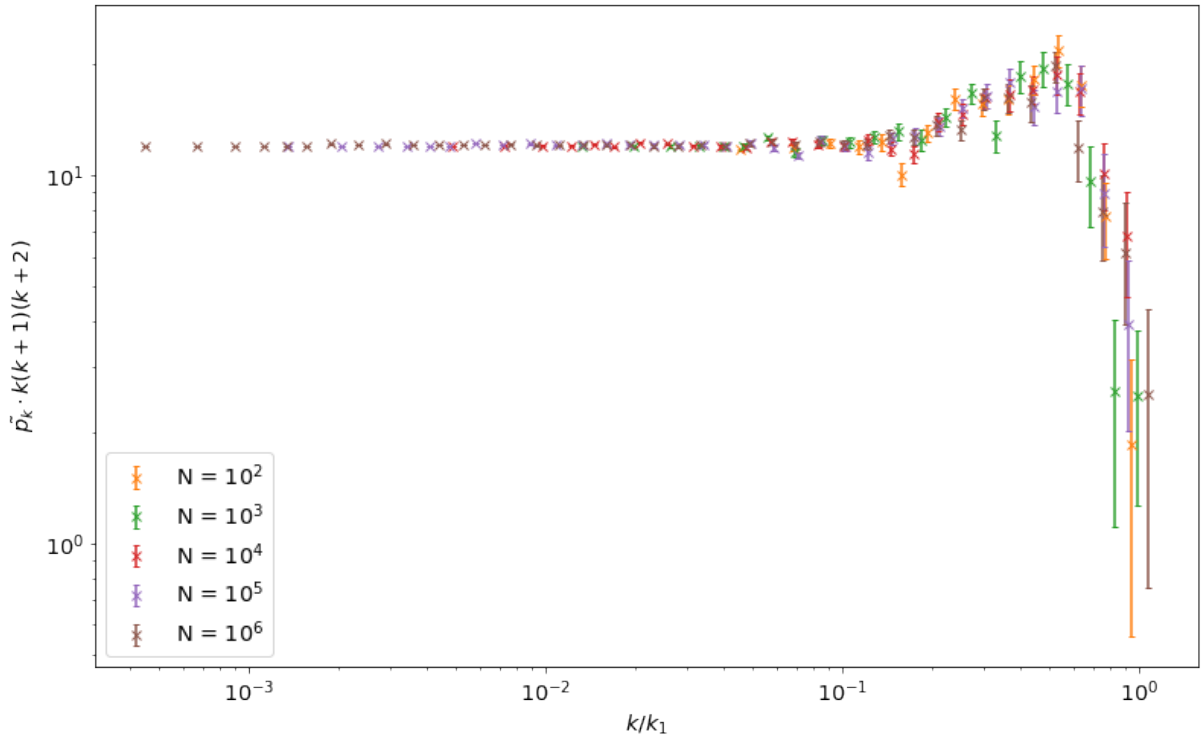

Figure 3: Data Collapse of $N = [10^2, 10^3, 10^4, 10^5, 10^6]$ for $m = 2$ using 100 repeats. $N = 10^1$ was excluded from the data collapse as the system was not sufficiently large, thus deviating from the collapse. Error bars were scaled in the same way as data points.

3 Phase 2: Pure Random Attachment Π_{rnd}

3.1 Random Attachment Theoretical Derivations

3.1.1 Degree Distribution Theory

Derivation of the degree distribution follows from the master equation Eq.(1), with similar steps as in section 2.2.1. The probability of a new node attaching to other nodes is now

$$\Pi_{\text{rnd}} = \frac{1}{N(t)}, \tag{22}$$

where all nodes are equally likely. Applying Eq.(3) and Eq.(22) to the master equation, the situations

$$(N+1)p_k(t+1) = \begin{cases} Np_k(t) + mp_{k-1}(t) - mp_k(t), & \text{if } k > m \\ Np_m(t) - mp_m(t) + 1, & \text{if } k = m \\ 0, & \text{if } k < m \end{cases} \tag{23}$$

are found. By taking $t \to \infty$, it is assumed that $p_k(\infty) = p_k$ such that probabilities at all times are equal. The situations in Eq.(24) are written as

$$p_k = \begin{cases} \frac{1}{m+1}, & \text{if } k > m \\ \frac{m}{m+1}p_{k-1}, & \text{if } k = m \\ 0, & \text{if } k < m \end{cases} \tag{24}$$

whereby observing subsequent terms

$$p_{m+1} = \Big(\frac{m}{m+1}\Big)p_m \ = \Big(\frac{m}{m+1}\Big)^1 \cdot \frac{1}{(m+1)}, \tag{25}$$

$$p_{m+2} = \Big(\frac{m}{m+1}\Big)p_{m+1} = \Big(\frac{m}{m+1}\Big)^2 \cdot \frac{1}{(m+1)}, \tag{26}$$

$$p_{m+3} = \Big(\frac{m}{m+1}\Big)p_{m+2} = \Big(\frac{m}{m+1}\Big)^3 \cdot \frac{1}{(m+1)}, \tag{27}$$

a recursive relation is seen which can be generalised as

$$p_k(\Pi_{\text{rnd}}) = \Big(\frac{m}{m+1}\Big)^{k-m} \cdot \frac{1}{m+1}, \text{ where } k \geq m. \tag{28}$$

giving a discrete degree distribution for random attachment.

This distribution must be normalised, and can be verified by taking a sum over all possible terms, that is from m to ∞. Starting from Eq.(28),

$$\begin{aligned} \sum_{k=m}^{\infty} p_k &= \sum_{k=m}^{\infty} \frac{m^{k-m}}{(1+m)^{1+k-m}}, \\ &= \frac{m^{-m}}{(m+1)^{1-m}} \sum_{k=m}^{\infty} \Big(\frac{m}{m+1}\Big)^k, \end{aligned} \tag{29}$$

presenting a geometric sum to infinity. The index k can be relabelled as $k = i + m$

$$\sum_{k=m}^{\infty} p_k = \frac{m^{-m}}{(m+1)^{1-m}} \Big(\frac{m}{m+1}\Big)^m \sum_{i=0}^{\infty} \Big(\frac{m}{m+1}\Big)^i, \tag{30}$$

where applying the geometric sum equation $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$ provides

$$\sum_{k=m}^{\infty} p_k = \frac{m^{-m}}{(m+1)^{1-m}} \Big(\frac{m}{m+1}\Big)^m \cancel{(m+1)} = 1. \tag{31}$$

showing the distribution as normalised.

3.1.2 Largest Degree Theory

Similar to the method used for preferential attachment, a sum from k_1 to ∞ across $p_k(\Pi\text{rnd})$ encapsulates the probability of the largest degree. This is equal to $\frac{1}{N}$ as it is assumed that one node has the largest degree

$$\begin{aligned} \sum_{k=k_1}^{\infty} p_k = \sum_{k=k_1}^{\infty} \frac{m^{k-m}}{(1+m)^{1+k-m}} &= \frac{1}{N}, \\ \frac{m^{-m}}{(m+1)^{1-m}} \sum_{k=k_1}^{\infty} \Big(\frac{m}{m+1}\Big)^k &= \frac{1}{N}, \end{aligned} \tag{32}$$

presenting a geometric sum to infinity. The index k can be relabelled as $k = i + k_1$

$$\frac{m^{-m}}{(m+1)^{1-m}}\Big(\frac{m}{m+1}\Big)^{k_1}\sum_{i=0}^{\infty}\Big(\frac{m}{m+1}\Big)^{i}=\frac{1}{N}, \tag{33}$$

where applying the geometric sum equation $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$ provides

$$\begin{aligned}\frac{m^{-m}}{(m+1)^{1-m}}\Big(\frac{m}{m+1}\Big)^{k_1}\cancel{(m+1)}&=\frac{1}{N},\\ \Big(\frac{m}{m+1}\Big)^{k_1-m}&=\frac{1}{N},\end{aligned} \tag{34}$$

giving an exponential relation. Rewriting with logarithms gives

$$k_1=\frac{\log\big|\big(\frac{m}{m+1}\big)^{m}\cdot\frac{1}{N}\big|}{\log\big|\frac{m}{m+1}\big|}. \tag{35}$$

as the largest degree for random attachment.

3.2 Random Attachment Numerical Results

3.2.1 Degree Distribution Numerical Results

Fig. 5 shows the degree distribution p_k with k for $N = 100{,}000$ and $m = [2,4,8,16,32,64,128]$.

Similar to preferential attachment, the observed degree probabilities agree with theoretical until it reaches degree values limited by system size. At large k relative to the system, finite size effects are seen once again. The gradient decreases for larger m values, sharing probabilities across a larger range of k. Errors were calculated using the same method as preferential attachment.

Goodness of fits were tested using K-S tests. Results of this are seen in Table 2. It is seen that for all m tested except for $m = 128$, D is smaller than the threshold and thus are good fits as they follow the same theoretical distribution. As seen on Fig. 4, $m = 128$ does not follow likely due to the finite size bump being too large.

3.2.2 Largest Degree Numerical Results

Largest degree k_1 for random attachment against number of node added is seen in Fig. 5. k_1 is divided by its dependence on N to reveal the separation between predicted and observed. The separation between observed k_1 and expected k_1 appear to reduce as $N \to \infty$, indicating the relationship between k_1 and N is consistent at large N.

Errors are calculated in the same way as for preferential attachment.

4 Phase 3: Random Walks and Preferential Attachment

4.1 Implementation

All definitions in section 1.1 apply to this model, and is initialised with a complete graph with nodes $m_0 = m$. This was implemented using an adjacency list, which records each

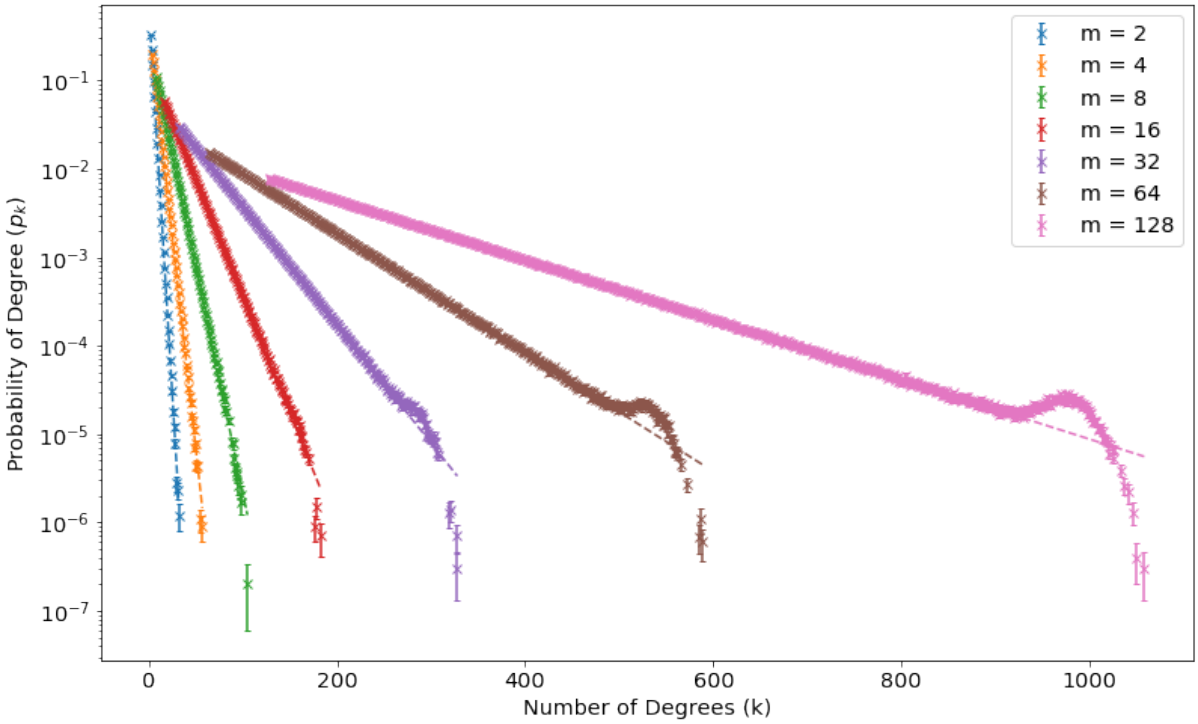

Figure 4: Plot shows probability of degree against number of degrees k, for the BA Model with random attachment. Crosses are observed points, whereas the dashed lines represent the theoretical equation Eq. (28). This distribution was linearly binned for clarity.

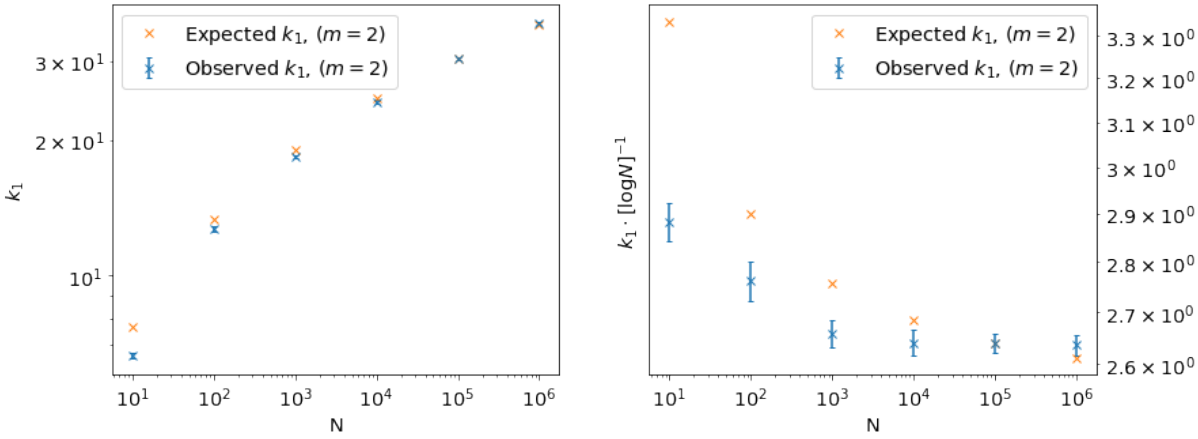

Figure 5: **Left**: Plot largest degree k_1 against number of nodes N added is shown. **Right**: k_1 is divided by its dependence on N as seen in Eq. (35) to display the separation between expected and observed clearly. Observed k_1 tends to expected k_1 for large N, indicating consistency at large N.

node and its corresponding neighbours. This was chosen as it was simple, and thus easy to debug. A probability q for the connection to walk to a neighbouring node in the network was introduced as a new parameter.

To implement the walk, a random number generator was used. While q is larger than a randomly generated number between 0 and 1, a neighbouring node will be randomly picked from the adjacency list and walked to. This is repeated until the walk ended.

New connections were then added to the adjacency list and this process of picking a

m	$D_{n,1}$	Thresholds ($\alpha = 0.05$)	Δ
2	0.000186	0.000607	-0.000421
4	0.000148	0.000607	-0.000459
8	0.000142	0.000607	-0.000465
16	0.000092	0.000607	-0.000515
32	0.000147	0.000607	-0.000461
64	0.000353	0.000607	-0.000254
128	0.000792	0.000607	0.000185

Table 2: K-S statistic values for different m and $N = 100,000$ were found when analysing random attachment. Thresholds represent the RHS of Eq. (17). The p-value α is 0.05. Δ represents the difference of D and the threshold, a negative Δ indicates that the null hypothesis is accepted, as $D <$ threshold, meaning the distributions are the same. D is smaller than the threshold for all m above except $m = 128$, due to an increasing finite size effect as seen in Fig. 4 as m increases. This result indicates most m values follow the same distribution as random attachment, as the null hypothesis cannot be rejected.

random node and walking was repeated for additional links up to m and for nodes up to N. A list of disallowed vertices was again implemented.

4.2 Numerical results

Probabilities of $q = [0, 0.5, 0.99]$ were investigated using $m = 2$ and $N = 100,000$ over ten repeats. $q = 1$ was not chosen as the walk would never end.

In Fig. 6, it can be seen that increasing the value of $q \to 1$ causes the distribution to tend to a preferential attachment.

For $q = 0$, this resembled a random distribution. This was expected as walking was impossible and nodes chosen to connect to were done so randomly. For $q = 0.5$, this represented a midpoint between random and preferential attachment. For $q = 0.99$, this was seen as preferential attachment as the observed data points matched the corresponding degree distribution seen in Eq. (9). This is expected as nodes with larger degrees provides more pathways for the a link to walk to it, increasing the likelihood of it being chosen, imitating preferential attachment.

Additional q values were investigated to measure the transition of random walk to random attachment and preferential attachment. This was done for $N = 1000$, $m = 2$ with no repeats. K-S tests were used to test if the distributions of different q values matched random or preferential attachment models. The results of this test is seen in Table 3, where the transition to random occurred at $q = 0.30 \pm 0.02$ and the transition to preferential occurred at $q = 0.76 \pm 0.02$. Fig. 7 shows network graph visualisations for the transition q values. $q = 0.30 \pm 0.02$ can be seen to resemble random attachment of having many nodes with similar degree size, and $q = 0.76 \pm 0.02$ can seen to have hubs, resembling preferential attachment.

4.3 Discussion of Results

Real-world networks are rarely perfectly preferential or random attachment. Random walks can be used to model networks that follow large-tailed distributions which lie in between with $0 \leq q < 1$,

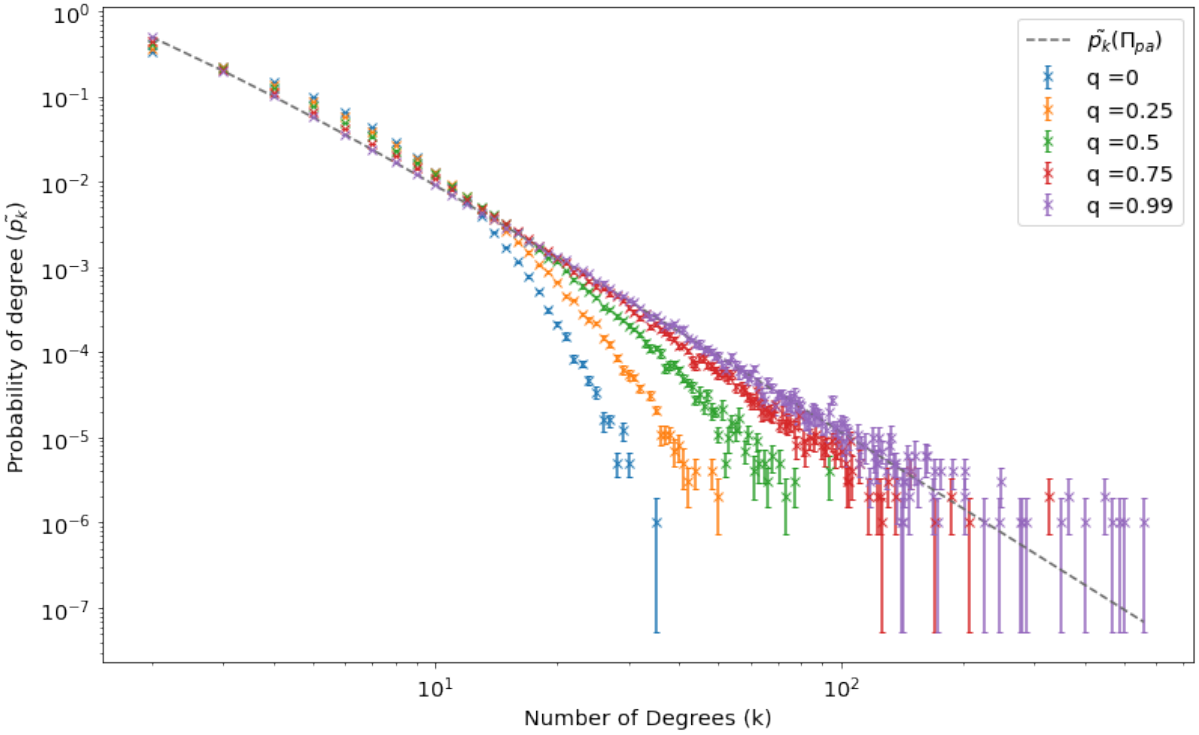

Figure 6: Degree Distribution of different random walks using different q values were tested. This shows the transition of from the random attachment model to the preferential attachment model. This is not log-binned in order to display characteristics clearly.

5 Conclusions

In conclusion, a greater understanding of the different networks that can be used to model real-world networks has been achieved.

References

[1] Barabasi, A. (2016). "Network Science". Cambridge University Press.

[2] Evans, T. (2020). "Networks Lecture Notes". Imperial College London.

[3] Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". Journal of the American Statistical Association.

[4] Cochran, W. G. (1952). "The Chi-square Test of Goodness of Fit". The Annals of Mathematical Statistics.

q	D	Thresholds ($\alpha = 0.05$)	Δ
0.00	0.007984	0.060676	-0.052691
0.30	0.044245	0.060676	-0.016431
0.31	0.060878	0.060676	0.000203
q	D	Thresholds ($\alpha = 0.05$)	Δ
0.75	0.067864	0.060676	0.007189
0.76	0.035928	0.060676	-0.024747
0.99	0.012974	0.060676	-0.047701

Table 3: K-S statistic values for different q comparing to preferential attachment. $N = 1000$, $m = 2$ was used. Thresholds represent the RHS of Eq. (17). The p-value α is 0.05. Δ represents the difference of D and the threshold, a negative Δ indicates that the null hypothesis is accepted, as $D <$ threshold, meaning the distributions are the same. **Top**: Compared to random attachment, transition occurs around $q = 0.30 \pm 0.02$, **Bottom**: Compared to preferential attachment, transition occurs around $q = 0.76 \pm 0.02$. It should be noted that using increasing values of N causes the 'transition' q to be closer to 0 or 1 for random and preferential respectively.

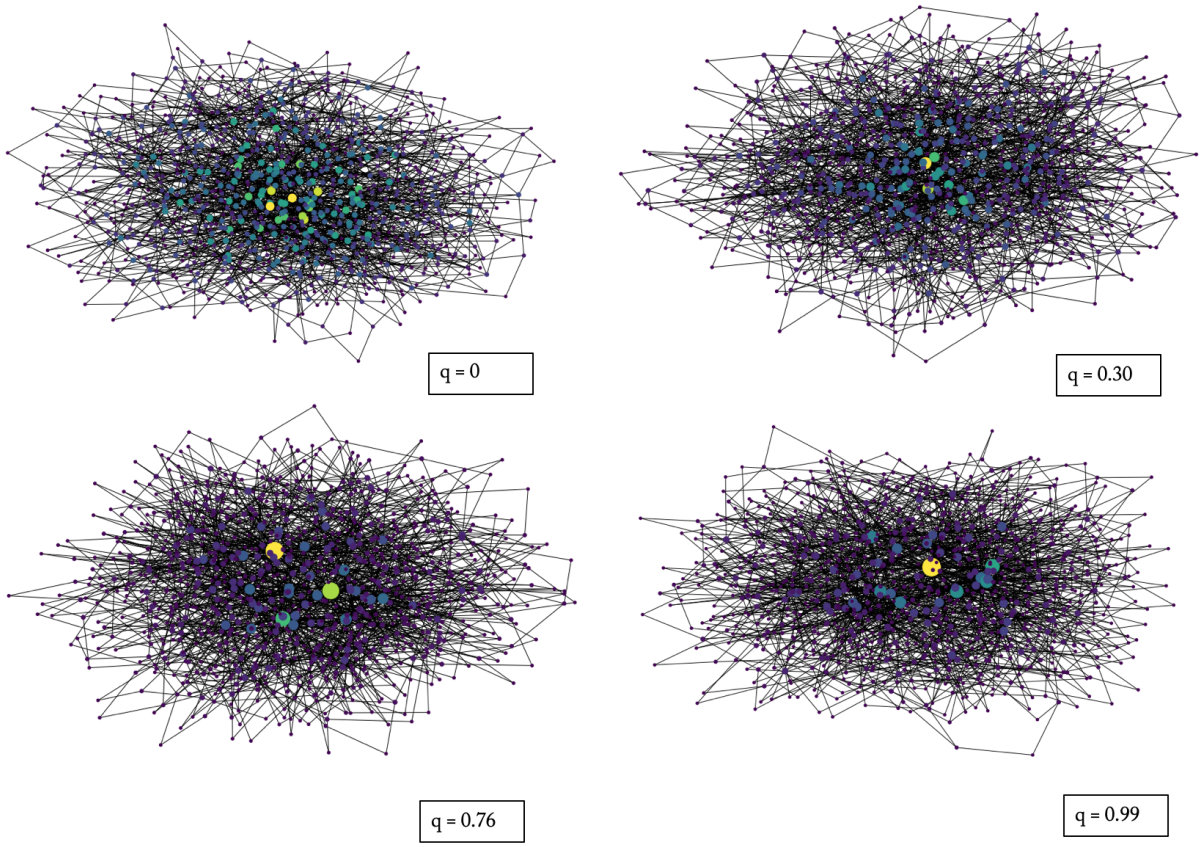

Figure 7: Different network graphs were plotted using different values of q to show the transition from random to preferential attachment. The node size on each graph is directly proportional to the number of degrees. $q = 0.30$ was shown to be random attachment as it has similar degree sizes as in $q = 0$. $q = 0.76$ was shown to be preferential attachment as it exhibited hubs as in $q = 0.99$. This was reinforced by using K-S tests in Table 3. These figures are for $N = 1000$, $m = 2$ and no repeats.