Networks Project: Emergent self-organised criticality in growing networks

Nikolaos Koukoulekidis
CID: 00950301

19th April, 2016

Abstract: Real networks often obey scale-free degree distributions. This report develops theoretical and numerical models for the cases of pure preferential attachment and pure random attachment of nodes in growing networks. Then, the process of random walk is used to bridge these two methods and illustrate a mechanism for emergent self-organisation of networks growing by random walkers to acquiring a global scale-free degree distribution.

Word count: 2488 - excluding front page, figure captions and references.

1 Introduction

Real complex networks can often be described by a model growing network which follows a rule that dictates how new nodes are "attached" to existing ones. A common example is the World Wide Web, where websites and links represent the nodes and the edges of the network respectively [1]. A new node v_{new} is generally expected to connect with m existing nodes v_{ex}, each time according to a probability that scales positively with the degree $k_{v_{ex}}$,

$$\Pi(k) \propto k^{\gamma}, \gamma > 0. \tag{1}$$

Therefore, nodes with large degrees become "hubs", attracting new nodes more often. The case of $\gamma = 1$ results in the well-known Barabási and Albert (BA) model of pure preferential attachment [2]. Interestingly, only this value of γ results in a scale-free degree distribution [3], although this is a very common phenomenon in real networks. Therefore, there must be an underlying mechanism which gives rise to this scale-free behaviour.

This report aims to investigate such growing networks for the critical case of $\gamma = 1$ and the case of random attachment where $\gamma \to 0^{+}$, examining the properties of the degree distribution. Finally, it is shown that if the attachment rule is a random walk, the BA model emerges from the random attachment model, at the limit of long walk.

2 Results and Discussion

All models are numerically implemented with two parameters, the final number of nodes N_f and the number of edges m attached to each new node. Initialisation of the network can be performed in multiple ways, in general by producing a random graph with a number of nodes $N_0 \geq m$, to ensure each node has degree $k \geq m$. This constraint on the degrees is important for consistency with the theoretical degree distribution which assigns zero probability to degrees less than m as explained below equation (4). The network is chosen to be initialised at time $t = 0$ simply as a complete network of $m + 1$ nodes. Therefore, all nodes have degree $k = m$, and thus no artificial bias towards a particular node is introduced. Then, the network grows, $t \to t + 1$, by adding a node according to a rule varying with the model. The attachment process is repeated until all m chosen nodes are distinct. In other words, the network is chosen to contain no multiple edges and no self-loops. Time t is advanced until the network contains N_f nodes. Testing that the algorithm works, includes checking for multiple edges and self-loops as well as nodes with degree not bounded by m and $N(t) - 1$ at any time t. Evidence for any of these elements disqualifies the algorithm.

2.1 Pure Preferential Attachment

The attachment rule in this case is to connect with m nodes each time with probability

$$\Pi(k, t) = \frac{k}{2E(t)}, \tag{2}$$

where $E(t)$ is the number of edges at time t in the network. Since each edge has two stubs, the degrees sum up to twice $E(t)$, leaving the probability correctly normalised. This is practically achieved by selecting a node from the list of stubs at time t.

The network growth is dictated by the master equation (3) which defines the number of nodes $n(k,t)$ with degree k at time t recursively through time. Specifically, the number of nodes with degree k at the next time step is equal to the sum of four contributions.

$$
\begin{aligned}
& n(k,t+1) = \\
& + n(k,t) && \text{number of nodes with degree } k \\
& + m\Pi(k-1,t)n(k-1,t) && \text{expected number of nodes with degree } k-1 \\
& && \text{to attach to new node} \\
& - m\Pi(k,t)n(k,t) && \text{expected number of nodes with degree } k \\
& && \text{to attach to new node} \\
& + \delta_{km} && \text{one node added with degree } m
\end{aligned} \tag{3}
$$

The probability $p(k,t)$ of a vertex having degree k at time t is defined by the relation $n(k,t) = p(k,t)N(t)$. The distribution of the network at long time limit $p_\infty(k)$ can be defined based on p so that $p_\infty(k) = \lim_{t\to\infty} p(k,t)$. This requires large values of N, since $t = N(t) - N_0$. Using the definitions of all quantities that appear in (3), rearranging yields the recursive relation

$$p_\infty(k) = \frac{1}{2}[(k-1)(p_\infty(k-1) - kp_\infty(k)] + \delta_{km}. \tag{4}$$

Given the initial conditions used in the numerical simulations and the fact that each new node is attached to m existing ones, $p_\infty(k) = 0, \forall k < m$. Then,

$$p_\infty(k) = \frac{2}{1+m}, \qquad k = m \tag{5}$$

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2} \Rightarrow p_\infty(k) \propto \frac{\Gamma(k)}{\Gamma(k+3)}, \qquad k > m \tag{6}$$

Equations (5), (6) lead to the normalised formula

$$p_\infty(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{7}$$

$$\Rightarrow \sum_{k=m}^{\infty} p_\infty(k) = 2m(m+1)\left(\frac{1}{2}\frac{1}{m} - \frac{1}{2}\frac{1}{m+1}\right) = 1, \tag{8}$$

since all terms in the sum, appearing after splitting the fraction in (7), cancel with each other except for the two in parentheses.

This result is compared to numerical data in figure (1 left) for $m = 1$. Parameter $N_f = 10^7$ is as large as possible to account for the limit of large time. The numerical distribution illustrated by the blue dots looks very consistent with the theoretical formula, while a fat tail appears at large k where statistics were poor. Since every new node is attached to

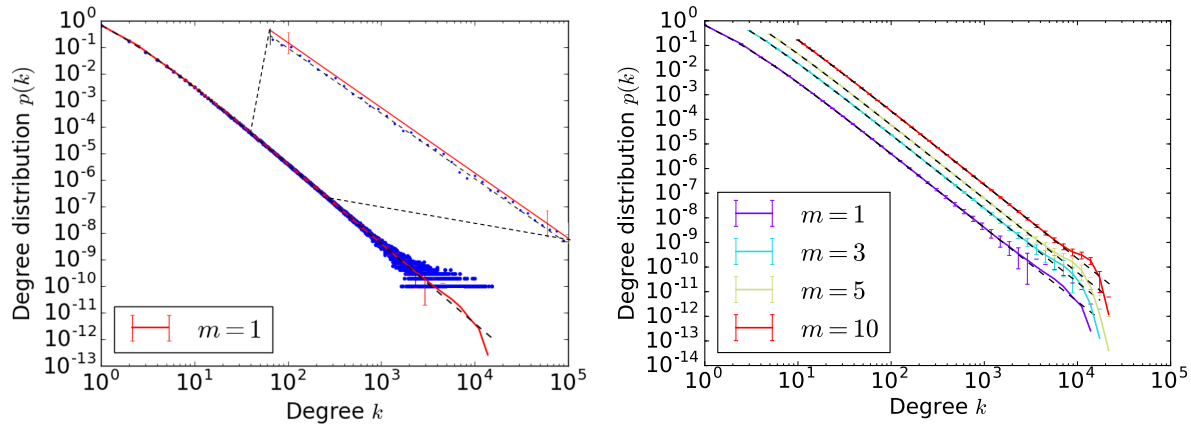

Figure 1: Numerical and theoretical degree distribution of BA model for $m = 1$ (left) and various m (right). Data are log-binned, always with base $a = 1.25$, and plotted for $N_f = 10^7$ starting from $k = m$. Results were averaged over 1000 networks to produce better statistics, hence the error bars. Left figure also plots raw data to demonstrate by zooming in on the distribution that the log-binning function loses some information about the data by shifting the distribution upwards.

m existing ones, it is expected that most nodes will have a degree close to m, while only few will have a degree $k \gg m$. Few nodes in the early network will attract several new ones by chance and then they will "snowball" to these large values of the distribution due to the preferential attachment used. This is "the rich get richer" principle described by Barabási and Albert [2].

Two ways of tackling the issue were used. Firstly, results were log-binned to smoothen out the tail. The red log-binned curve in figure (1 left) illustrates the smooth cut-off that replaces the fat tail. Log-binning is used throughout the report for illustrative purposes. However, there is no inverse procedure to return to the raw data, therefore some information is lost. This is shown in figure (1 left), by zooming in on the distribution. The raw data randomly lie on both sides of the theoretical curve, while the log-binned data are all consistently slightly above it, due to the approximations used to normalise them. The second solution used is to average data over a large ensemble of networks, in this case a thousand networks, therefore increasing the number of observations at large degrees. This gives rise to statistical uncertainty encapsulated by error bars.

Figure (1 right) shows data for various values of m. All distributions closely follow the corresponding theoretical one and have a rapid cut-off at large k. However, there are three key points that favour low values of m, used hereinafter. Large values do not expose the start of the distribution as is clear in the figure and are computationally more expensive.

In addition, the distribution can be rewritten as

$$p_\infty(k) = \frac{2m(m+1)}{k^3\left(1+\frac{3}{k}+\frac{2}{k^2}\right)} = \frac{2m(m+1)}{k^3}\left(1-\frac{3}{k}+\mathcal{O}(k^{-2})\right), \tag{9}$$

so it is clear that corrections to scaling are more evident for large m.

Statistical tests for goodness-of-fit were used to quantify to what extent data and theory match. Since log-binning loses information about the data, statistical analysis was performed only on raw data directly.

The Pearson χ^2 test is a commonly used non-parametric test [4]. The latter property is desirable, since parametric tests assume that the theoretical distribution is parametrised in specific ways, usually assuming gaussianity. It is a measure of relative distance between numerically observed and theoretically expected data suitable for discrete datasets. It requires a minimum sample size at each degree k, which means that it does not work at the tail of the distribution, as illustrated in figure (2.1). The p-value plotted is the probability that the numerical data obey the null hypothesis, i.e. fit the theoretical distribution. Indicating the significance level at 10% on the graph, a p-value higher than that suggests strong evidence of the null hypothesis. Looking at figure (2.1), the p-value starts having a negative trend at roughly $k = 10^3$ which corresponds at the value where the fat tail starts to emerge in figure (1 left). Excluding the tail a p-value greater than 0.5 is obtained. A reduced χ^2 is not suitable since it leads to very small statistics ($\sim 10^{-18}$) which suggests overfitting of the data due to the huge number of numerical data points.

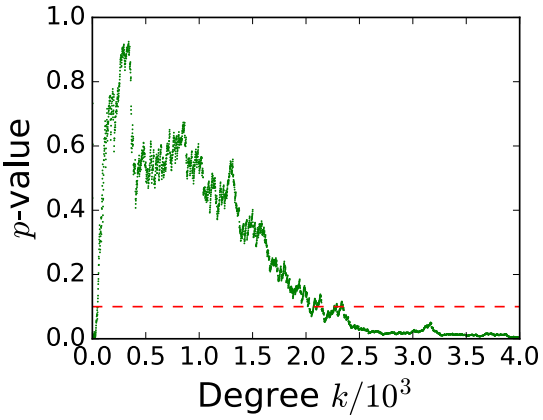

Figure 2: Scatter plot of the χ^2 p-value against degree for $N = 10^7$ and $m = 3$. Including the tail reduces the p-value drastically, since the test is not valid in this regime. The red line indicates the 10% significance level.

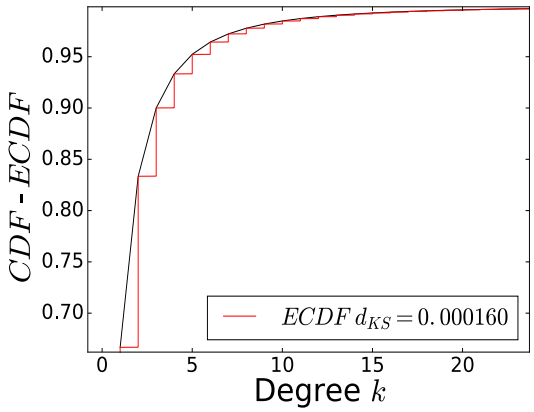

Figure 3: Plot of the CDF of the theoretical distribution and the indicator $ECDF$ function of a sample with size 10^6 extracted according to the numerical distribution for $N = 10^7$ and $m = 3$. The KS statistic quoted in the legend is the maximum vertical difference between the two functions.

Another commonly used statistic can be obtained by the Kolmogorov-Smirnov (KS) test. This is also a non-parametric test, but only applies to continuous probability distributions. It calculates the maximum difference between the cumulative theoretical distribution and the indicator function of the observed data. The indicator function describes how many of the observed data lie up to a certain degree. The plot of figure (2.1) shows that the two functions are very close to each other with a maximum difference of $d_{KS} \approx 1.6 \times 10^{-4}$ which is well below the critical value 1.2×10^{-3} in this case [5]. The main problem with the test is that the numerical distribution is highly binned at low k which disqualifies it from being approximated as continuous.

Extending the KS test to discrete distributions is non-trivial [6], but a two-sampled KS test can be used instead. This implies that a sample from the theoretical distribution is obtained as well as a sample from the numerical. This relaxes the conditions on continuity, but it is a test of low statistical power, that is low probability of rejecting the null hypothesis given it is false. This is reflected by the fact that ignoring the fat tail, the test yield a p-value of 1.0.

It is instructive to investigate how the mechanism of the preferential attachment depends on the network size. Figure (4) plots the degree distribution for $m = 1$ over different sizes. The distributions have the same slope, approximately close but less than -3 due

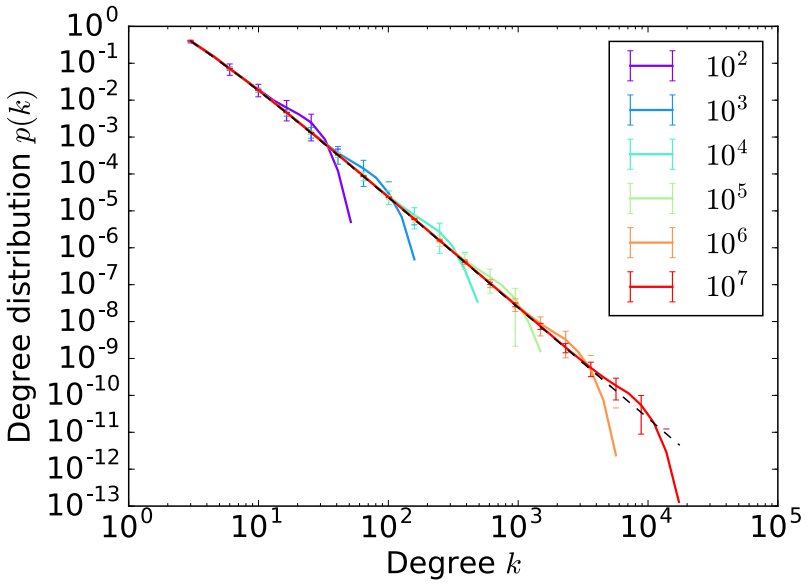

Figure 4: Degree distribution of BA model for $m = 3$ and different N_f. Data are log-binned with base $a = 1.25$ and averaged over 1000 networks, hence the error bars. Distributions seem to have the same slope, but incrementing cut-offs.

to the corrections to scaling of equation (9), as well as cut-off values, due to the finiteness of the system, that are equidistant, with larger networks accessing larger values of degrees.

Defining k_{max} as the degree at which the cut-off appears, it seems to scale with N. Assuming the expected value of nodes with degree $k \geq k_{max}$ is one and using equation (7),

$$N \sum_{k=k_{kmax}}^{\infty} p_\infty(k) = 1 \Rightarrow k_{max} = \frac{-1 + \sqrt{1 + 4Nm(m+1)}}{2} \tag{10}$$

$$N \to \infty, \quad k_{max} \to \sqrt{m(m+1)}N^{\frac{1}{2}}. \tag{11}$$

Step (10) is obtained by solving the exact difference equation as in step (8) where all terms except two cancel, but equation (11) is equally useful and simpler when working with large networks ($N = 10^7$). Figure (5 left) provides a log-log plot of the numerical cut-off scaling. The slope of the plot is (0.502 ± 0.018) and corresponds to the exponent of

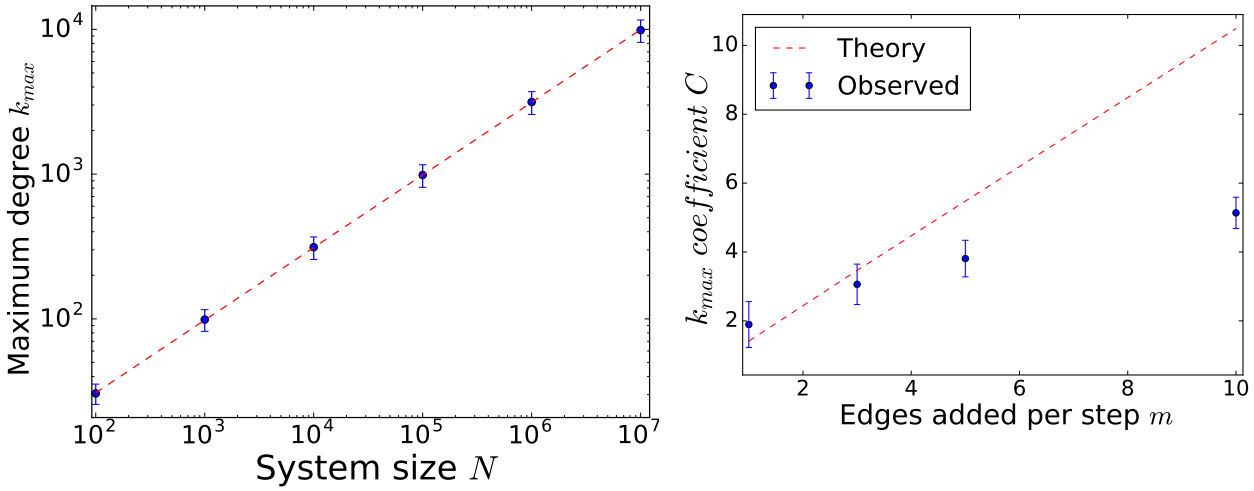

Figure 5: Cut-off scaling for $m = 3$. The exponent of network size, obtained as the slope of the left plot is (0.502 ± 0.018). The right plot shows how the coefficient of network size diverges with m from the theoretical distribution.

network size N, hence being consistent within error with the theoretical derivation (11). The error is carefully propagated from the large ensemble of networks considered at each size, through k_{max} data points to the slope. The constant exponent dictates the equal distance between cut-offs in figure (4). The plot on the right shows the dependence of the coefficient $C(m)$, defined as $k_{max} = C(m)N^{\frac{1}{2}}$ for numerical data. In this case there is a rapid departure from the theoretical prediction for large m, reinforcing the argument suggested by equation (9) that large values of m are associated with more profound corrections to scaling. Low m stay close to theory within error.

Rescaling the observed distribution with the theoretical one as well as rescaling the degrees with k_{max} attempts a horizontal and vertical collapse of the distribution respectively. The result is presented in figure (6). The distributions are indeed collapsed, so that

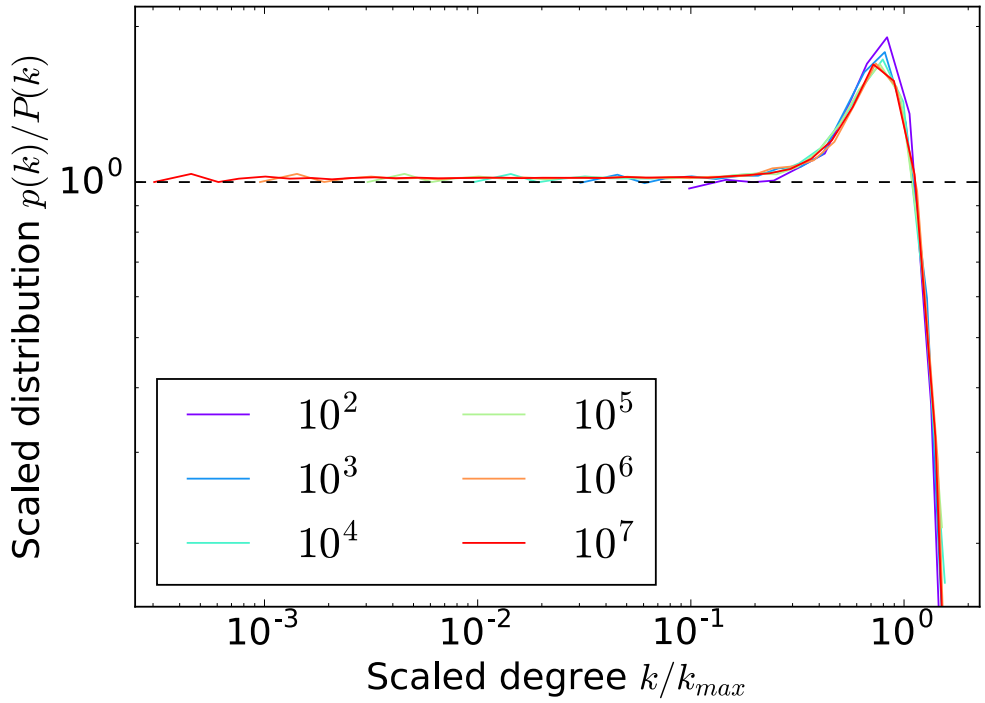

Figure 6: Data collapse of data in figure (4). The mechanism of preferential attachment is independent of network size.

$$\frac{p_{observed}(k)}{p_{\infty}(k)} = \mathcal{G}\left(\frac{k}{k_{max}}\right) \tag{12}$$

The scaling function $\mathcal{G}(x)$ is 1 for $x \ll 1$, then has a bump before rapidly decaying for $x \gtrsim 1$. Clearly, the decay is a finite-size effect that disappears in the true infinite time limit as the system is critical. The mechanism of preferential attachment evidently does not depend on network size and the BA model forms a universality class.

2.2 Pure Random Attachment

The probability of attachment is now

$$\Pi(t) = \frac{1}{N(t)}, \tag{13}$$

which is normalised correctly at any given time t, as it is clear by summing over all nodes. The results of a random attachment model are briefly presented with comments on the differences between the BA model.

The probability in this limiting case does not depend on the degree of the node at all and hence we would expect a degree distribution that is not a power law [3] as illustrated in figure (7).

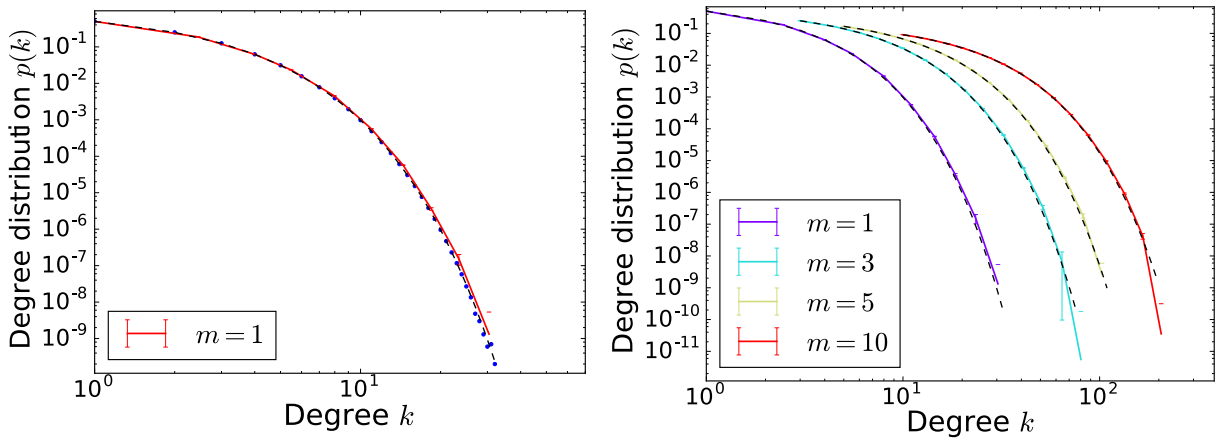

Figure 7: Numerical and theoretical degree distribution of random attachment model for $m = 1$ (left) and various m (right). Data are log-binned, always with base $a = 1.25$, and plotted for $N_f = 10^7$ starting from $k = m$. Results were averaged over 1000 networks to produce better statistics, hence the error bars. Left figure also plots raw data to demonstrate.

Regardless of m, the distributions seem to decay rapidly for random attachment. The plot on the left shows clearly that the tail of the distribution is not fat, since the raw data reamin close to the theoretical distribution. The log-binning still possesses a, less sharp, cut-off due to the finiteness of the system, evident in the case of $m = 10$ on the right plot.

The theoretical derivation of the degree distribution is similar to that in the BA model, only now using the new attachment probability given in (13). Starting from the master equation (3) and using identical definitions,

$$p_\infty(k) = \frac{m p_\infty(k-1) + \delta_{km}}{1+m}. \tag{14}$$

Taking $p_\infty(k) = 0, \forall k < m$,

$$p_\infty(k) = \frac{1}{1+m}, \qquad k = m \tag{15}$$

$$p_\infty(k) = \frac{m}{1+m} p_\infty(k-1), \qquad k > m \tag{16}$$

yielding,

$$p_\infty(k) = \frac{1}{1+m}\left(\frac{m}{1+m}\right)^{k-m} \Rightarrow p_\infty(k) \propto e^{-k\ln\left(1+\frac{1}{m}\right)}. \tag{17}$$

which is normalised, setting $n = k - m$, since the geometric sum converges,

$$\sum_{k=m}^{\infty} p_\infty(k) = \frac{1}{1+m}\sum_{n=0}^{\infty}\left(\frac{m}{1+m}\right)^n = 1. \tag{18}$$

The distribution (17) is in fact an exponential decay, which confirms that it is not fat-tailed. It also fits the numerical data quite well, with a p-value of ~ 0.95 from the χ^2 test before the tail. The one-sample KS test gives a statistic $d_{KS} = 6.0 \times 10^{-4}$, which is still less than the critical value which remains at 1.2×10^{-3}. Finally, the two-sample KS test also gives a positive result with p-value of 0.96.

Figure (8) illustrates the scaling of the distribution with network size. The cut-offs do

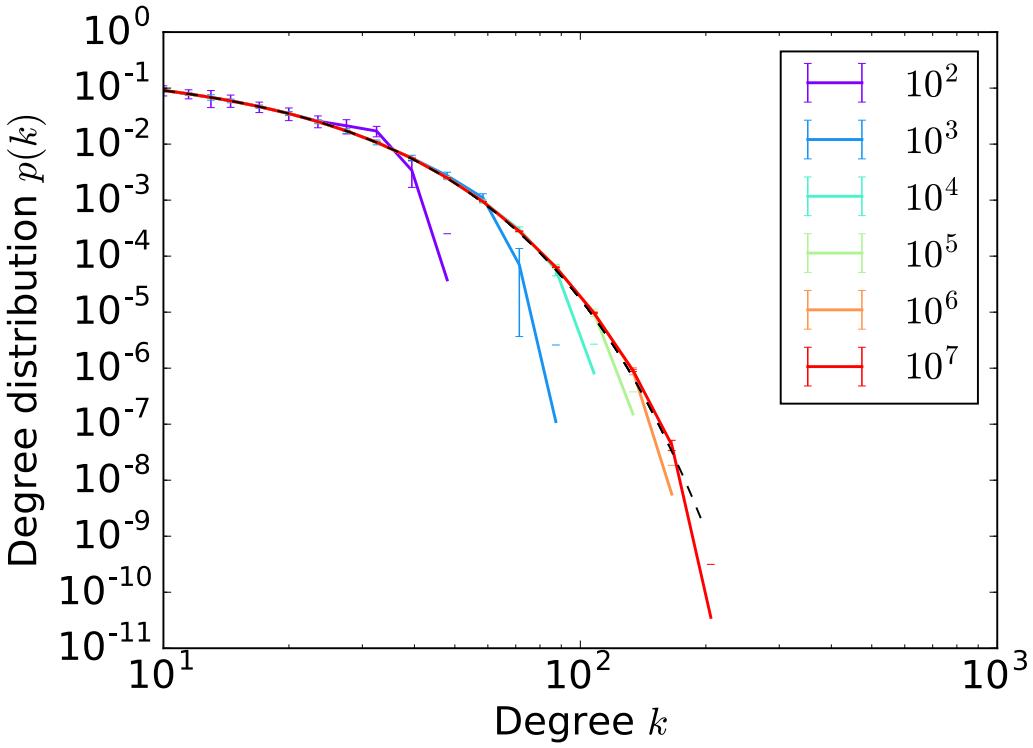

Figure 8: Degree distribution of random attachment model for $m = 3$ and different N_f. Data are log-binned with base $a = 1.25$ and averaged over 1000 networks, hence the error bars. Distributions seem to have the same slope, but incrementing cut-offs.

not look equidistant in this case, which suggests that they do not scale as a power law of N. Using distribution (17) and the geometric sum at step (20),

$$N \sum_{k=k_{kmax}}^{\infty} p_\infty(k) = 1 \tag{19}$$

$$\Rightarrow \frac{1}{1+m}\left(\frac{1+m}{m}\right)^m \sum_{k=k_{kmax}}^{\infty} \left(\frac{m}{1+m}\right)^k = \frac{1}{N} \tag{20}$$

$$\Rightarrow k_{max} = m + \frac{\ln N}{\ln\left(1+\frac{1}{m}\right)} \tag{21}$$

$$N \to \infty, \quad k_{max} \to \frac{\ln N}{\ln\left(1+\frac{1}{m}\right)}. \tag{22}$$

Figure (9) plots formula (22) for $m = 3$.

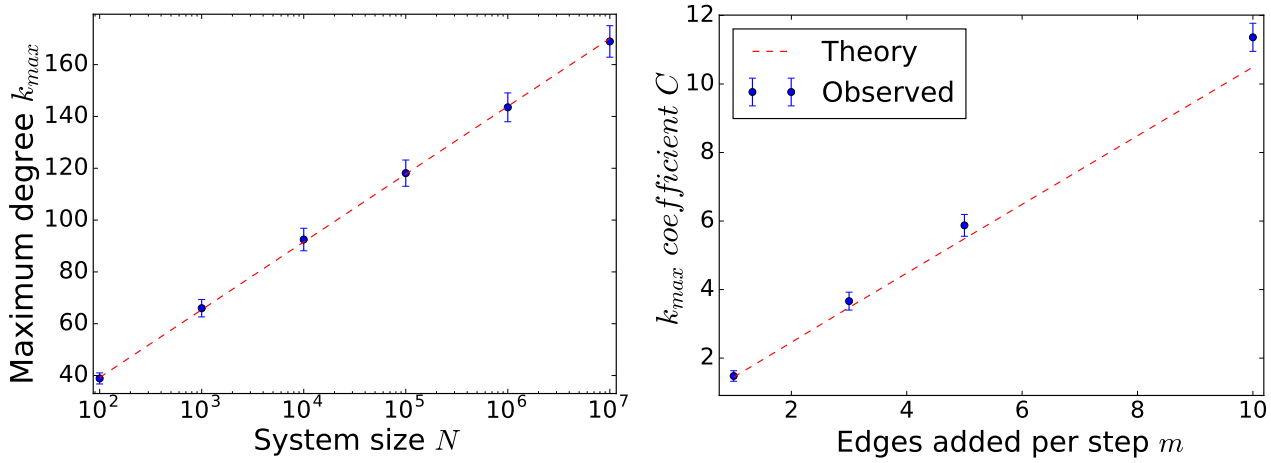

Figure 9: Cut-off scaling for $m = 3$. The coefficient of network size in the left plot is (3.66 ± 0.26), in agreement with the theoretical value. The right plot shows the coefficient diverges with m from the theoretical distribution.

The slope of the left plot is the coefficient $C(m)$, where $k_{max} = C(m) \ln N$. It is estimated to be (3.66 ± 0.26), while equation (22) suggests that $C(m = 3) = 3.48$, meaning that the data are in agreement with the theoretical formula within error. The plot on the right in figure (9) indicates how $C(m)$ diverges with m.

An attempt for data collapse in figure (10) does not have a clear result, but it is possible that all sizes collapse on a scaling function, identical in behaviour as equation (12).

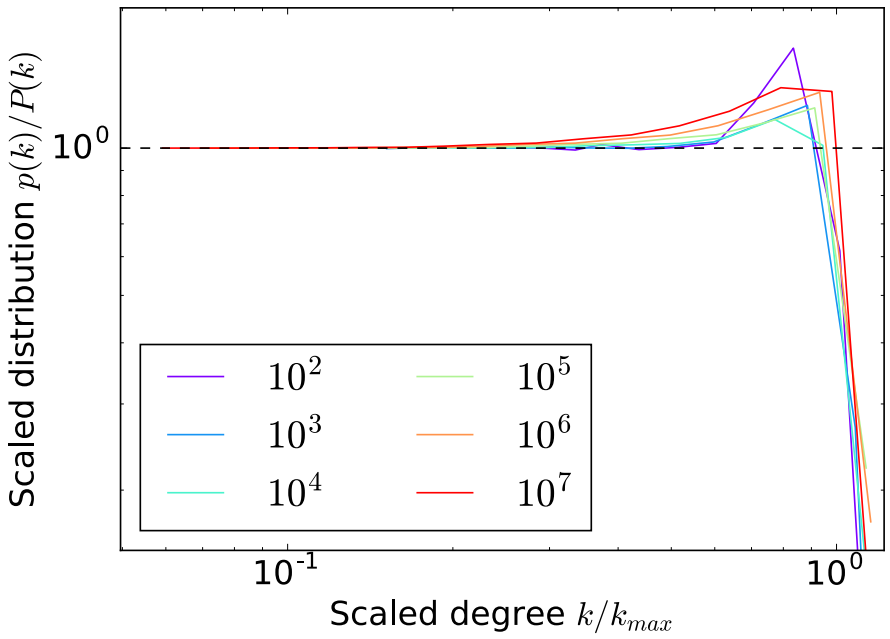

Figure 10: Data collapse of data in figure (4). The mechanism of random attachment is possibly independent of network size.

2.3 Random Walks

The random attachment seems to be a very natural way to connect in a real network. New websites joining the Web link to other nodes without having knowledge of global properties [8] like how many edges the network has. In principal, they link to a small fraction of similar websites and the links of these websites etc. Therefore, they "attach" through a random walk of some length ℓ with pure random attachment being the special case of $\ell = 0$.

Degree distributions for various walk lengths are plotted in figure (11) for $m = 3$. For all non-zero cases of length, the distribution is fat-tailed. Length one in particular includes a lot of high degrees. As length increases further, the distributions in figure (11 left) tend to coincide with equation (7) for preferential attachment. Lengths 5 and 10 already do not have a clear difference and give p-values well above 0.5 in a KS two-sampled test with the pure preferential attachment distribution, excluding the tails. The slope in figure (11 right) for $\ell = 1$ overshoots over -3, while $\ell = 5$ is already converging to -3 that corresponds to preferential attachment.

This tendency of the network to change behaviour from random to preferential attachment makes sense mathematically for a strongly connected network, which is the case in a growing unweighted, undirected network. Defining $\boldsymbol{w}(\ell)$ such that w_i is the probability of walking to the i^{th} node after a walk of length ℓ, then

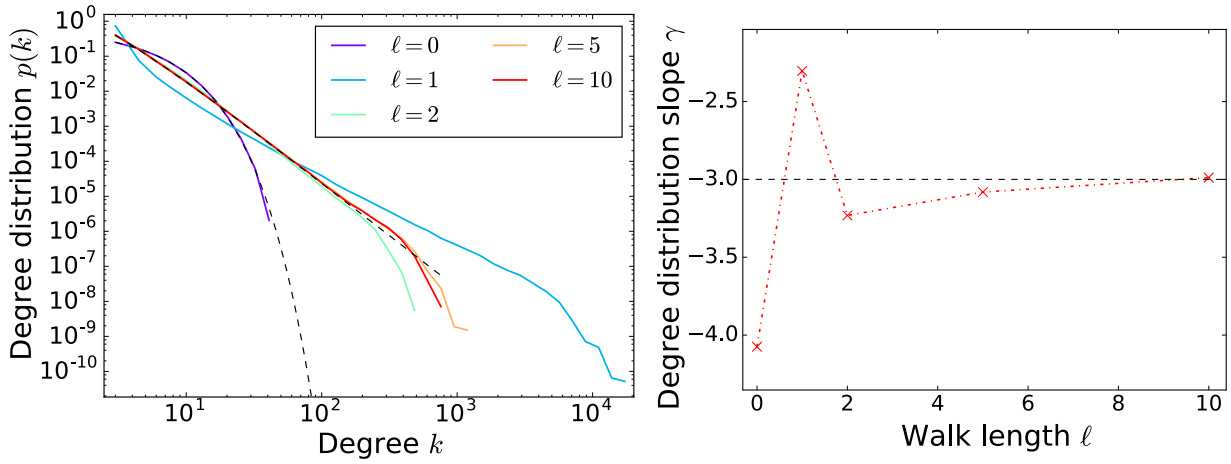

Figure 11: Growing network with attachment via random walk of different lengths for $m = 3$ and $N = 10^4$. Data are log-binned with base $a = 1.25$ and averaged over 500 networks. As the length increases, the degree distributions tend to the distribution of pure preferential attachment.

$$\boldsymbol{w}(\ell) = \boldsymbol{T}^{\ell}\boldsymbol{w}(0), \tag{23}$$

where $\boldsymbol{T}$ is the transfer matrix. The Perron-Frobenius theorem allows decomposition of the vector $\boldsymbol{w}(0)$ into eigenvectors $\boldsymbol{v}_n$ of the transfer matrix which span the space, so that

$$\boldsymbol{w}(\ell) = \sum_{n=1}^{N} c_n \boldsymbol{T}^{\ell}\boldsymbol{v}_n \to c_1 \lambda_1^{\ell}\boldsymbol{v}_1 \text{ as } \ell \to \infty, \tag{24}$$

where only the largest eigenvalue λ_1 survives at the limit, since the theorem suggests that $\boldsymbol{T}$ is diagonalisable for strongly connected networks and has a unique largest eigenvalue. But the corresponding eigenvector $\boldsymbol{v}_1$ has entries proportional to the degree of the associated node. Therefore, the probability distribution for each node flows from pure random to preferential.

$$\lim w_i(\ell)_{\ell\to\infty} \propto k_i. \tag{25}$$

This analysis breaks down for a bipartite networks which realistically appear only for $m = 1$, since in this case any node partitions the network into nodes that are either even or odd number of edges away. This is clear in figure (12), where distributions for random walks of even and odd lengths tend to different limits.

Therefore, a random walk is a local process which tends to lead to a power law distribution. Since the process is local, it can be concluded that the network self-organises to a steady state which is critical, i.e. scale free.

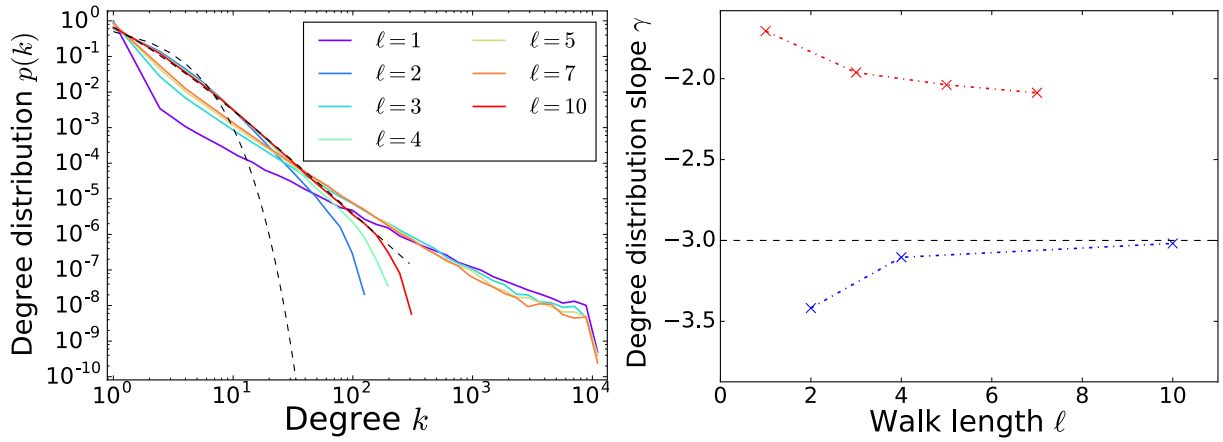

Figure 12: Growing network with attachment via random walk of different lengths for $m = 1$ and $N = 10^4$. Data are log-binned with base $a = 1.25$ and avearged over 500 networks. As the length increases, the degree distributions tends to two different limits for even and odd lengths.

3 Conclusion

The BA model of preferential attachment and the case of random attachment were studied with respect to degree distribution in the context of growing network models. The numerical data obeyed the theoretical expectations, with the two cases presenting a power law and an exponentially decaying behaviour respectively. At large network sizes, corrections to scaling were mainly important for large values of m.

The more general case of a random walk as an attachment process was then investigated and proven to be a mechanism that bridges random attachment to the emergence of power law behaviour in the degree distribution of the networks. Results can be generalised to directed and weighted graphs [7], while the mechanism is significant in explaining real complex networks like citations and WWW [8].

References

[1] A. Vázquez, "Knowing a network by walking on it: emergence of scaling", Havana University, Cuba, 2000.

[2] A.L. Barabási and R. Albert, "Emergence of Scaling in Random Networks", Science, 1999, Vol 289 509-512.

[3] P.L. Krapivsky, S. Redner and F. Leyvraz, "Connectivity of Growing Random Networks", Phys.Rev.Lett., 2000, **85** 21 4629.

[4] Chakravarti, Laha, and Roy, *Handbook of Methods of Applied Statistics*, Volume I, 1967, John Wiley and Sons.

[5] L. Sachs, "Angewandte Statistik", Springer, 1997 427-431,

[6] T.B. Arnold and J.W. Emerson, "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions", The R Journal, Vol 3/2,

[7] T.S. Evans and and J.P. Saramäki, "Scale Free Networks from Self-Organisation", Imperial College London, London, 2005. 2011.

[8] J. Saramäki and K. Kaski, "Scale-Free Networks Generated By Random Walkers", Helsinki Univerity of Technology, Helsinki, 2004.