Network Project
A Growing Network Model

Shona Curtis-Walcott

23rd March 2020

Abstract: This report investigates models of a growing network using various edge attachment strategies: namely pure preferential attachment, pure random attachment, and random walk attachment. Using a comparison of predicted theory against numerical results, the maximum degree trend and degree probability distribution are studied for each attachment method. The degree distribution for the random walk model is seen to tend towards that for pure random or preferential attachment depending on a parameter q. It is found that for $k < k_1$, the numerical data for all systems matches the theoretical expectations. For $k \sim k_1$ however effects due to the finite-size of the systems become prominent.

Word Count: 2493

1 Introduction

Networks science builds predictive models for systems containing non-trivial connections between discrete elements. The models are described by graphs, where nodes represents elements and edges represent connections. A range of physical, biological, and social phenomena can be modelled using networks science, an example of which is the citation network [1]. This report explores the nature of a growing network with three different methods of attachment; preferential, random, and a random walk with preferential attachment.

1.1 Definition

The Barabasi Albert (BA) algorithm has been implemented to model a growing network. Citation networks are theorised to be scale-free [2] and thus the BA algorithm is suitable as it builds a scale-free network where nodes are connected via preferential attachment. Preferential attachment is the increased likelihood of an already highly-connected node being involved in further connections; for a citation network this would be seen as an increased likelihood of a highly-cited scientific paper being found, read, and cited by another author compared with a paper on the same topic but with very few citations.

The model was set up as follows:

1. Graph created at $t = 0$
2. Time is incremented $t \to t + 1$
3. One new node is added to the graph
4. The new node forms m edges with existing nodes, with probability Π

5. Steps $2 \rightarrow 4$ are repeated until a total of N nodes have been added to the graph

In the case of pure preferential attachment, Π will be proportional to the degree of a node k; for pure random attachment, Π will be constant.

2 Phase 1: Pure Preferential Attachment Π_{pa}

2.1 Implementation

2.1.1 Numerical Implementation

The system is initialised with $2m+1$ nodes, all with degree $2m$. Steps $2-4$ as stated in section 1.1 are looped over with $N-(2m+1)$ iterations, each time adding a node and m edges, so that the total size of the graph will be N.

A list $\mathcal{G}$ is created to represent the graph. It exists as a list of lists, where the i^{th} element corresponds to the i^{th} node of the graph, and the numbers contained in the i^{th} sublist correspond to nodes with which this i^{th} node is connected via an edge.

A second list, $\mathcal{G}_{Conn}$, is created to which the index of a node is added every time it is involved in the formation of a new edge. As such, there will be more instances of a highly-connected node in $\mathcal{G}_{Conn}$ than a node with only a few connections. Each time one new node is added, it forms m edges. The nodes to which these edges connect are selected randomly from $\mathcal{G}_{Conn}$. Given the nature of this list, the probability of connecting to a node with many existing connections is higher than that of connecting to a node with very few existing connections: this correctly reflects the preferential attachment probability Π_{pa}.

To eliminate the chance of randomly selecting the same node twice within one iteration, the chosen nodes are added to a temporary list which can be searched through using a conditional statement. This repeats until the list has a length m, indicating the correct number of edges to the new node have been formed.

2.1.2 Initial Graph

By ensuring that no node in the initial graph has a higher degree than any other, a skewness in the final graph can be avoided. To avoid multiple edges forming between a new and existing node during initialisation, it is required that for a BA system adding m edges to each new node, that the number of nodes added during initialisation $N(0)$ must satisfy $N(0) > m$. Furthermore, $N(0)$ must satisfy $N \gg N(0)$, where N is the total number of nodes added to a system overall. In the written BA model, $2m+1$ nodes were added to the graph upon initialisation, and each node was connected to every other.

Starting with $2m+1$ nodes satisfies both $E(t) = mN(t)$ from theoretical derivations of section 2.2.1, and the requirement that every node on the final graph has a degree of $k \geq m$. m is the minimum degree number as the probability of finding a node with degree smaller than m is 0, and hence initialisation with an empty graph, star graph, or cycle graph is ruled out as they have a respective minimum degree of $0, 1$ and 2 for all m.

2.1.3 Type of Graph

The models produce are simple graphs; they are unweighted, undirected, and contain no self-loops or multiple edges. Most real-world networks the BA algorithm models are directed -for instance one paper cites another- however the algorithm does not consider

the direction of connections as it should not affect the final degree distribution. Self-loops are not allowed as papers do not cite themselves. A pair of scientific papers will generally not cite each other multiple times, and as such multiple edges on the graph are not allowed.

After initialisation, the graph is predicted to grow into a sparse graph with most nodes having a degree m and some nodes with particularly high degrees ('hubs') will be seen. This is due to $N(0) > m$ where m remains constant.

2.1.4 Working Code

When working correctly, the BA model should add one new node per iteration, and connect it to m existing nodes via an edge. To ensure the numerical implementation is working correctly, the lengths of $\mathcal{G}$ and $\mathcal{G}_{Conn}$ are measured at the first 10 time steps after initialisation. The starting value of $\mathcal{G}$ is expected to be $2m + 1$ (section 2.1.2), and as all nodes are connected to each other during initialisation, the starting value of $\mathcal{G}_{ConnL}$ should be $2(2m+1)$. The value of $\mathcal{G}_{ConnL}$ is expected to increase by $2m$ on each iteration as for each new connection made, both nodes involved are added to $\mathcal{G}_{Conn}$.

Timestep	1	2	3	4	5	6	7	8	9	10
$\mathcal{G}_L$	5	6	7	8	9	10	11	12	13	14
$\mathcal{G}_{ConnL}$	20	24	28	32	36	40	44	48	52	56
$E(t)/N(t)$	2	2	2	2	2	2	2	2	2	2

Table 1: Properties of the graph $\mathcal{G}$ are measured at the first 10 time steps after initialisation for a system of $m = 2$. $\mathcal{G}_L$ corresponds to the number of nodes in the graph, and is seen to increment by 1 as expected. $\mathcal{G}_{Conn}$ contains all instances of a node connection by an edge and as expected increments by $2m$ each timestep.

The value of $E(t)/N(t)$ should also converge to m (Eq. (6)). From the data for $\mathcal{G}_L$ and $\mathcal{G}_{ConnL}$ in Table 1, where $N(t) = \mathcal{G}_L$ and $E(t) = \frac{\mathcal{G}_{ConnL}}{2m}$, it is clear the ratio does equate to the correct value, m.

2.1.5 Parameters

The parameters defined in this investigation are the number of edges added to each new node m, the final number of nodes in the graphs N, and the number of numerical runs. Although their values vary between the different investigations, N was generally chosen to be large to satisfy theoretical approximations, and m was chosen to be small to reduce the finite-size effects present. When testing a range of m values, those chosen were $m \in \{2, 4, 8, 16, 32, 64\}$ so as to have data linearly spaced on a logarithmic scale. The simulations were run 50 times for each task, as this produced good data in a reasonable amount of time.

2.2 Preferential Attachment Degree Distribution Theory

2.2.1 Theoretical Derivation

The master equation describes the degree distribution of a growing network to which one node is added at each time step,

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

where $n(k,t)$ is the number of nodes of degree k at time t, $\Pi(k,t)$ is the probability that a node of degree k forms an edge with the new node, and $\delta_{k,m}$ guarantees the equation is coherent with the addition of a new node of degree m at each time step. Note that Eq. (1) neglects the case of multiple edges attaching to the same node in one time step; this is only compatible for the case of multigraphs if the probability of it occurring is negligible. Given,

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

where $N(t)$ and $p(k,t)$ are the number of nodes in the graph and the probability of a given node having degree k at time t respectively. Eq. (1) is rewritten:

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

In a system to which one node and m edges are added in each time step, it can be stated:

$$N(t) = N(t-1) + 1 = N(0) + t \tag{4}$$

$$E(t) = E(t-1) + m = E(0) + mt \tag{5}$$

where $E(t)$ is the number of edges at time t. Considering now the ratio of $E(t)$ over $N(t)$ in the long-time limit,

$$\lim_{t\to\infty}\left(\frac{E(t)}{N(t)}\right) = \lim_{t\to\infty}\left(\frac{E(0)}{N(0)+t}\right) + \lim_{t\to\infty}\left(\frac{mt}{N(0)+t}\right) = m, \tag{6}$$

the ratio approaches a constant as time goes to infinity, regardless of initial conditions. Considering now the same ratio applied to a finite system,

$$\lim_{t\to\infty}\left(\frac{E(t)}{N(t)}\right) = \frac{E(0) + m(N - N(0))}{N} = m + \frac{E(0)}{N} - \frac{mN(0)}{N} \tag{7}$$

from which it can be read that,

$$E(t) = mN(t) \tag{8}$$

in the case of $E(0) = mN(0)$ or $E(0), mN(0) \ll N$.

Pure preferential attachment is defined such that $\Pi(k,t) \propto k$, where to ensure normalisation,

$$\Pi(k) = \frac{k}{\sum_{i=1}^{N} k_i}. \tag{9}$$

Furthermore, as every edge connects together two nodes, each of which has their degree increased by 1, it can be stated $\sum_{i=1}^{N} k_i = 2E(t)$. Substituting Eq. (2), (8) and (9) into the master equation obtains

$$N(t+1)p(k,t+1) = N(t)p(k,t) + \frac{1}{2}(k-1)p(k-1,t) - \frac{1}{2}kN(t)p(k,t) + \delta_{k,m}. \tag{10}$$

The degree distribution $p(k,t)$ is assumed to become time-independent when t tends to infinity, i.e. $\lim_{t\to\infty} p(k,t) = p_\infty(k)$, ensuring convergence in the equation solutions. In this limit $N(t)$ is infinite and therefore so is $E(t)$. Considering this, and the relation $N(t+1) = N(t) + 1$ (Eq. (4)), Eq. (10) becomes:

$$p_\infty(k) = \frac{1}{2}[(k-1)p_\infty(k-1) - kp_\infty(k)] + \delta_{k,m}. \tag{11}$$

Three cases are now considered:

1. $k < m$: $p_\infty(k) = 0$ as each new node must form m edges.

2. $k > m$: $\delta_{k,m} = 0$ and the master equation gives $\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2}$.

3. $k = m$: $p_\infty(k)$ adopts same form as for case 2.

The general solution to the difference equation in case 2, $\frac{f(k)}{f(k-1)} = \frac{k+a}{k-b}$, is found using the form of gamma function [3],

$$k + c = \frac{\Gamma(k + c + 1)}{\Gamma(k + c)} \tag{12}$$

hence,

$$\frac{f(k)}{f(k-1)} = \frac{\Gamma(k + a + 1)\Gamma(k + b)}{\Gamma(k + b + 1)\Gamma(k + a)}. \tag{13}$$

It is noticed that by defining

$$h(k) = \frac{\Gamma(k + a + 1)}{\Gamma(k + b + 1)}, \tag{14}$$

it can be stated

$$\frac{f(k)}{f(k-1)} = \frac{h(k)}{h(k-1)} \tag{15}$$

and therefore, once the initial condition $f(1)$ is defined,

$$f(k) = \frac{f(1)}{h(1)}h(k) = A\frac{\Gamma(k + a + 1)}{\Gamma(k + b + 1)} \tag{16}$$

where A is a constant. Applying this general solution the difference equation for case 2 obtains,

$$p_\infty(k) = \frac{\Gamma(k)}{\Gamma(k + 3)} = \frac{A}{k(k + 1)(k + 2)} \tag{17}$$

Substituting this into Eq. (11) for the case of $k = m$ returns,

$$A = 2m(m + 1). \tag{18}$$

The finite size of systems considered in this report are suspected to have an affect on the degree distribution. Therefore a scaling ansatz is proposed of the form,

$$p_N(t) = p_\infty \mathcal{F}\left(\frac{k}{k_m(N)}\right) \tag{19}$$

where $\mathcal{F}$ is a scaling function and $k_m(N)$ is the maximum k value for which finite size effects in the system can be neglected.

2.2.2 Theoretical Checks

The theoretical degree distribution has been found as,

$$p_\infty(k) = \frac{2m(m + 1)}{k(k + 1)(k + 2)}, k \geq m. \tag{20}$$

For large k, Eq.(20) is a power law relation with leading order term k^{-3}. This is coherent with a scale-free distribution, $p(k) \sim k^\alpha$.

For $k \to \infty$, Eq. (20) tends to zero as nodes must not have an infinite degree. Additionally, $p_\infty(k) < \frac{2}{m+2}\ \forall k, m$ and given m is an integer, $p_\infty(k) < 1$.

As $p_\infty(k)$ is a probability distribution function, it must satisfy the normalisation condition. The sum from m to infinity is taken:

$$\sum_{m}^{\infty} p_\infty(k) = 2m(m+1) \sum_{m}^{\infty} \frac{1}{k(k+1)(k+2)} \tag{21}$$

The majority of the terms inside the sum cancel with one another, leaving just

$$m(m+1)\left(\frac{1}{m} - \frac{1}{m} + 1\right) = 1 \tag{22}$$

as expected for normalisation.

2.3 Preferential Attachment Degree Distribution Numerics

2.3.1 Fat-Tail

Data collected for $p_N(k)$ is seen to be a fat-tailed distribution as it contains very few cases of nodes with a large k, and a disproportionate amount of noise for nodes with low k values. This is a key characteristic of a scale-free network. To reduce the statistical noise and obtain reliable values for $p_N(k)$, the data was log-binned. A log-bin scale of 1.3 was used so that each bin was 1.3 times larger than the previous. This scale was a compromise between the amount of data retained at low scales and the cleanliness of data at high scales.

2.3.2 Numerical Results

The BA algorithm was run 50 times for $m \in \{2, 4, 8, 32, 64\}$ and on each occasion a total of $N = 100,000$ nodes were added to the network. The results were log-binned and plotted against k, as shown in Fig. 1. Overall the collected results match predictions as seen by the close fit of points to the line. However, for larger values of k a deviation of points from the line occurs, suggesting a deterioration of the scale-free theory at some cutoff k arising due to the finite system sizes. This occurs due to the breakdown of the infinite N assumption used when deriving Eq. (11).

The obtained results are also plotted against their theoretical predictions in Fig. 2 where the line $y = x$ has been plotted to illustrate the deviation of expected vs. obtained results for low $p_N(k)$ values. It can be expected that these finite-size effects have a dependence on the chosen type of starting-graph. An extension to this report would be to investigate properties of a system beginning with a different graph type.

2.3.3 Statistics

To test the goodness of fit of the numerical data, Kolmogorov-Smirnov statistic values have been evaluated and are displayed in Table.2. This test was chosen as it determines whether to reject the hypothesis that a given sample came from a theoretical distribution. It is appropriate as the raw data collected in the algorithm can be approximated as continuous. The two-sample K-S test used is not ideal however, as it only checks whether the samples come from the same distribution, without knowledge of what the distribution is.

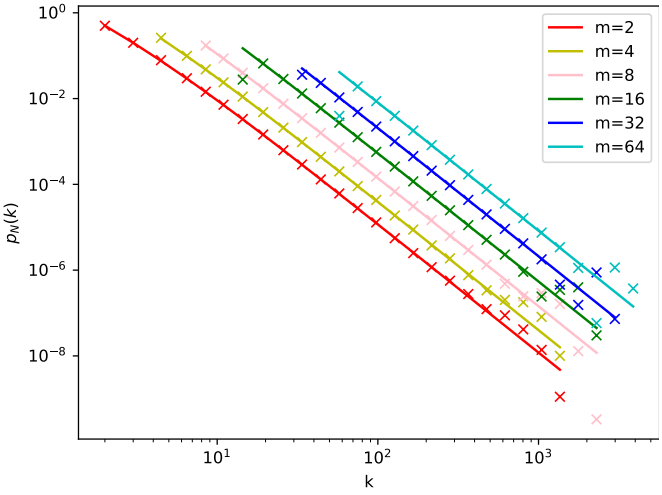

Figure 1: The degree probability distribution $p_N(k)$ collected from running the BA model with preferential attachment is plotted against degree k for $m \in \{2, 4, 8, 32, 64\}$. The theoretical expectation for each m is plotted over the data as a straight line.

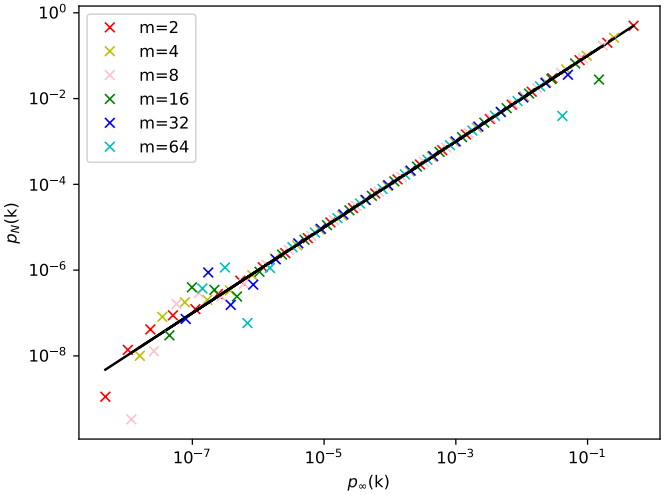

Figure 2: The theoretical degree probability distribution $p(k)$ is plotted in black against the measured data $p_N(k)$ for $m \in \{2, 4, 8, 32, 64\}$. The line $y = x$ has also been plotted to highlight deviations in the data from what was theorised.

m	2	4	8	16	32	64
K-S statistic	0.0417	0.0909	0.0909	0.1	0.0555	0.0588
p-value	0.9999	0.9999	0.9999	0.9999	1	1

Table 2: The BA model has been run 50 times for $N = 100,000$ and $m \in \{2, 4, 8, 16, 32, 64\}$. The K-S statistics of the numerical data are presented along with their corresponding p-values.

The K-S statistics were compared to the critical statistic value $D_c = \frac{1.36}{\sqrt{n}}$ [4] at the $\alpha = 0.05$ significance level for each m. As all K-S statistics were smaller than their corresponding D_c, and the p-values are 1 or very close to 1, the hypothesis that the distributions of the two samples are the same cannot be rejected at even a 5% significance level.

2.4 Preferential Attachment Largest Degree and Data Collapse

2.4.1 Largest Degree Theory

The largest expected degree, k_1 is defined as the degree after which only one node exists with degree k_1 or higher. The expectation value illustrates this:

$$\sum_{k=k_1}^{\infty} Np_\infty = 1. \tag{23}$$

Once again, the contents of the summation are reduced to smaller fractions and evaluated,

$$\sum_{k=k_1}^{\infty} \frac{1}{k+2} - \frac{2}{k+1} - \frac{1}{k} = \frac{1}{2mN(m+1)}, \tag{24}$$

$$k_1(k_1+1) = Nm(m+1). \tag{25}$$

Solving for the physical solution with the quadratic equation obtains,

$$k_1 = \frac{-1+\sqrt{1+4mN(m+1)}}{2} \tag{26}$$

so in the limit of large N it is expected that k_1 would scale as $\sqrt{N}$.

Taking the limit of large N, Eq. (26) becomes,

$$k_1 \simeq \frac{-1\pm\sqrt{Nm(m+1)}}{2} = \sqrt{m(m+1)}\sqrt{N} - \frac{1}{2}. \tag{27}$$

2.4.2 Numerical Results for Largest Degree

An investigation into the dependency of the maximum largest degree on N is undertaken. The simulation is run 10 times for $m = 2$ and $N = 1000,000$, and k_1 is measured every time $N(t)$ is a square number so as to record linearly spaced data on the final graph. The value of m was chosen as it is small compared to the N values studied. The systems are plotted against $\sqrt{N}$ in Fig. 3 where the linear relationship indicates a scaling of k_1 with $\sqrt{N}$, as theorised.

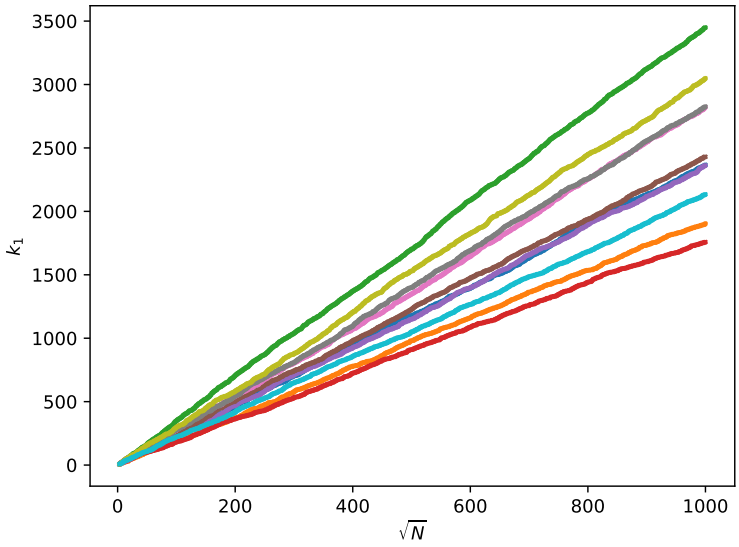

Figure 3: The largest degree for 10 growing graphs, each plotted in a different colour, is measured at a number of intervals. Each graph grows to a size $N = 1000,000$ for $m = 2$. k_1 is plotted against $\sqrt{N}$, where the linear relationships indicate k_1 must scale with $\sqrt{N}$.

The average k_1 values are calculated and plotted in Fig. 4 against $\sqrt{N}$, where the error bars correspond to the standard errors [5]. The numerical results match the theoretical predictions well; the slight deviation will be due to having used an infinite sum in the theoretical derivation whereas only finite-size systems have been measured.

An alternative approach to the measuring of k_1 is undertaken whereby the maximum degree of a graph is only measured at the end of each numerical run, i.e. when $N(t) = N$. The code executes 100 repeats for a range of values of N; and the averages of the maximum degree $\langle k_1 \rangle$ for each run are plotted against $\sqrt{N}$ with associated errors in Fig. 5. The

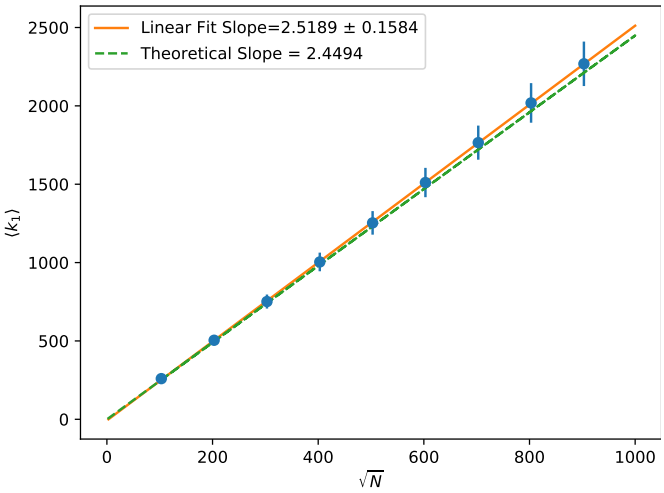

Figure 4: The average maximum degree $\langle k_1 \rangle$ are plotted against $\sqrt{N}$ with standard errors. These errors are calculated as the standard deviation of maximum degrees collected divided by $\sqrt{repeats}$. The linear fit returns a gradient of 2.5189 ± 0.1584.

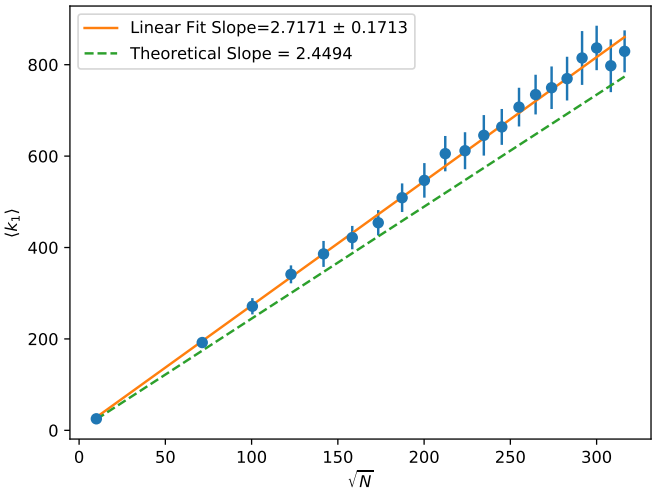

Figure 5: k_1 values are measured at the end of each numerical simulation for N values between $N = 100$ and $N = 100,000$ and averaged over 100 iterations. The error bars are calculated as the standard error. The linear fit returns a gradient of 2.7171 ± 0.1713.

results can be seen to deviate even further from theory so this method is not investigated further.

The Pearsons coefficient of data in Fig. 4 returned a value $r = 0.999981509$, demonstrating the linear correlation between the two variables. The p-value returned was $p < 0.001$. The null hypothesis of the two variables having no linear correlation is therefore rejected, as it is generally rejected with a 95% confidence interval.

As a further investigation, the standard deviations σ are plotted against $\sqrt{N}$ in Fig. 6, and display a linear relationship with $\sqrt{N}$ themselves, as may be expected by the scale-free behaviour of the BA model.

The fractional error of k_1 is shown to be constant over all system sizes N: Fig. 7 displays $\frac{\langle k_1 \rangle}{\sigma}$ against $\sqrt{N}$ where the data tends towards a constant value of 4.8073. Therefore, $\frac{1}{4.8073} = 0.2080$ can be taken as the amount by which the measured value of k_1 is expected to vary for each system.

2.4.3 Data Collapse

The predicted scaling ansatz of Eq.(19) has a dependency on the cutoff degree size $k_m(N)$, defined as the maximum degree size before which finite size effects are significant. The values of k_1 illustrated in Fig. 5 are the largest present degree in each graph, and it is expected that finite sizing effects will have an effect on their probability. It is therefore predicted that the cutoff degree size $k_m(N)$ will scale with k_1. To investigate the effects of finite scaling on the system, the code is run for $N \in \{10^3, 10^4, 10^5, 10^6\}$ and data being collected over 100 repeats for $m = 2$. $\frac{p_N(k)}{p(k)}$ is plotted against $\frac{k}{\sqrt{N}}$ in Fig. 8, where it can be seen for $k \ll k_1$ the collected degree distribution data is well approximated by the theoretical degree distribution, as their ratio remains constant and equal to 1. The data

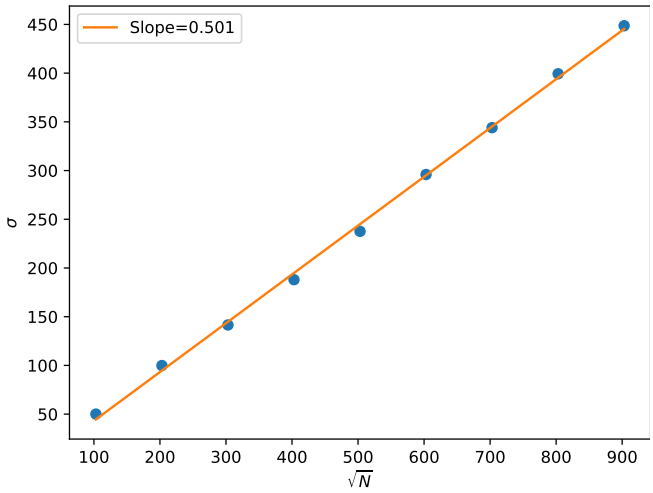

Figure 6: The standard deviation errors in $\langle k_1 \rangle$ data from Fig. 4 is plotted against $\sqrt{N}$ and illustrates a linear relationship. This is expected from the scale-free behaviour of the BA model.

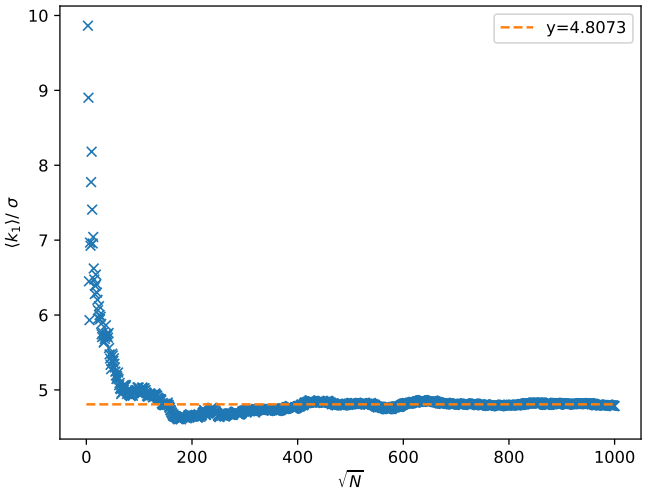

Figure 7: The fractional error in k_1 tends to a constant, indicated by the dashed orange line at $y = 4.8073$. This suggests the fractional error in each system is $\frac{1}{4.8073} = 0.2080$.

diverges from theory at $k \approx k_1$, experiencing a bump and steep decline owing to the finite size of systems considered in this model; the size of k is capped by the value of N and therefore a build-up of nodes in the vicinity of N is seen. These nodes would have had higher degrees if further nodes and edges had been added to the system.

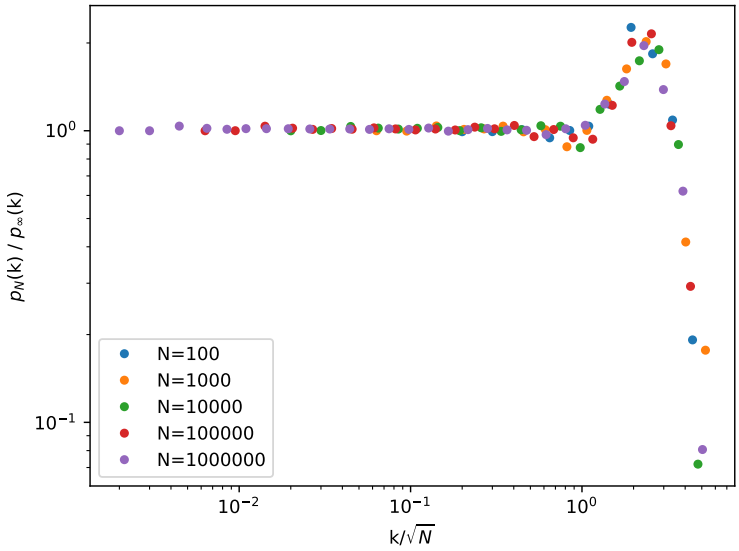

Figure 8: A data collapse is performed; $\frac{p_N(k)}{p_\infty(k)}$ is plotted against $\frac{k}{\sqrt{N}}$ as per the scaling function stated in Eq. (19). A range of system sizes are plotted with $m = 2$ and the code was run 50 times for each system size to improve the statistics. The data has been log-binned, and is plotted on a logarithmic scale.

3 Phase 2: Pure Random Attachment Π_{rnd}

3.1 Random Attachment Theoretical Derivations

3.1.1 Degree Distribution Theory

When all nodes have an equal probability of being chosen, $p(k,m)$ remains as in Eq. (10) but now $\Pi(t) = \frac{1}{N(t)}$. Substituting this into the master equation:

$$N(t+1)p(k,t+1) = N(t)p(k,t) + mp(k-1,t) - mp(k,t) + \delta_{k,m} \tag{28}$$

In the limit $t \to \infty$:

$$(1+m)p_\infty(k) = mp_\infty(k-1) + \delta_{k,m} \tag{29}$$

Considering once again three cases:

1. $k < m$: $p_\infty(k) = 0$ as each new node must form m edges.
2. $k > m$: $p_\infty(m) = \frac{1}{m+1}$.
3. $k = m$: Eq (29)simplifies to a trivial recurrence relation $p_\infty(k) = \frac{m}{m+1}p(k-1)$.

Therefore, combining cases 2 and 3,

$$p_\infty(k) = \frac{1}{m+1}\left(\frac{m}{m+1}\right)^{(k-m)} \quad k \geq m, \tag{30}$$

where it can be said $\frac{m}{m+1} \ll 1$ and $\frac{1}{m+1} \ll 1 \; \forall m$ as $p_\infty(k)$ is a probability density function and therefore must correspond to a positive integer. Furthermore, $p_\infty(k) \to 0$ as $k \to \infty$. Eq. (30) is normalised and describes a geometric sequence for consecutive k,

$$\sum_{k=m}^{\infty} p_\infty(k) = \frac{(m+1)^{-1}}{1-(m+1)^{-1}} = 1 \tag{31}$$

from above it can be confirmed that $p_\infty(k)$ coheres with expectations. $p_\infty(k)$ does not, however, describe a scale-free network as was studied before; the degree distribution is not uniform as the network is growing.

3.1.2 Largest Degree Theory

As stated in section 2.4.1, the expectation value of the largest degree k_1 must equal 1. In the case of pure random attachment,

$$\sum_{k=k_1}^{\infty} p_\infty(k) = \frac{1}{N}. \tag{32}$$

Rearranging and plugging in Eq. (30) for $p_\infty(k)$,

$$\sum_{k=k_1}^{\infty}\left(\frac{m}{m+1}\right)^{k} = \left(\frac{m}{m+1}\right)^{m}\frac{m+1}{N}. \tag{33}$$

The summation index on the left hand side of Eq. (33) is now re-expressed as $k = k_i + i$,

$$\sum_{k=k_1}^{\infty}\left(\frac{m}{m+1}\right)^{k} = \sum_{i=0}^{\infty}\left(\frac{m}{m+1}\right)^{k_1+i} = (m+1)\left(\frac{m}{m+1}\right)^{k_1} \tag{34}$$

where the final step is the evaluation of the simple geometric sum to infinity. Therefore, Eq. (34) becomes:

$$\left(\frac{m}{m+1}\right)^{k_1} = \left(\frac{m}{m+1}\right)^{m} \frac{1}{N}. \tag{35}$$

Finally, by applying the natural logarithm to both sides and rearranging for k_1:

$$k_1(ln(m) - ln(m+1)) = m(ln(m) - lm(m+1)) - ln(N) \tag{36}$$

$$k_1 = m + \frac{1}{log\left(\frac{m}{m+1}\right)} log(N) \tag{37}$$

For the case of pure random attachment the largest degree scales with $log(N)$ for $N \gg m$, rather than as a asymptotic power law.

3.2 Random Attachment Numerical Results

3.2.1 Degree Distribution Numerical Results

The simulation is run for $m \in \{2, 4, 8, 16, 32, 64\}$ and $N = 100,000$ over 50 iterations. The degree data does not form a fat-tailed distribution hence the data is not log-binned. Fig. 9 displays $p_N(k)$ data against k, with theoretical predictions from Eq. (30) overlaid as lines.

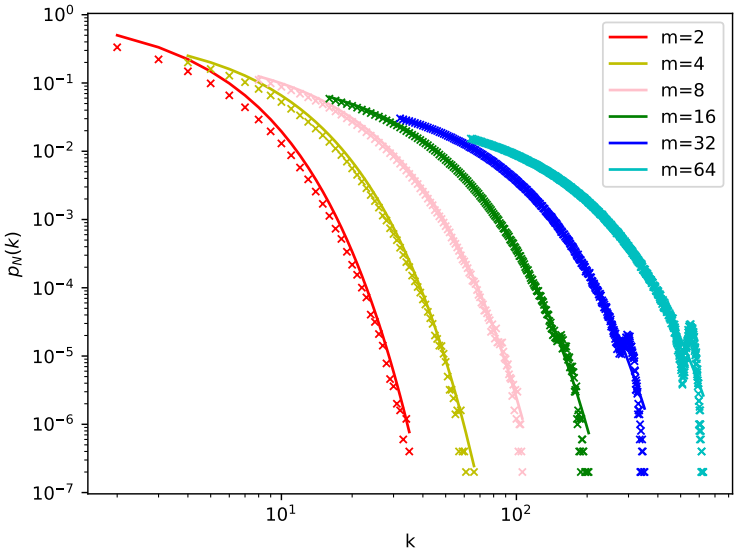

Figure 9: The degree probability distribution $p_N(k)$ collected the BA model with pure random attachment is plotted against k for various m. The kinks seen are finite-size affects which grow with increasing k.

The general trend of points lie close to theory, however for the higher values of k in each m plot distinct kinks can be seen owing to the effects of finite-size scaling. Once again, a K-S test has been carried out on the data, the results of which are displayed in Table. 3. The K-S statistics are small and the p-values for m up to $m = 8$ are close to 1, from which it would be inferred that the null hypothesis of the data matching numerical predictions cannot be rejected. However for larger m values, the p-values vary significantly from those for preferential attachment due to the prominent presence of finite-scaling effects. These

effects of course exist both for preferential attachment and random attachment, however as no log-binning of the data was conducted here the effects are more obvious and hence the statistics are substantially worse.

m	2	4	8	16	32	64
K-S statistic	0.0588	0.0678	0.0606	0.0604	0.0701	0.0396
p-value	0.9999	0.9994	0.9939	0.8950	0.4245	0.7763

Table 3: The simulation has been run 50 times for $N = 100,000$ and $m \in \{2, 4, 8, 16, 32, 64\}$. The K-S statistics of the numerical data are presented along with their corresponding p-values.

3.2.2 Largest Degree Numerical Results

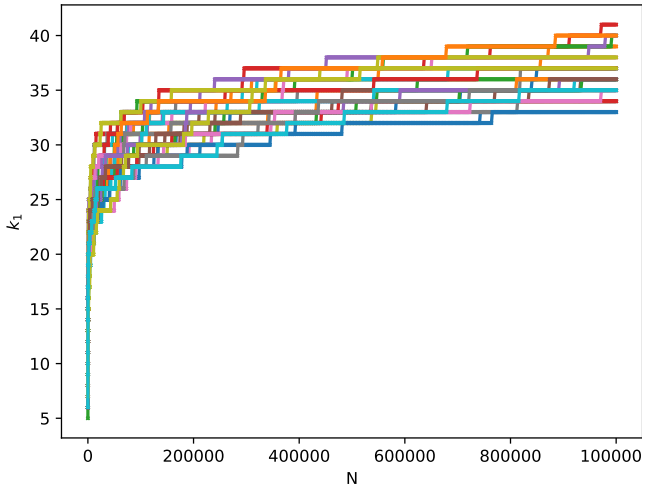

Figure 10: The maximum degree size k_1 is plotted against N for 30 systems run with $m = 2$ and $N = 1000,000$. k_1 is measured at multiple intervals during the growth of each graph.

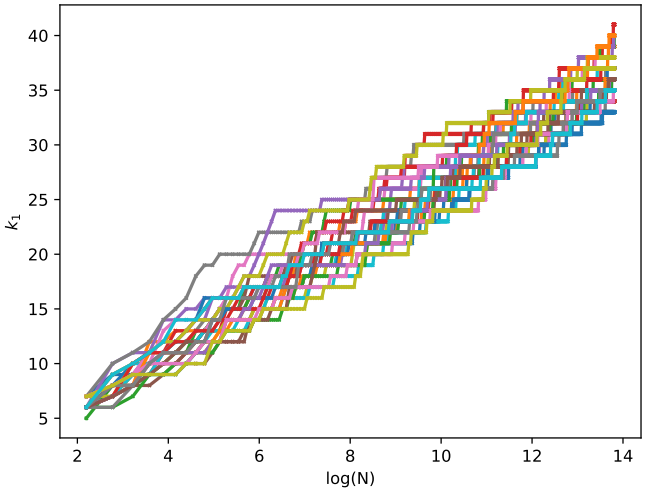

Figure 11: The maximum degree size k_1 is plotted against $log(N)$ for 30 systems run with $m = 2$ and $N = 1000,000$. The linear relation highlights the scaling of k_1 with N.

The simulation was run 30 times for $N = 1000,000$ and $m = 2$. The maximum degree of each graph was measured multiple times during each run: once again every time $N(t)$ was a square number. The values collected for k_1 are plotted against N in Fig. 10, and against $log(N)$ in Fig. 11 which displays a linear fit. The averaged data is therefore plotted against $log(N)$ in Fig. 12. The theoretical expectation according to Eq. (37), plotted in green, fits increasingly well for larger N. The deviation may be due to the use of an infinite sum over k during the theoretical derivation.

A data collapse has been attempted, plotted as $\frac{p_N(k)}{p_\infty(k)}$ vs. $\frac{k}{log(N)}$ in Fig. 13. Despite the collapse seeming successful for low k values, clear finite-size effects are visible for high k, especially for higher N values tested. As such, the system does not exhibit scale-free behaviour.

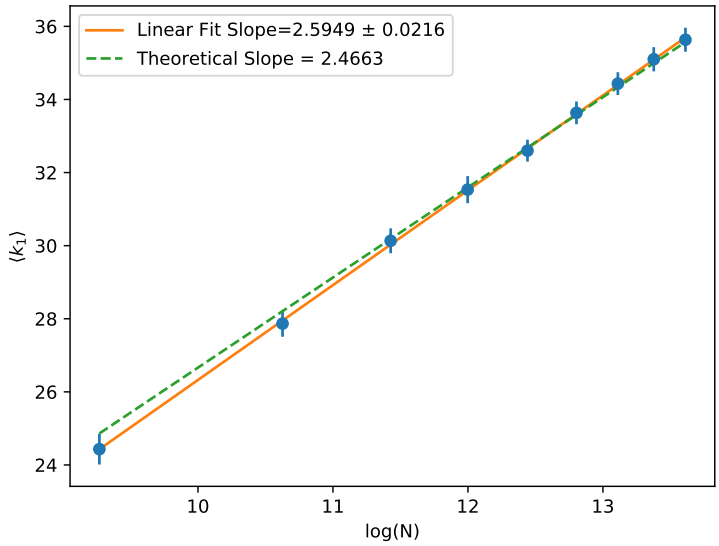

Figure 12: The average maximum degree $\langle k_1 \rangle$ plotted against $log(N)$ for the BA model run with pure random attachment. The error bars are calculated as the standard error, and the linear fit returns a gradient of 2.5949 ± 0.0216.

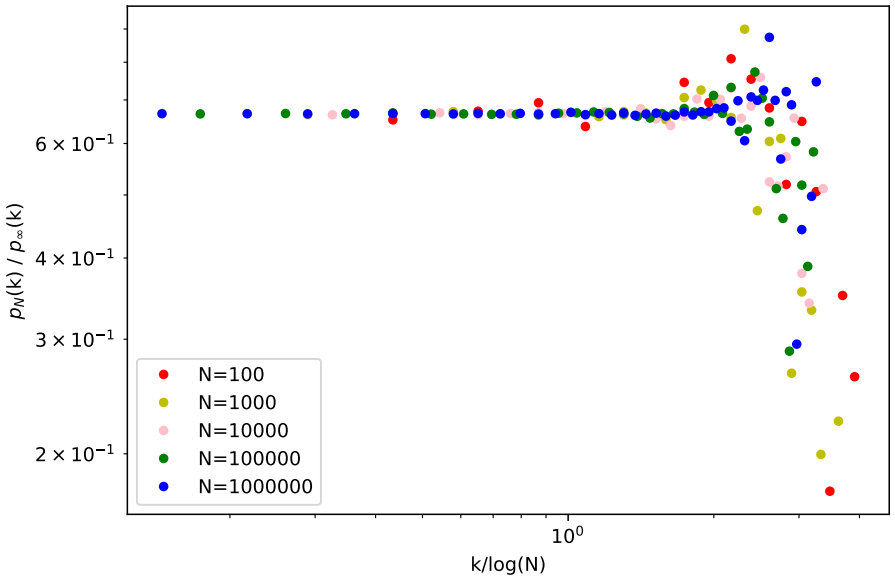

Figure 13: $\frac{p_N(k)}{p_\infty(k)}$ is plotted against $\frac{k}{log(N)}$ for various values of N. A data collapse is not achieved; the system does not exhibit scale-free behaviour.

4 Phase 3: Random Walks and Preferential Attachment

4.1 Implementation

The model was set up with a random walk attachment method: an existing node n_i in the graph $\mathcal{G}$ is chosen uniformly at random and is the node from which the random walk

begins. The probability that the current node is accepted and becomes the node with which a new edge is formed is defined $p_{acc} = (1 - q)$, where q is a parameter of the algorithm. A random number R between 0 and 1 is chosen and a while loop is entered: while $R < p_{acc}$ the algorithm randomly chooses a node to which n_i is connected via an edge, this new node becomes n_i, and one step in the random walk is taken. The algorithm continues updating the random number R, and taking steps in the random walk, until $R > p_{acc}$ at which point the node n_i is accepted and an edge is formed.

It is seen that parameter q will heavily influence the average length of each random walk: for instance a low q will result in a high p_{acc} and hence the length of the random walk d is expected to be small. It can be expected that q will also determine the computational time required to collect data, as for a high q each new edge requires a long random walk to be completed before formation and hence the simulation will take longer to run. As such, the model is run for a low value of m, $m = 2$, and a system size $N = 100,000$. The maximum value of q investigated was 0.9 as for $q \to 1$, $\langle d \rangle \to \infty$ and hence a numerical simulation is not possible.

4.2 Numerical results

The theoretical degree distribution for the random walk method would have a similar form to that in Eq.(28) but with a consideration of the degree of its nearest neighbours. For a given random walk, this degree distribution of a neighbour d edges away is,

$$p_{RW}(k,t) = \left(\frac{k}{\langle k \rangle_t}\right)^d p(k,t) \tag{38}$$

where $\langle k \rangle_t = \sum_{k=m}^{\infty} p(k,t)k$. A complete theoretical description of random walk attachment is beyond the scope of this report, as a discrete differential equation of $\langle k \rangle_t$ would need to be solved.

It is expected that for $q = 0$ the degree distribution will become identical to that of pure random attachment, i.e. $p_{RW}(k) = p(k)$, as $\langle d \rangle = 0$ always. The degree distribution is collected for $q \in \{0, 0.5, 0.9\}$ and plotted in Fig. 14. Note the data here has been log-binned as for $q = 0.9$, a fat-tailed distribution is produced which obscures the other results. As expected, the data for $q = 0$ matches that of the pure random attachment method. Additionally, as $q \to 1$, the data tends to that of a pure preferential attachment method.

The K-S statistics for the $q = 0$ distribution indicate the hypothesis of the results originating from the pure random attachment model can not be rejected at even a 5% significance level, and the statistics for the $q = 0.9$ distribution indicate the preferential hypothesis could also not be rejected at a 5% significance level. For the case of $q = 0.5$ the results are incomplete; the distribution lies equally between that expected for preferential and random attachment.

The expectation value of the random walk can be expressed $\langle d \rangle = \frac{q}{1-q}$. The system is run for $q \in \{0.1, 0.2, ...0.9, 1\}$ and a counter d keeps track of the number of steps taken in the random walk before a new edge is formed. The results are averaged and plotted against q in Fig. 15, where they are seen to align exactly with the theoretical expectation.

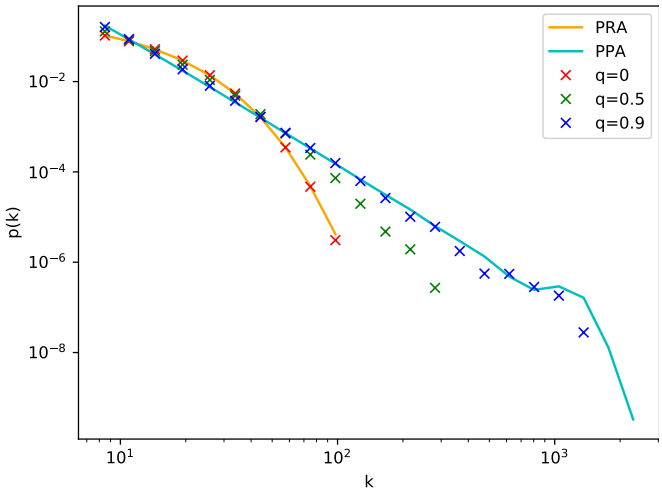

Figure 14: The degree probability distribution is collected for a system of size $N = 100,000$ and of $m = 2$ for parameter values $q \in \{0, 0.5, 0.9\}$. The data collected for pure preferential attachment and pure random attachment are also plotted as lines with the labels PPA and PRA respectively.

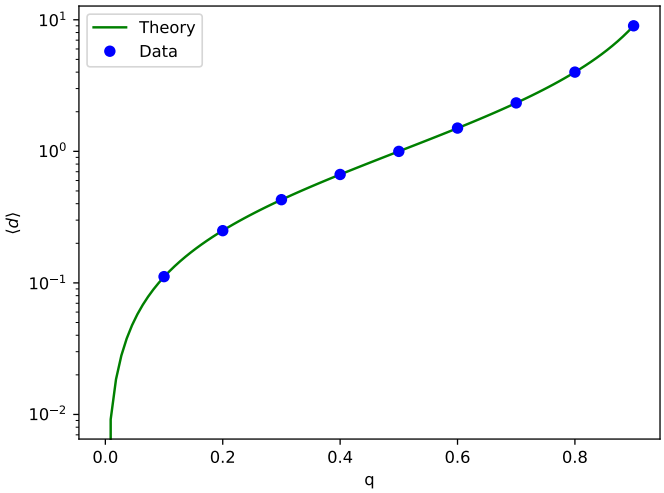

Figure 15: The average random walk distance is plotted as a function of q for a system of size $N = 100,000$ and of $m = 2$. The theoretical expectation for $\langle d \rangle$ is also plotted. The random walk algorithm is seen to tend towards preferential attachment: as $\langle d \rangle$ tends to zero, q also tends to zero.

4.3 Discussion of Results

It is the dependence of probability Π on the degree of a selected node which leads to fat-tailed degree distributions. This was seen when investigating preferential attachment in section 2, where Π_{pa} was proportional to k, and in section 4 for random walks with a parameter value of q close to 1. The behaviour of the latter case can be traced back to Eq. (38), which shows p_{RW} has a dependence on k as long as $d \neq 0$. As Π_{rnd} for pure random attachment had no dependence on k, a fat-tailed distribution was not seen. This type of behaviour is observed in social systems as the "rich get richer" effect: when the likelihood of a connection being made to a person is proportional to their wealth or status.

5 Conclusions

This report has investigated three methods of node attachment in a growing network: pure preferential attachment, pure random attachment, and random walk attachment. Theoretical models were proposed and confirmed by numerical results. The BA model with preferential attachment behaved as a scale-free network, whereas random attachment did not. The results obtained for random walk attachment indicate the degree distribution varies with a dependence on the parameter q between behaviour of the random attachment model and the preferential attachment model.

References

[1] National Research Council, "*Committee on Network Science for Future Army Applications* doi:10.17226/11516, ISBN 978-0309653886." *The National Academies Press.*

[2] S. Cohen, Reuven; Havlin, "*Scale-Free Networks Are Ultrasmall*," *Physical Review Letters.*

[3] E. Weisstein, "Gamma functions," Available at http://mathworld.wolfram.com/GammaFunction.html [Accessed: 1.3.2020].

[4] S. Rick Wicklin, "Critical values of the kolmogorov-smirnov test," Available at https://blogs.sas.com/content/iml/2019/05/20/critical-values-kolmogorov-test.html [Accessed: 20.3.2020].

[5] Investopedia, "Standard error of the mean," Available at https://www.investopedia.com/ask/answers/042415/what-difference-between-standard-error-means-and-standard-deviation.asp [Accessed: 20.3.2020].