Degree distribution of preferential and random attachment network models

Priyanka Mocherla

27th March 2017

CID: 00929775

Words: 2480

***This paper studies the degree distribution of simulated networks. Networks are built using three methods of attachment: pure preferential attachment, using the Barabási-Albert model, pure random attachment, and random walk attachment. The degree probability distribution and the cut-off degree for both preferential and random networks are derived, and compared to results measured from the simulation. For preferential attachment networks, the relationship between the cut-off degree and system size, N, is verified to scale with N1/2, with the exponent measured as 0.500 ± 0.002. The models for both network types produce a high goodness of fit when compared to theoretical results, however they show evidence of finite-size scaling in the distribution tail. The random walk attachment networks show that preferential attachment can be generated through random local processes and display scale-free behaviour.***

# Introduction

Networks give structure to the bilateral relationships present between objects. They are not universal and can be classified in different ways. Two main types of networks are preferential attachment networks and random networks.

The World Wide Web is an example of a preferentially attached network. A network of this nature displays scale-free growth and the existence of hubs, which means that there exist a few objects with many links to them. In fact, it is found that many real life systems converge to this scale-free model, whereby making links to other objects is more likely if the object has a large number of links already.

This paper will use the Barabási-Albert model to study preferential attachment, and compare results to a random network. It will also discuss random walk attachment networks, and how they are able to display preferential attachment.[1]

# Theory

To compare the models investigated in this paper, two main metrics were used: degree probability distribution and cut-off degree. Both were derived for the Barabási-Albert model and a random network model, starting from the master equation.

## Master Equation

To find the best theoretical form of the degree distribution for preferential attachment, the master equation, representing the growth of a network as a function of time and degree, is presented as:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where is the number of nodes of degree at time , is the number of stubs per new node, and is the probability that a single edge from the new node attaches to a node of degree k. It is noted that the master equation here only permits an increase of one degree per node for any m value.[2]

By defining , the time dependent degree probability distribution, as:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where is the total number of nodes in the system at time , and letting , Equation 1 can be rewritten as:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Now, an asymptotic stationary solution is assumed for the probability, where:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Therefore giving:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

In this model, a node is added with a specified number of edges, , per time-step. Therefore, a constraint is imposed that is only specified for , and is zero otherwise. We also require that the total number of nodes added is much larger than the number of edges added per node, to maintain a sparse network, and to eliminate the effect of initial conditions.

## Preferential Attachment

### Degree Distribution

According to the Barabási-Albert model, must be proportional to the degree of the node, and therefore can be equated as the following:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

Where the last term makes use of the approximation that the number of edges, and that the total degree of the system is equal to for large t.[2] Substituting this into Equation (5), we obtain:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

Considering the rate equations for the conditions k > m and k = m, Equations 8 and 9 are respectively produced:

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Here, with a variable change from for Equation 8, a simple recursion relation can be noticed. By starting from Equation (9), applying Equation (8) and changing , the discrete degree distribution is obtained as:

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

For large this tends to the continuous limit where , producing a fat-tailed distribution.

### Cut-off degree,

To obtain the best theoretical estimate of the largest expected degree, , the cut-off degree should be defined as the value of onwards where there is only one occurrence. In other words:

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

Inserting Equation 10 into Equation 11, and rewriting the result as partial fractions, the following is obtained:

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

which can be split into:

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

Explicitly evaluating Equation 13, and taking the limit , produces:

|  |  |  |
| --- | --- | --- |
|  |  | (14) |

Using the quadratic equation and recognising that only the positive solution corresponds to a physical system, the theoretical cut-off degree is:

|  |  |  |
| --- | --- | --- |
|  |  | (15) |

As the number of nodes in the system, , becomes large, the cut-off degree scales as .

## Random Attachment

### Degree Distribution

For random attachment, the probability of an edge attaching to a target node is uniform:

|  |  |  |
| --- | --- | --- |
|  |  | (16) |

Substituting this into the master equation, the following equation is obtained:

|  |  |  |
| --- | --- | --- |
|  |  | (17) |

Evaluating this equation for and recognising that for, we find:

|  |  |  |
| --- | --- | --- |
|  |  | (18) |

Next, evaluating Equation 17 for , we obtain:

|  |  |  |
| --- | --- | --- |
|  |  | (19) |

giving the following recursion relation by induction:

|  |  |  |
| --- | --- | --- |
|  |  | (20) |

So the asymptotic degree probability is:

|  |  |  |
| --- | --- | --- |
|  |  | (21) |

This does not follow a power law decay; therefore a fat-tail distribution is not expected.

### Cut-off degree,

Using the same method as for Section 2.2.2, the following sum has to be solved:

|  |  |  |
| --- | --- | --- |
|  |  | (22) |

Renumbering the index and extracting out constants gives:

|  |  |  |
| --- | --- | --- |
|  |  | (23) |

Using the standard result for the sum of an infinite geometric series and rearranging:

|  |  |  |
| --- | --- | --- |
|  |  | (24) |

Taking the logarithm of Equation 24 and rearranging, we obtain:

|  |  |  |
| --- | --- | --- |
|  |  | (25) |

Therefore the cut-off degree scales according to , smaller than the scaling expected for preferential attachment.

# Method

## The Algorithm

To simulate network growth, the following algorithm was produced:

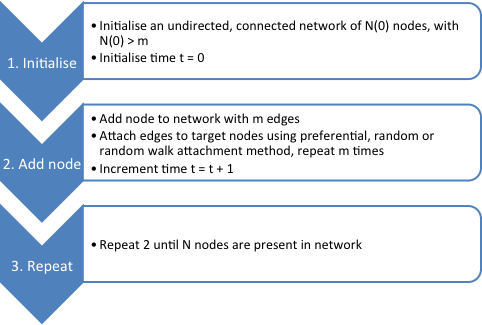


Figure : The simplified algorithm for network simulation. The attachment of nodes is specified according to the mechanism required. The network is initialised with N(0) > m to reduce the probability of multi-links and self-loops are not permitted. The approximation is made that the number of nodes in the system at time t is equal to the time.

The network created in this simulation allowed for multi-links and was undirected. Self-loops were not permitted. In general the value of N was set to be much greater than m, thereby producing a sparse network. The parameters of the model were (equal to the time), and ,which was set to , to reduce the probability of multi-links.

For preferential attachments, Equation 6 stipulates that the probability of attaching to a node is proportional to . To achieve this, a list of nodes was maintained, with node frequency equal to the node degree. Therefore, the number of entries of a node in the list is proportional to its degree. Upon addition of a node, a target was chosen from this list, and the index of the added node and its target were subsequently added to the list.

Random attachments were made by randomly picking a target node in the system, where all nodes had a uniform probability of being chosen.

Random walk attachment involved initially choosing a target node as per pure random attachment. A neighbouring node was then chosen at random and moved to and this was repeated based on the step length specified.

## Testing the Algorithm

In deriving the form of the asymptotic probability from the master equation, the approximation made is that multi-loops and self-loops do not occur. This simulation however does permit multi-links. To check the significance of this approximation, a network with 105 nodes and = 4 was created, and the number of multi-links were calculated as a fraction of the total edges. Results showed that less than 0.05% of the edges were multi-links, therefore allowing the use of Equation 1. The Barabási-Albert model was checked statistically by calculating the average number of edges, , and degree size, , for large .[2] Running the simulation for 106 nodes, = 2, 4, 8 returned 4.0, 8.0 and 16.0, and 2.0, 4.0, 8.0, as required.

To test that the initial graph configuration did not affect results, a sparse graph and a complete graph were both initialised for the models. The probability distributions plotted produced no evidence of being affected by the initial conditions. Finally, by setting a random number seed, the model was run for a small size and manually checked to ensure correct procedure during network creation.

# Results and Analysis

## Preferential Attachment

### Probability Distribution and Edge number

Using the preferential attachment scheme in conjunction with the algorithm outlined in Section 3.1 initialised with a sparse graph, a network was grown to N = 105 nodes. By calculating the frequency probability of node degrees, the data was plotted for between 1 and 64, along with the theoretical distribution, given by Equation 10. These values were chosen to ensure a sufficient range was chosen for investigation, with sufficient data points included.

Figure 2 presents this data. The measured data follows theoretical results, however the statistical noise present at the tail of the fat-tail obscures underlying distribution behaviour. To overcome this, the data was processed using logarithmic binning, where the data was input into bins of logarithmically increasing sizes depending on the multiplication factor, optimised at 1.2.[3] The processed data is displayed in Figure 3. In general, the distribution follows a power law, predicted by Equation 10, and decays rapidly after a cut-off due to finite system size.

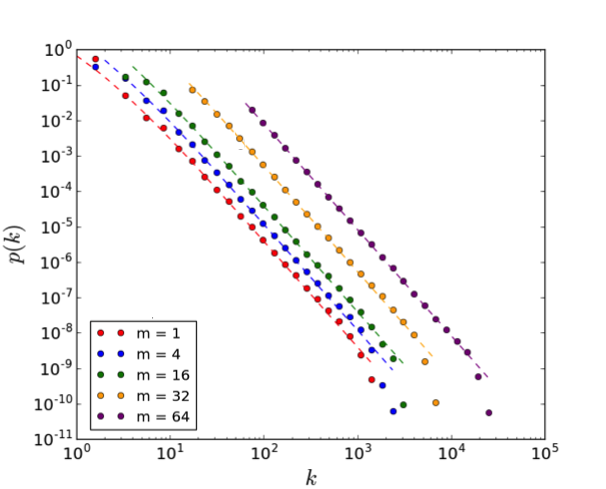
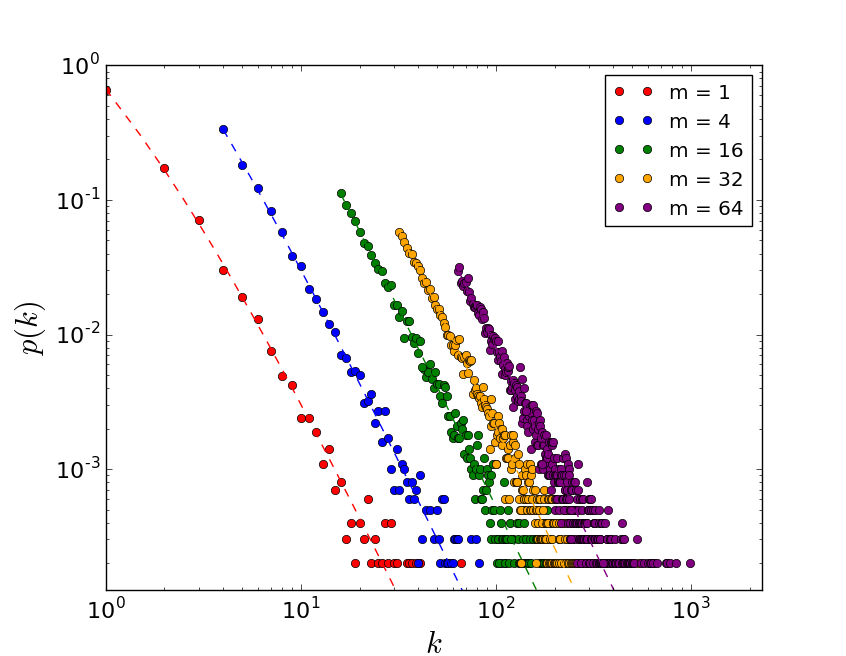
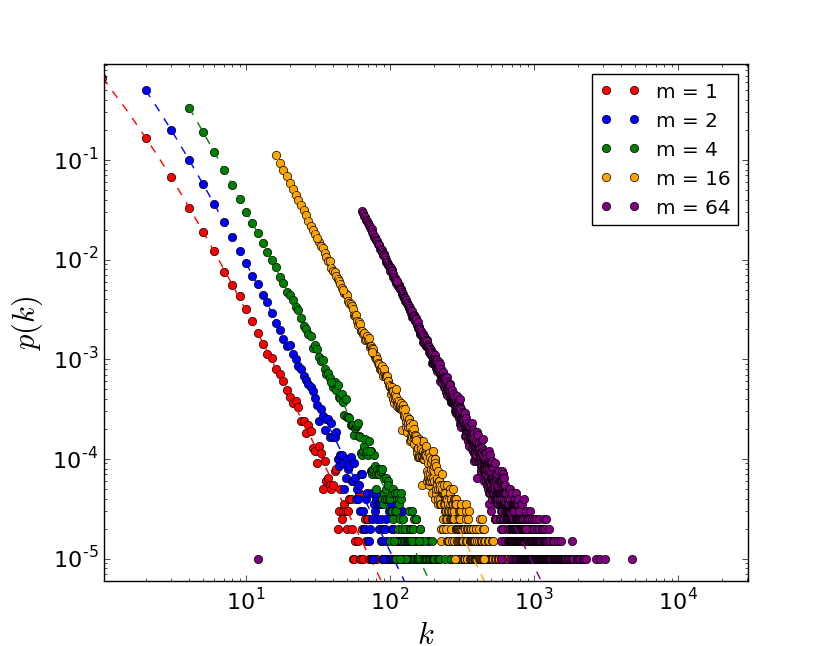
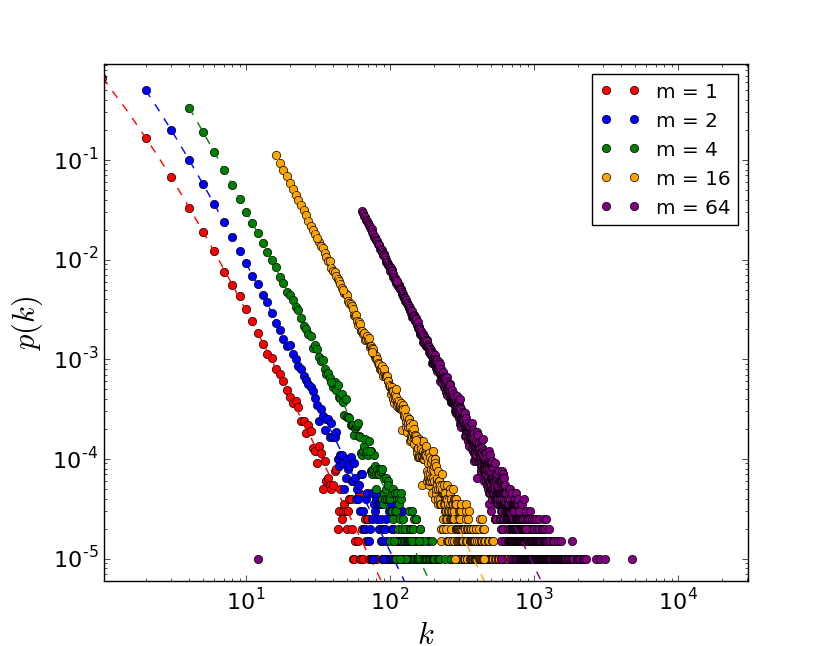


Figure : The unprocessed probability distribution data for 105 nodes, and different m values. The data corresponds roughly to the theoretical results – given by the dashed lines. Data processing is required to remove the statistical noise of the fat-tail.

Figure : Log-binned data for 105nodes. The bulk of the data follows the scale-free relationship presented by the continuum limit of Equation 10. For small values of k, the model breaks down. The data shows effects of finite size scaling from its rapid decay upon reaching a cut-off value.



A goodness of fit test with the theoretical distribution was performed for each dataset. For simplicity, the scaling region of the log-binned data was fit using linear regression, and compared to theory. As the logarithmic scale invalidates the result any linear residual analysis, a χ2 test was used instead. The results are shown in Table 1. By setting a significance level of 5%, the p-value shows that between = 2 and = 16, the data fit successfully and the best fit was achieved for = 4. This, along with optimisation of run-time, determined that = 4 was used for subsequent analysis.[4]

|  |  |  |  |
| --- | --- | --- | --- |
| **m** | **Gradient** | **χ2** | **p-value** |
| 1 | -2.930 | 4.855 | 0.847 |
| 2 | -2.959 | 1.452 | 0.963 |
| 4 | -2.983 | 0.088 | 0.993 |
| 16 | -3.001 | 0.099 | 0.952 |
| 64 | -2.995 | 0.126 | 0.939 |

Table : The results for the chi-square test performed on 106 node networks, for various m values. The gradient of the scale-free region, relating to the exponent of the continuous degree distribution approximation, of each data set was measured for a quantitative check with theoretical results, largely corresponding to the prediction of -3. For a significance level of 5%, the p-value shows that m = 2, 4, 16 strongly correspond to theory, whilst m = 4 has the best goodness of fit value.

### Finite Size Effects

To observe finite size effects in the data, N was varied between 102 and 106 nodes for = 4. The data was collected over a number of runs, and processed using log-binning. Results are presented in Figure 4.

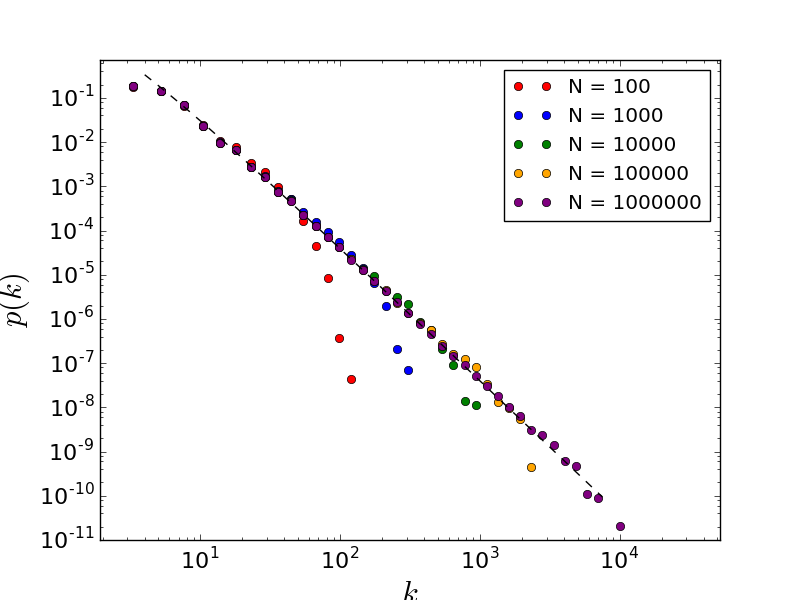
From Equation 15, in the large system-size limit, the cut-off degree, k1, is expected to scale with N0.5. For each N, each k1 of the runs were measured and averaged, whilst calculating the standard deviation. The data is plotted in Figure 5, where the gradient of this line was fitted using linear regression. The gradient returned was 0.500 ± 0.002, agreeing with the theoretical exponent of 0.5. By scaling degree with k1, and the measured probabilities with the Equation 10, a data collapse was performed, shown in Figure 6. At , there is a characteristic bump in probability, followed by a rapid decay.

Figure : Log-binned data for varying N and m = 4. For each N, the algorithm was averaged over [10000, 1000, 100, 50, 25] runs respectively. The effect of the finite system-size is more prominent for smaller N, with the probability decaying rapidly. There is a slight bump in each data set before this decay occurs. For small degree sizes, the data shows deviation from expected behaviour. Larger N follow the theoretical relation, given by the dashed line, more closely.

## Random Attachment

Figure : The relationship between k1 and N follows a power law. The gradient of the fitted data points was calculated as 0.500 ± 0.002, strongly corresponding to the theoretical result, given by the blue dashed line. Each data point was averaged over [10000, 5000, 1000, 500, 50] runs, respectively.

Figure : A data collapse of the data in Figure 5, using the theoretical asymptotic probability and the scaling of the cut-off degree. A bump, characteristic of self-organised critical systems, occurs for k/k1 ~ 1, thereafter the probability rapidly decays. The successful data collapse shows that the network exhibits scale-free behaviour.

### Probability Distribution and System Size

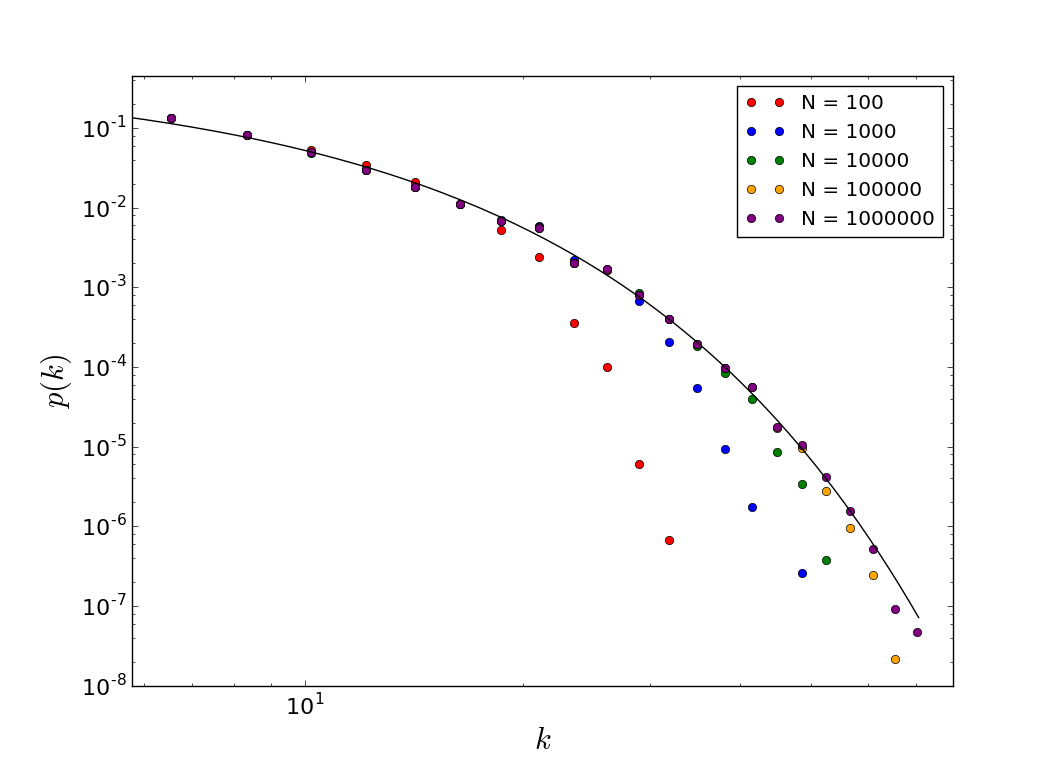
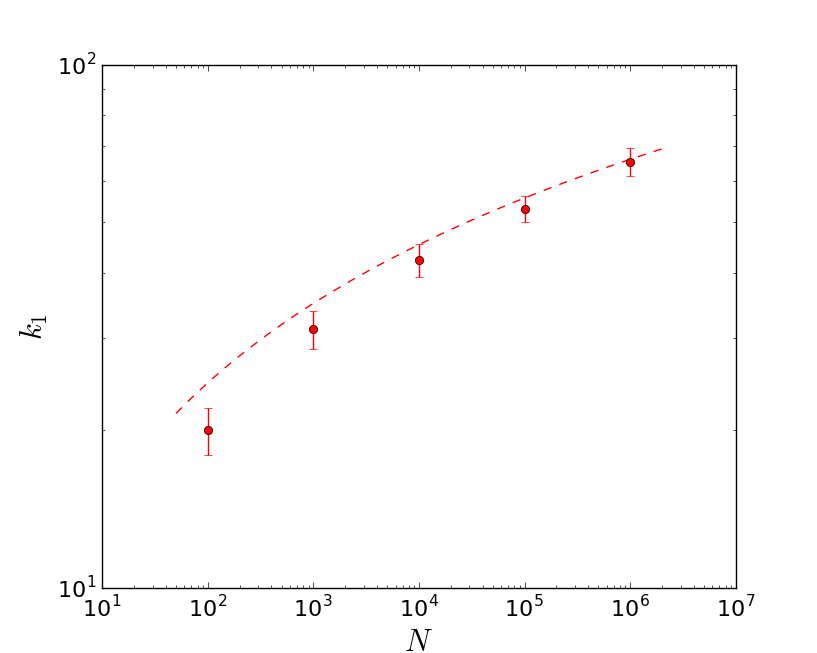
Equation 21 states that the random attachment scheme will produce an exponential network, exhibiting growth. Plotting the log-bin processed data for the probability distribution with varying N, Figure 7 presents their relationship with = 4. For small system sizes, the tail of the distribution does not tend to the asymptotic probability distribution. Given that Equation 21 is only valid for , smaller system sizes are not expected to follow it exactly. Compared to the preferential attachment model, the cut-off degree for each system size is around an order of magnitude lower, suggesting a slower growth rate.

Figure 8: The relationship between k1 and N follows a logarithmic relation, which displays less growth than the power law given for preferential attachment. Each data point was averaged over [10000, 5000, 1000, 500, 50] runs, respectively. For low N, there is a deviation of results from theory (red dashed), however as N becomes large, the data tends to the true values.

Figure : The probability distribution for m = 4 at various N. For small system sizes, the probability decays earlier, compared to the theoretical probability, given by the black line. Unlike for preferential attachment, no obvious bump occurs before decay.

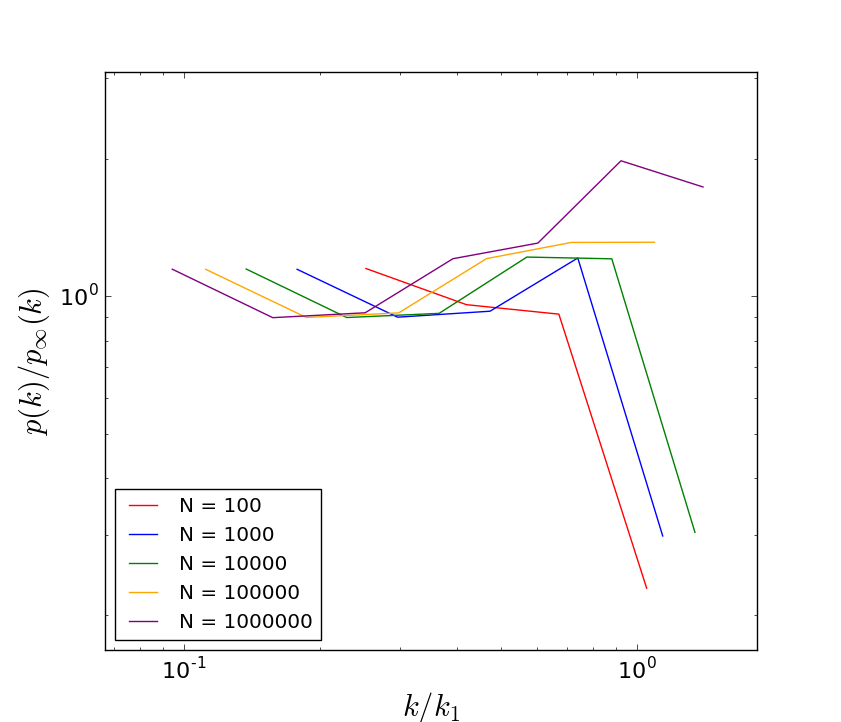
Equation 25 shows that there is a logarithmic relationship between N and k1. Taking multiple runs for each N, the average k1 and its uncertainty were calculated. Figure 8 shows how k1 varies with N. For small N, k1 measured falls below that of the theoretical values. As N becomes larger, the data conforms more to the theoretical logarithmic relationship, displaying a requirement of correction to scaling. By scaling the probability with the theoretical distribution, and the degree with the theoretical k1, a collapse of the datasets was attempted. Figure 9 shows its failure, thus we find that networks, which exhibit growth but not preferential attachment, are not scale-free.

Figure : A data collapse of the data displayed in Figure 7, done by scaling the probability with theoretical values and the degree size by theoretical cut-off. The data collapse fails. This shows that scale-free growth does not occur in this random network.

### Probability Distribution and Edge number

Fixing N = 106, the probability distribution of the random network was measured for between 1 and 16. The data, presented in Figure 10, follows theoretical expectations well. A χ2 test was carried out to assess goodness of fit, with results displayed in Table 2. All results show a low chi-square value with p-values close to unity, suggesting that the processed data fits well with the theoretical results. The best fit occurred for = 8.

|  |  |  |
| --- | --- | --- |
| **m** | **χ2** | **p-value** |
| 1 | 0.084 | 0.9991 |
| 2 | 0.004 | 0.99999994 |
| 4 | 0.003 | 0.99999997 |
| 8 | 0.001 | 0.99999999998 |
| 16 | 0.009 | 0.999999990 |

Table : The results for the chi-square test performed on 106 node networks, for various m. For a significance level of 1%, the p-value shows that all data sets have a high goodness of fit and correspond strongly to theoretical data. Whilst m = 8 produces the best goodness of fit value, m = 4 was used in analysis to be consistent and to optimise code execution time and good statistics.

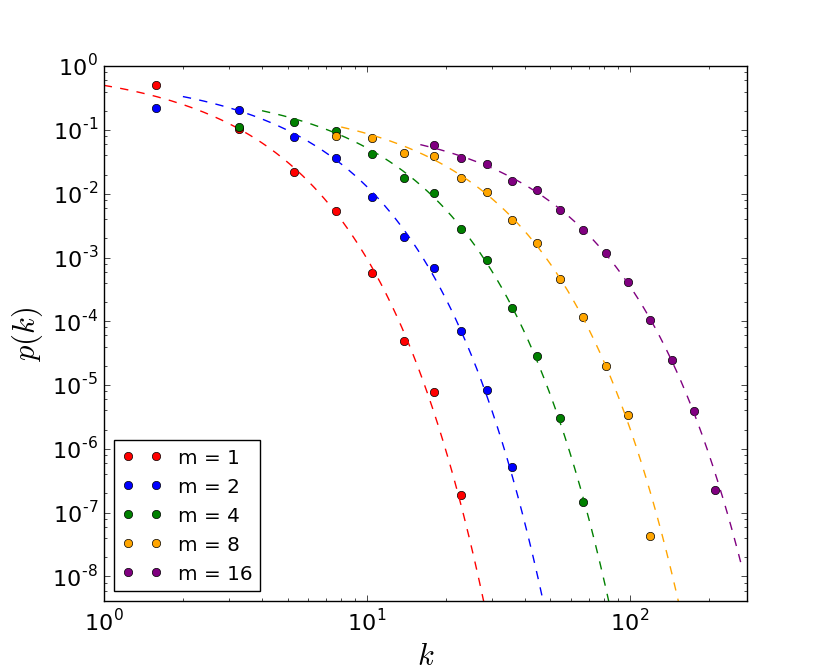


Figure : The variation of the probability distribution with m, with N = 106. Finite-size scaling is not obvious, however the cut-off begins deviate from expected values, particularly with increasing m. Small values of k also deviate from theoretical results, however this effect reduces with increasing m. The data was processed using log binning to remove statistical noise.

## Random Walks and Preferential Attachment

Using the random walk attachment scheme the probability distributions for N = 105 and = 4 were plotted for different step lengths, L. L = 0, 1 and 5, were analysed, with L = 5 corresponding to the diameter of the network, given by:[5]

|  |  |  |
| --- | --- | --- |
|  |  | (26) |

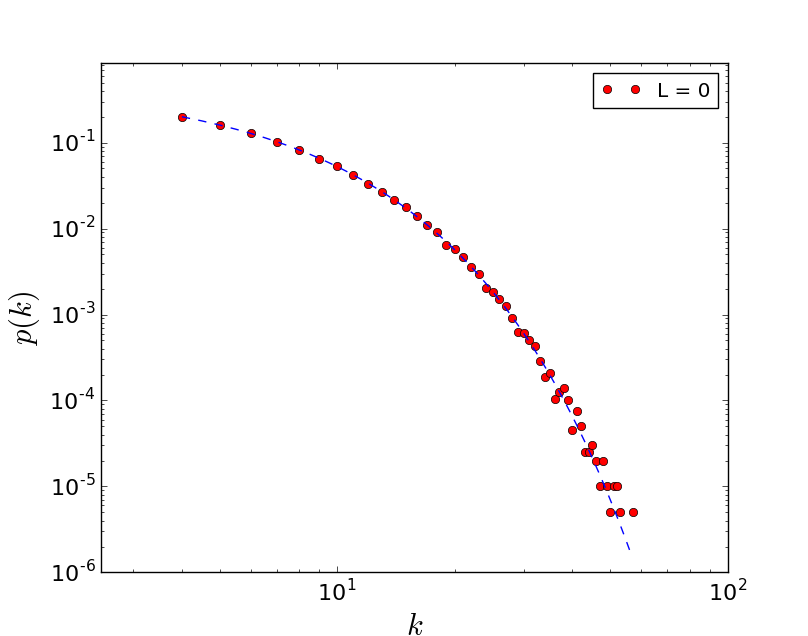
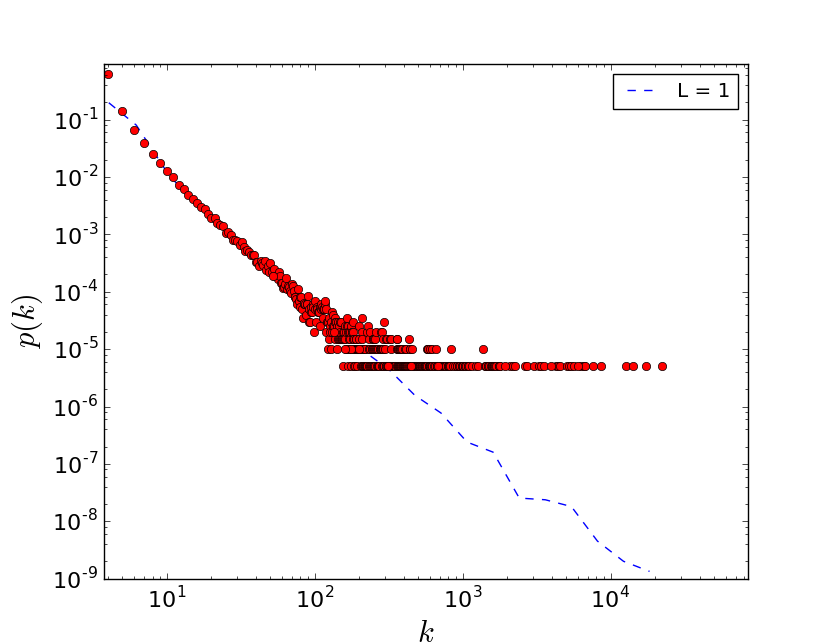
L = 0 produced an exponential distribution, as expected. Numerical results, presented in Figure 11, show that the raw data conformed to a theoretical random distribution.

Figure : Probability distribution for L = 1, m = 4 and N = 105. The blue dashed line represents the log-binned data. The data has strong characteristics of a fat-tailed distribution, however the slight curve in probability decay suggests that there are random effects still present in the data.

Figure : Probability distribution for L = 0, m = 4 and N = 105. The data is unprocessed. As expected, the plot follows the theoretical relationship for a pure random attachment network.

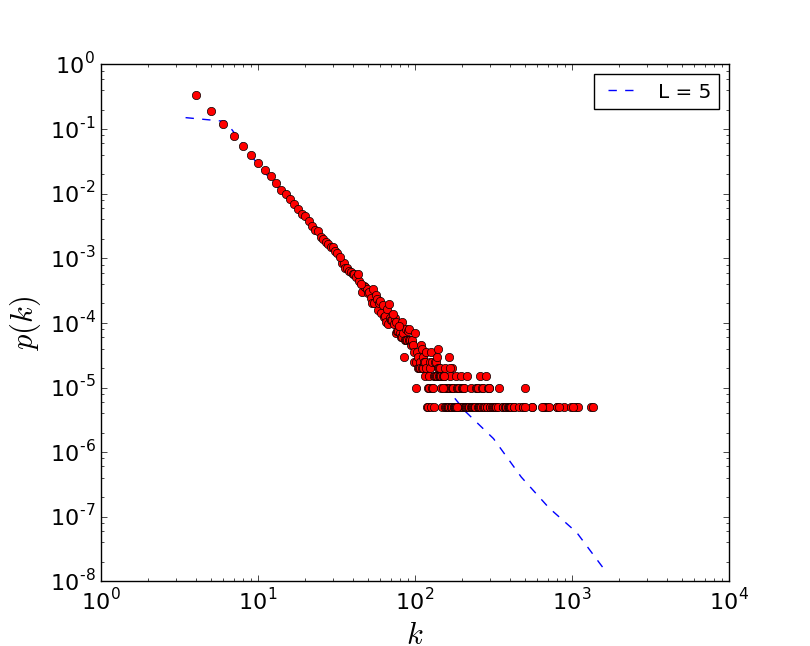
Increasing to L = 1 immediately transformed the data into a fat tailed distribution, implying preferential attachment. The data can be found in Figure 12. The scaling region was fitted after processing, and a gradient of -2.1 ± 0.1 was obtained, corresponding to a power decay of k-2.1. The exponent does not contain the Barabási-Albert continuum exponent of -3.0 to within three standard deviations of the mean, suggesting an element of randomness is inherent. L = 5 also produced a fat-tailed distribution, and fitting the scale-free region produced a gradient of -3.0
± 0.5. This more strongly suggests that the process displays preferential attachment and growth, as shown in Figure 13.

Figure : Probability distribution for L = 5, m = 4 and N = 105. The blue dashed line represents the log-binned data. The data has strong characteristics of a fat-tailed distribution and shows clear power law decay, calculated to be -3.0
± 0.5, therefore corresponding to Equation 10 in the large degree limit. The tail of the distribution produces a small double bump, which has possibly been inherited from the random walk, as L = 1 displays much more pronounced bumps. This effect is not present in pure preferential attachment.

This emergence of preferential attachment through seemingly random processes can be explained by considering node centrality. During a walk, particularly longer walks, hubs are more likely to be encountered, as they connected to more pathways. Thus, new nodes generate preferential attachment through local, random mechanisms; global knowledge of the network is not required. This corresponds to real life networks such as the Internet, where global network knowledge is impossible, however search engines can direct you to hubs.

# Conclusion

To conclude, the Barábasi-Albert model was implemented to study growth and preferential attachment in networks. The probability distribution and the cut-off degrees were measured as a function of system size, N, and edges created per node, . Using a chi-square statistic, the fat-tailed probability distribution shown to follow theoretical results to within a 10% significance level for > 1. The exponent of k1 relationship with N was calculated as 0.500 ± 0.002, strongly corresponding to the expected N0.5. Using this, and the theoretical probability distribution, a data collapse was successful performed.

After simulating a random network, an exponential distribution was obtained instead of a fat-tail. For different m, the simulation data fit the theoretical model to 1% significance. Measuring k1 with N showed that, for large N, the data followed the theoretical prediction. Data collapse was not successful for the random network.

Lastly, the random walk attachment method was implemented for L = 0, 1, 5. For L = 0, a random network probability was obtained, however for L = 1 and L = 5, a fat-tailed distribution emerged, implying the generation of preferential attachment through random, local processes.

# References

[1] A.-L. Barabási and R. Albert, (1999).“Emergence of scaling in random networks Science”, 286 173

[2] A.-L. Barabási (2014) “Network Science: The Barábasi-Albert Model” p. 1-12

[3] K. Christensen and N.R. Moloney, (2005), “Complexity and Criticality”, *Imperial College Press.* Appendix E

[4](2017) Stat Trek: Chi-Square Goodness of Fit Test, Available at: <<http://stattrek.com/chi-square-test/goodness-of-fit.aspx?Tutorial=AP>> [Accessed on: 14/03/2017]

[5] Cohen, Reuven; Havlin, Shlomo (2003). "Scale-Free Networks Are Ultrasmall". *Physical Review Letters*. 90 (5): 058701